KTH Computer Science
and Communication

Investigating Communicative Feedback Phenomena across Languages and Modalities

LOREDANA CERRATO

Doctoral Thesis
Stockholm, Sweden 2007

KTH Computer Science and Communication
Department of Speech, Music and Hearing
10044 Stockholm, Sweden

GSLT Graduate School of Language technology
Faculty of Arts, Göteborg University
40530 Göteborg, Sweden

TRITA-CSC-A 2007:3
ISSN-1653-5723
ISRN-KTH/CSC/A--07/03—SE
ISBN 978-91-7178-632-6

Akademisk avhandling som med tillstånd av Kungliga Tekniska Högskolan framlägges till offentlig granskning för avläggande av filosofi doktorexamen i tal och musikkommunikation tisdagen den 15 maj 2007 klockan 15.00 i sal F3, Sing-Sing, Kungliga Tekniska Högskolan, Lindstedtsvägen 26, Stockholm

Tryck: Universitetsservice US AB

Abstract

This thesis deals with human communicative behaviour related to feedback, analysed across languages (Italian and Swedish), modalities (auditory versus visual) and different communicative situations (human-human versus human-machine dialogues).

The aim of this study is to give more insight into how humans use communicative behaviour related to feedback and at the same time to suggest a method to collect valuable data that can be useful to control facial and head movements related to visual feedback in synthetic conversational agents.

The study of human communicative behaviour necessitates the good quality of the materials under analysis, the support of reliable software packages for the audio-visual analysis and a specific coding scheme for the annotation of the phenomena under observation.

The materials used for the investigations presented in this thesis span from spontaneous conversations video recorded in real communicative situations, and semi-spontaneous dialogues obtained with different eliciting techniques, such as map-task and information-seeking scenarios, to a specific corpus of controlled interactive speech collected by means of a motion capture system. When motion caption is used it is possible to register facial and head movements with a high degree of precision, so as to obtain valuable data useful for the implementation of facial displays in talking heads.

A specific coding scheme has been developed, tested and used to annotate feedback. The annotation has been carried out with the support of different available software packages for audio-visual analysis.

The procedure followed in this thesis involves initial analyses of communicative phenomena in spontaneous human-human dialogues and human-machine interaction, in order to learn about regularities in human communicative behaviour that could be transferred to talking heads, then, for the sake of reproduction in talking heads, the investigation includes more detailed analyses of data collected in a lab environment with a novel acquisition set-up that allows capturing the dynamics of facial and head movements.

Finally the possibilities of transferring human communicative behaviour to a talking face are discussed and some evaluation paradigms are illustrated. The idea of reproducing human behaviour in talking heads is based on the assumption that the reproduction of facial displays related to communicative phenomena such as turn management, feedback production and expression of emotions in embodied conversational agents, might result in the design of advanced systems capable of effective multi-modal interactions with humans.

Alla vita

Acknowledgements

I would like to thank all the people who, in one way or another, have contributed to make my time as PhD student memorable, and all the people who have guided me towards the enormous achievement that a PhD thesis means for me.

In the first place I wish to thank my two excellent supervisors: Professor David House and Professor Jens Allwood.

David House, my first supervisor, has always been ready to listen to me and to experience my swinging moods. He has enormously helped me shaping my ideas, structuring my findings and discussing my results, always contributing with motivating and constructive comments.

Jens Allwood, my second supervisor, provided inspiration of the topic of this thesis: communicative feedback. He has always been supportive with his profusion of enthusiasm for my research, his stimulating, constructive comments and his strong and inflexible, and often provocative, points of view. To both my deepest gratitude.

Besides the supervisors, many people have been helpful in one way or another during the work on this thesis.

Colleagues are important elements in my PhD student formation, and, especially for me, being a linguist among engineers and technicians at TMH, the help of Jonas Beskow, Gunilla Svanfeldt, Kåre Sjölander and Giampiero Salvi has been indispensable, as well the helping hand of Niclas Horney who has given me the possibility to have a good relationship with my PC.

In particular I would like to express gratitude to Jonas Beskow and Gunilla Svanfeldt.

Jonas has helped me many times with technical matters and constantly amazed me with his brightness. For this, for productive collaboration during the years, for useful comments to the first draft version of chapter 8 and for his contribution to chapter 9, I wish to gratefully acknowledge him.

Gunilla has been a fantastic roommate, always glad and positive, for this and for her contribution to chapter 8, I wish to gratefully acknowledge her.

Colleagues can also help in other ways: to give emotional support, to brighten up dark days (which is very necessary when living in Sweden!) to make lunch and coffee breaks pleasant. For this I have to thank all my colleagues and ex-colleagues at TMH, the lively people of the language group and the jovial Maestro Musices.

In particular I would like to show appreciation for:

Botond Packucs for his contagious laugh, his good mood and shared interest in research, good food and wine. Rebecca Hincks, Beata Megeysy and Tina Magnuson for their emotional support and nice corridor chitchats; Rebecca Hincks deserves my deepest gratitutde also for proofreading of the whole thesis.

Giampiero Salvi, my Italian colleague, for giving me the possibility to practise my native language and discuss trivial matters related to Italian football and politics as well as more complex matters as statistics!

Kjell Gustavsson, a perfect gentleman, an authentic linguist, merits a special thanks, in many languages: *grazie, merci, tack, kiitos*… for being able to always find interesting topics of conversation related to language, linguistics, travels and Italian literature.

One of the most rewarding aspects of my experience as a PhD student at TMH has been the opportunity to be in contact with lots of interesting people met at the department as well as at seminars, conferences, summer schools, in the frameworks of Nordic collaborations and EU-projects, and of course within the Graduate School of Language Technology.

Having the honour of sitting in the room next to Gunnar Fant has made my passion for research even stronger, and all the people I have had fruitful and enjoyable collaborations with during the years have also contributed to keep my passion alive over the years. I would like to express my thankfulness to all these people, among which in particular: Jens Allwood, Piero Cosi, MariaPaola D'Imperio, Susanne Ekeklint, Kristiina Jokinenen, Costanza Navarretta, Nadia Mana, Patrizia Paggio and Mustafa Skhiri.

Many thanks to the people who provided me with the materials used to carry out some of the studies reported in this thesis:

Federico Albano Leoni and his collaborators at CIRASS, University of Naples, Pètur Helgason from Stockholm and Uppsala University, Sweden, for the map-task dialogues analysed in chapter 4.

Jens Allwood and his collaborators at Gothenburg University, in particular Magnus Gunnarsson and Leif Grönqvist, for the dialogues selected from the GSLC corpus analysed in chapter 5 and for technical support with Multitool.

Researchers from TMH and Telia Research who collected and analysed the AdApt databases that I have used in studies presented in chapter 6 and 7 in this thesis, in particular Joakim Gustafson, Linda Bell, Johan Boye, Matts Wirén, Magnus Nordstrand, and especially Jens Edlund and Anna Hjalmarsson.

Thanks also to my colleagues from TMH with whom I worked hard to collect the 3D data analysed in chapter 8: Jonas Beskow, Björn Granström, David House, Mickael Nordeberg, Magnus Nordstrand, and Gunilla Svanfeldt. The data collection was carried out at the Qualisys Lab at Linköping University, which was kindly made available by Bertil Lyberg.

Special thanks to the subjects who participated in the data acquisition sessions and to the several people who took part in the evaluation tests.

Joakim Nivre, vice-chairman of the GSLT board, and Lars Ahrenberg my official "granskare", deserve my gratitude for their engagement in the school, their supportive attitudes towards the students, and invaluable help in the compilation of my yearly study plans.

What makes a PhD student life easier is the affection of family and friends. For this I would like to thank my family and friends in Italy, Sweden and the rest of the world for being so supportive especially in the latest difficult months of my life.

In particular:

grazie ai miei amati genitori, Vittorio e Rosalba, e ai miei adorati fratellini Pierluca e Giorgio; alle mie cugine Antonella, Annamaria e Paola, ai miei amici di sempre Attilio, Francesco, Paola e Michela, per essermi stati vicini anche da lontano.

Thanks to Maria for her support and contagious enthusiasm for life, thanks to "magic" Romany for being always close even if faraway from Sweden.

Tack till min familj i Sverige: Erik, Lars and Anita Sundberg, for being very understanding, supportive and helpful during my studies. Anita needs a special thank also for meticulous proofreading of the whole thesis.

Erik, my husband, deserves my deepest thankfulness not only for repeatedly reading through this thesis and providing me with generous feedback to improve some of the "Italian biased" formulations, the graphical details and the format of the thesis, but also for being such a thoughtful, encouraging and helpful partner during my studies. He is actually the real reason why I ended up in Sweden and became a PhD student at KTH…and, last but not least I have to acknowledge my wonderful daughter Elisa, for being the best result of my PhD student career!

This study was carried out at KTH with the support of the Swedish Graduate School of Language Technology (GSLT), the Centre for Speech Technology (CTT) and the Nordic Academy of Advanced Study (NorFA). The Ragnar and Astrid Signeuls Foundation has contributed to some of the traveling and conference expenses.

Grazie a tutti!
Loredana

Loredana Cerrato's publications

Cerrato, L. (2002) A comparison between feedback strategies in Human-to-Human and Human-Machine communication. In *Proc. of International Conference of Speech and Language Processing (ICSLP),* Denver, Co: 557-560.

Cerrato, L. (2002) A Study of Verbal Feedback in Italian. In Henrichsen P.,J. (ed) *Proc. of the NORDTALK Symposium on Relations between Utterances*, Copenhagen, Danmark Copenhagen Working papers in LSP: 80-97.

Cerrato, L. (2002). Some characteristics of feedback expressions in Swedish. In *Proc. of Fonetik* Stockholm, Sweden Speech, Music and Hearing Quarterly Progress and Status Report: 41-44.

Cerrato, L. & Ekeklint, S. (2002) Different ways of ending human-machine interactions. *AAMAS Workshop on Embodied Conversational Agents: "Embodied conversational agents - let's specify and evaluate them",* Bologna, Italy.

Cerrato, L. & Skhiri, M. (2002) Quantifying non-verbal communicative behaviour in face-to-face human dialogues. In First Pan-American/Iberian Meeting on Acoustics Cancun. Mexico (only abstract).

Allwood, J. & Cerrato, L. (2003) A study of gestural feedback expressions. In *Proc. of the First Nordic Symposium on Multi-modal Communication.* Copenhagen, Danmark, Paggio, P., Jokinen, K. & Jönsson, A. (Eds.):7-20

Cerrato, L. & D'Imperio, M. (2003) Duration and tonal characteristics of short expressions in Italian. In Solé, M. J., Recasens, D. & Romero, J. (Eds.) *Proc. of the International Conference of Phonetic Sciences (ICPhS)* Barcelona, Spain: 1213-1217.

Cerrato, L. & Paoloni, A. (2003) Utilizzo dei parametric della fonetica acustica nell'identificazione del parlante in ambito forense. In Cosi, P., Magno Caldognetto, E. & Zamboni, A. (Eds.), *Voce, Canto Parlato*, *Studi in Onore di Franco Ferrero*, Unipress: 59-66.

Cerrato, L. & Skhiri, M. (2003) A method for the analysis and measurement of communicative head movements in human dialogues. In *Proc. of Audio-Visual Speech Processing (AVSP) ISCA Tutorial and Research Workshop on Audio-Visual Speech Processing,* St Jorioz, France: 251-256.

Cerrato, L. (2003) On the acoustic, prosodic and gestural characteristics of "m-like" sounds in Swedish. *Gothenburg Papers in Theoretical Linguistics, Feedback in Spoken Interaction-Nordtalk Symposium*: 18-31.

Beskow, J. & Cerrato, L. (2004). Evaluation of the expressivness of a Swedish talking head in the context of human-machine interaction. In *Atti del convegno del Gruppo di Studio sulla Comunicazione Parlata (GSCP)* Padua, Italy (in press).

Beskow, J., Cerrato, L., Cosi, P., Costantini, E., Nordstrand, M., Pianesi, F., Prete, M. & Svanfeldt, G. (2004). Preliminary cross-cultural evaluation of expressiveness in synthetic faces. In André, E., Dybkjaer, L., Minker, W. & Heisterkampf, P. (Eds.), *Proc. of Tutorial and Research Workshop on Affective Dialogue Systems* (*ADS*), Kloster Irsee, Tyskland: 240-243.

Beskow, J., Cerrato, L., Granström, B., House, D., Nordenberg, M., Nordstrand, M. & Svanfeldt, G. (2004). Expressive Animated Agents for Affective Dialogue Systems.. In André, E., Dybkjaer, L., Minker, W. & Heisterkampf, P. (Eds.), *Proc. of Tutorial and Research Workshop on Affective Dialogue Systems* (*ADS*) Kloster Irsee, Tyskland: 301-304.

Beskow, J., Cerrato, L., Granström, B., House, D., Nordstrand, M. & Svanfeldt, G. (2004). The Swedish PF-Star Multi-Modal Corpora. In Martin, J.C. (Ed.), *Proc. of LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multi-modal Input and Output Interfaces*, Lisboa, Portugal: 34-37.

Cerrato, L. (2004). A coding scheme for the annotation of feedback phenomena in conversational speech. In Martin, J.C. (Ed.), *Proc. of LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multi-modal Input and Output Interfaces,* Lisboa: 25-28.

Cerrato, L. (2004). A comparative study of verbal feedback in Italian and Swedish map-task dialogues. In Copenhagen, P. & Hernrichsen, J. (Eds.), *Proc. of the Nordic Symposium on the Comparison of Spoken Languages*, Copenhagen Working Papers in LSP: 99-126.

Cerrato, L. & Ekeklint, S. (2004). Evaluating users reactions to human-like interfaces: Prosodic and paralinguistic features as new evaluation measures for users' satisfaction. In Ruttkay, Zs. & Pelachaud, C. (Eds.) *From Brows to Trust Evaluating Embodied Conversational Agents*. Kluwer's Human-Computer Interaction Series 7: 101-125.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. & Paggio, P. (2005). The MUMIN Annotation Scheme for Feedback, Turn management and Sequencing. In *Gothenburg papers in Theoretical Linguistics 92: Proc. from The Second Nordic Conference on Multi-modal Communication,* Gothenburg University, Sweden: 91-109.

Cerrato, L. (2005). Linguistic functions of head nods. In Allwood, J. & Dorriots, B. (Eds.), *Gothenburg papers in Theoretical Linguistics 92: Proc. from The Second Nordic Conference on Multi-modal Communication*, Gothenburg University, Sweden: 137-152.

Cerrato, L. (2005). The communicative function of "sì" in Italian and "ja" in Swedish: an acoustic analysis. In *Proc. of Fonetik* Göteborg, Sweden: 41-44.

Cerrato, L. & Svanfeldt, G. (2005). A method for the detection of communicative head nods in expressive speech. In Allwood, J., Dorriots, B. & Nicholson, S. (Eds.), *Gothenburg papers in Theoretical Linguistics 92: Proc. from The Second Nordic Conference on Multi-modal Communication,* Gothenburg University, Sweden: 153-165.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. and Paggio, P. (2006) A Coding Scheme for the Annotation of Feedback, Turn management and Sequencing Phenomena. In Martin et al (Eds.) *Proc. of the LREC Workshop on Multi-modal Corpora. From Multi-modal Behaviour to Usable Models.* Genova, Italy: 38-42.

Cerrato, L. & D'Imperio, M. (2006). An Investigation of the Communicative Functions of Short Expressions in Italian and Swedish. In La comunicazione parlata, Liguori, Naples, Italy: 183-203 (in press)

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. & Paggio, P. (2007) "The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena". Submitted to Special issue of the *International Journal of Language Resources and Evaluation:* Multi-modal Corpora for Modelling Human Multi-Modal Behaviour.

Beskow, J. & Cerrato, L (2007) "Synthesis and Evaluation of Expressiveness in a Swedish Talking Head Trained on a Multi-modal Expressive Speech Corpus", submitted to Special issue of the *International Journal of Language Resources and Evaluation*: Multi-modal Corpora for Modelling Human Multi-Modal Behaviour.

Abbreviations

3D	Three-dimensional
ASR	Automatic Speech Recognition
A-V	Audio-Visual
CLIPS	Corpora and Lexicon of Written and Spoken Italian
DAT	Digital Audio Tape
DH	Disfluence
DV	Digital Video
ECA	Embodied Conversational Agent
FAPs	Facial Animation Parameters
FB	Feedback
FBElChA	Feedback Elicit Check Attention
FBElM	Feedback Elicit More information
FBElRA	Feedback Elicit Require Acceptance
FBGiA	Feedback Give Acceptance
FBGiCI	Feedback Give Continuation I go on
FBGiCY	Feedback Give Continuation You go on
FBGiEx	Feedback Give Expressive
FBGiR	Feedback Give non Acceptance (refusal)
GSLC	Gothenburg Spoken Language Corpus
KTH	Kungliga Tekniska Högskolan (Royal Institute of Technology)
H-H	Human to Human
H-M	Human to Machine
IT	Italian
MPEG	Moving Picture Experts Group
MT	Map Task
PCA	Principal Component Analysis
RN	Reply Negative
RP	Reply Positive
SW	Swedish
TMn	Turn Management
TMH	Tal, Musik och Hörsel (Dept. of Speech, Music and Hearing)
WOZ	Wizard of Oz

Contents

1 Introduction and Outline

1.1 Introduction

In human communication, signals transmitted via different modalities are blended to convey meanings. Communication takes place not only via the auditory modality, but also via the visual modality, which includes facial expressions, head and hand movements, gaze direction and body postures.

The advantage of using multiple modalities in interpersonal face-to-face communication lies in the resulting ease and robustness of communication; the use of several complementary modalities improves recognition accuracy and makes communication redundant. For people with a hearing impairment the benefit of multi-modal transmission is very obvious: they use the visual information in support of the audio information they lack. Also in case of communication occurring in particular condition of noise, the support from the visual modality can play an important role in conveying the message even for people with normal hearing.

Human-machine interaction can also benefit from modelling several modalities, since the advantage of multiple modalities in this case is increased usability: the weaknesses of one modality are compensated by the strengths of another. Bringing this multi-modal communication ability to the field of human-machine communication has recently become a big challenge in the design of advanced human-like computer interactive systems. A way to exploit the multi-modal nature of speech communication in human-machine interactions is to endow interactive systems with embodied conversational animated agents (ECAs) able to produce speech and communicative gestures in human-like ways. The assumption is that anthropomorphic systems would allow for more natural human-machine interactions.

While a great deal of research has been carried out to investigate conversational behaviour in human communication, much still needs to be done to get more insight into how to exploit human conversational behaviour to design more advanced systems capable of effective multi-modal interactions with humans.

The contribution of this thesis is two-fold: to give more insight into human conversational behaviour by analysing in particular feedback phenomena, and to suggest a feasible method to provide valuable data to

control facial and head movements related to visual correlates of communicative feedback in synthetic conversational agents.

1.2 The Focus of the Thesis

In this thesis the focus is on communicative behaviour produced in spoken interactions for the purpose of signalling feedback. The feedback phenomena studied are: words, phrases, utterances, facial expressions and head movements that serve the function of managing communicative interaction.

Feedback has been chosen among the several observable human communicative behaviour because it is a pervasive phenomenon in human communication: participants in a conversation continuously give and/or elicit feedback as a way of explicitly exchanging information about the state of communication, for instance to show attention, understanding, misunderstanding, acceptance, non-acceptance and so on, in order to make communication more efficient and robust.

Feedback is a sort of explicit "running commentary to what the current speaker is saying or doing" [Poyatos 2002, p. 241], as a consequence the success or failure of a conversation relies much on feedback and for this reason feedback phenomena can be interpreted as index of conversational fluency. The notion explicitness of information about the state of communication is of fundamental importance to understand the criterion for the identification and annotation of feedback followed in this thesis. If explicitness is not taken into account, the risk is that every contribution in a dialogue can be considered as feedback, since every contribution can indeed be interpreted as a reaction to the previous contribution in terms of implicit feedback.

Communicative feedback phenomena are investigated in this thesis across languages (Italian and Swedish) and across modalities (auditory versus visual), but also in different communicative situations (human-human versus human-machine interactions).

Particular attention in this thesis is paid to the realization of short feedback expressions, such as words and head nods. These are widely produced in the course of spontaneous conversation and seem to carry a variety of semantic-pragmatic functions. For instance a short expression such as "yes" or a head nod can be used as a feedback to signal the willingness to go on in the interaction and/or indicate that the current speaker has understood so far what s/he has been told.

Feedback phenomena are often interwoven with other communicative phenomena, such as turn management and expression of emotions and attitudes, and for this reason some marginal analyses of these phenomena have also been carried out in this thesis.

1.2.1 Terminological Clarification

In this thesis the terminological distinction made between verbal and non-verbal communicative behaviour is based on the modality in which the phenomena are transmitted. Verbal behaviour include the speech phenomena transmitted via the auditory modality (production of words, utterances, prosodic phenomena)[1], and non-verbal behaviour refers to the communicative phenomena transmitted via the visual modality (facial expressions, head movements, gaze direction and other gestures).

Non-verbal behaviour can be produced to serve the following primary functions in human communication according to Argyle [1988]:

- Conveying interpersonal attitudes
- Presenting one's personality to others
- Expressing emotion
- Accompanying speech for the purpose of controlling turn management and feedback.

However none of these communicative functions is limited to non-verbal behaviour alone, in fact we can show emotions and attitudes, present ourselves in a certain way and manage interaction using verbal cues, as well [Knapp & Hall 2002]. As a consequence, whatever definition and classification might be given of non-verbal behaviour, it has to be borne in mind that non-verbal communication should not be studied as an isolated phenomenon, but as an interwoven, inseparable part of the total communication process. In this thesis the term communicative behaviour is used to refer to the broader process of interactive communication [McNeill 2000].

1.3 Aims of the Thesis

Despite much interest in language and speech communication, only in recent times has the relevance of non-verbal behaviour in communication been underlined by studies that have reported on the different communicative functions that non-verbal behaviour can carry out [Argyle & Cook 1976; Bolinger 1989; Kendon 1993; McNeill 1992; 2000].

Recently the interest for non-verbal communicative behaviour has received particular attention also in the field of speech technologies, since it has been demonstrated that the presence of non-verbal signals enhances the naturalness and effectiveness of embodied dialogue systems, as the embodied agents are perceived as more cooperative, human-like and natural in their interactive style when showing non-verbal communicative behaviour [Cassell & Thorrisón 1999, Cassell et al. 1999; Nakano et al. 2003].

[1] This is what in the literature is defined as verbal/vocal.

The aim of this thesis is to give more insight into how humans use verbal and non-verbal behaviour related to feedback and at the same time to suggest a feasible method that can provide valuable data useful to control facial displays related to feedback in synthetic conversational agents.

Facial displays include phenomena such as changes in eyebrow position, expressions of the mouth, movement of the head and eyes [Chovil 1992; Cassell 2000]; in this thesis particular attention is paid to head movements related to feedback.

The idea of reproducing human behaviour in talking heads is based on the assumption that the implementation of facial displays related to communicative phenomena such as feedback, might result in the design of more advanced systems capable of effective multi-modal interactions with humans [Takeuchi & Nagao 1993]

The studies presented in this thesis try to answer the following main questions:

- Is it possible to categorise the semantic-pragmatic function of feedback expressions by using the coding scheme specially designed for this study? Are these categories independent of the modality in which feedback is expressed?
- Is it possible to identify a specific relationship between the semantic-pragmatic function of the feedback expressions and its verbal and non-verbal realization? In particular: can the acoustic characteristics of short feedback expressions, such as duration and pitch contour, and the visual characteristics, such as shape, entity and velocity of the movement, be regarded as cues to the interpretation of the semantic-pragmatic function that feedback conveys in the given context?
- Do the acoustic characteristics of short expressions extracted from their context help in the perceptual identification of their semantic-pragmatic function even out of the context?
- Is it reasonable to interpret the production of non-verbal behaviour that signal feedback and turn management, and the physical signals of emotional attitude of the users in interaction with a multi-modal dialogue system as an index of the fluency and naturalness of the interaction?
- Is it feasible to consider the production of non-verbal behaviour that signal feedback and turn management, and the physical signals of emotional attitude of the users as additional metrics for user satisfaction?

1.4 Materials and Method

The study of verbal and non-verbal human communicative behaviour related to feedback in human-human and human-machine interactions presupposes

sufficiently high quality of the materials under analysis, a specific coding scheme for the annotation of the phenomena under observation and the support of reliable tools for the analysis.

One of the assumptions on which this thesis is based is that understanding how humans behave in human-human communication can give more insight in how to design human-machine interactions that are experienced as natural as possible for the human part. As a consequence a great deal of the materials analysed in this thesis were chosen from existing speech databases in which the degree of spontaneity would mirror the actual human communicative behaviour, this to find regularities in human communicative behaviour that could be reproduced in talking heads.

Finally, for the sake of reproduction in talking heads, it was also necessary to collect data in a lab environment with a novel acquisition set-up that allows capturing the dynamics of facial displays. These data represent a valuable potential source for the training and testing of a data-driven model of non-verbal behaviour in talking heads.

The method of analysis is constant throughout the thesis. Once a feedback expression is identified, in terms of reaction to the previous communicative act, in Allwood's terms [Allwood 2001b] it is coded by means of a specific coding scheme that provides categories for the annotation of feedback expressions according to their direction, type and semantic-pragmatic functions.

The coding scheme is the key to all the investigations carried out in this thesis. It allows coding and analysing verbal and non-verbal feedback expressions and categorising them according to their type, direction and the semantic-pragmatic function they convey in the given context.

The coding scheme provides detailed features to annotate facial displays, in particular head movements related to feedback.

Of course other gestures, besides facial display, can be employed to signal feedback (for instance movements of the shoulders, hands and trunk). However the motivation of putting the focus on facial displays in his thesis is due to the advocated final application of the results of these analyses, that is the possibility to propose models for the reproduction of facial displays in talking heads, which do not have shoulders, trunks and hands.

The appropriateness, feasibility and reliability of the categories designed to code the semantic-pragmatic functions of feedback expressions are tested across languages (Italian and Swedish) and modalities (auditory and visual), moreover the reliability of the coding scheme is tested following the strategy described in Carletta et al. [1997].

The tags used for the annotation help to automatically retrieve several quantitative measures that provide an overall picture of the distribution and typology of feedback expressions. A more detailed picture of the specific functions that feedback expressions can carry out in the given context is

given by the coded explicit semantic-pragmatic function of each identified feedback expression.

The coding scheme has also been implemented in three tools for audio-visual analysis: Anvil [Kipp 2001], Multitool [Allwood et al. 2003] and WaveSurfer [Sjölander & Beskow 2000].

1.5 Thesis Outline

Chapter 1 provides a short introduction to the thesis and its outline.

Chapter 2 gives a historical perspective and reviews the state of the art of studies about feedback phenomena in human-human and human-machine interactions.

Chapter 3 gives an overview of the materials used for the analyses.

Chapter 4 illustrates the method of analysis. In particular the coding scheme developed for the purpose of coding phenomena related to feedback is presented. In this chapter a short description of the tools used to carry out the analyses of the materials is also provided.

Chapter 5 reports the investigation of verbal feedback phenomena carried out on dialogues recorded with the map-task technique in Swedish and Italian, and includes the results of the reliability test run to evaluate the semantic-pragmatic functions proposed in the coding scheme used to annotate verbal feedback phenomena.

Chapter 6 reports the results of the analysis of verbal and non-verbal feedback phenomena carried out on dialogues selected from a Swedish corpus of real spontaneous human-human interactions recorded in a travel agency.

Chapter 7 reports the results of the investigation of verbal and non-verbal feedback phenomena in human-machine interactions between users and an experimental Swedish dialogue system.

Chapter 8 deals with the collection and analysis of tri-dimensional data acquired with a motion capture system. Three studies are presented in this chapter: the first and second report on detailed investigation of head nods related to feedback. The third one proposes a method for automatic detection of head movements, in particular head nods.

Chapter 9 discusses the possibility to use the obtained results in a technological application and illustrates some evaluation paradigms that could be used to assess the appropriateness and effectiveness of the eventual models of non-verbal behaviour reproduced in embodied conversational agents.

In **Chapter 10** the thesis is summarised and conclusions are drawn.

2 Historical Background

2.1 Historical Perspective and Terminology

The starting point for the analysis of linguistic communicative feedback is the notion of feedback used in cybernetics and control engineering as early as in 1948 by Wiener [1948]. He used the term feedback to refer to the ability of a machine to use the results of its own performance as self-regulating information and so to adjust itself as part of an ongoing process. Thanks to Wiener the concept of feedback penetrated almost every aspect of technical culture, and has even been applied to human communication in a broad holistic sense by several researchers [Bateson 1972].

Although the term feedback is relatively new, and its application in linguistics is even more recent, the concepts denoted by the term are not as novel, in fact the rhetorical tradition of Plato, Aristotle, Cicero and Quintilian for instance already provides many reflections on how to elicit attention, understanding and emotional reaction in public speaking.

What we call feedback expressions have been referred to by several different names, among which: "interjections" [Beckman 1968; Poggi 1981] or "discourse markers" [Schiffrin 1994; Bazzanella 1994], "accompaniment signals" [Kendon, 1967], "back-channels" [Yngve 1970; Duncan 1972; Clark & Schaefer 1989] "responsive tokens" [Fries 1952; Gardner 2001; Caspers 2003].

In the grammatical tradition of the West, feedback phenomena have mostly been included under the grammatical category of interjections. The perspective of the grammatical tradition of interjection is however, the perspective of rational non-interactive written discourse [Allwood 1993]. To move from this perspective we have to jump in time up to 1950's, when, with the advent of new technology (such as audio and video-recording devices) it started to be possible to investigate spoken language. One of the first authors who noticed and described some of the interactive phenomena that we nowadays call feedback was Fries [1952] who analysed a corpus of telephone conversations in which he distinguished a set of "listener responses". These are unobtrusive response tokens such as *yeah*, *ok*ay, and m-like words. More recently, a thorough study of the characteristics and functions that response tokens carry out in interactive talk in English was carried out by Gardner [2001].

At the beginning of the seventies, it was Yngve [1970] who first coined the term back-channel to describe these tokens. He defined back-channel as the channel "over which the person who has the turn receives short messages such as yes and uh-huh without relinquishing the turn" (p. 568).

Dittmann [1972] used the term “listener responses” in the same manner that Yngve used “back-channel”, describing them as "specific signals that the listener is paying attention to the speaker, is keeping up with him, or that he has understood what was just said" (p. 405).

Although the notion of back-channel is well spread and much used in the research community, it is often still debated which types of utterances can be considered back-channel activity. The short expressions like *mm, yeah, right* (which are common in English and have similar counterparts in many languages) clearly qualify because they add a great deal to the quality and success of the interaction without really adding meaning to the conversation, at least when they are not produced with the intention of requesting the floor. In fact, back-channels are not viewed as speaker turns, but as “optional” utterances occurring during the turn of another speaker [Yngve 1970; Duncan & Fiske 1977; Koiso et al. 1998; Ward & Tsukahara 2000].

However, Yngve, aside from utterances that are primarily displays of recipiency and/or listenership, includes in the group of back-channels also questions such as, "You've started writing it, then,…your dissertation?" and short comments such as, "Oh, I can believe it".

Duncan and Fiske [1977] categorized back-channels into five types: m-like words, sentence completion, requests for clarification, brief restatement, and head nods and shakes. They underline that the purpose of back-channels is not actually to claim a turn, but to provide the speaker with needed feedback. It is important to note that Duncan and Fiske include both non-verbal and verbal responses in their types.

In addition to head nods, back-channels can be comprised of gestures and facial expressions [Krauss et al. 1977]. Smiles are a common source of back-channel communication. It has been found that smiling often perpetuates feelings or thoughts of being understood [Brunner 1979]. In an earlier study, Dittmann and Llewellyn [1968] also found that when a short smile was produced simultaneously with a head nod it tended to signal attention.

Schlegoff [1982], in his study of back-channels pointed out that speakers seem to wait for their interlocutor to produce a continuer back-channel which might indicate that back-channels are not to be considered as optional, but rather as compulsory.

In this thesis the term back-channel is not used. All the responsive phenomena (verbal and non-verbal) are grouped under the term feedback. This means that what is usually termed as back-channel is considered here

as feedback signal given or elicited while the other speaker is uttering his/her contribution.

The notion of contribution is in this thesis preferred since the notion of turn or utterance can be sometimes misleading for feedback analysis [Allwood 2001a]. A great deal of short feedback expressions are in fact produced as an insertion in the turn of the current speaker and not as a real turn of their own.

Schiffrin [1987] drew attention to the dynamicity of conversation, in which both speakers and listeners must be constantly alert to give and pick up a number of subtle signals that refer to changes in the conversational topic as well as to understanding and interest by the participants.

Generally speaking it is possible to distinguish between two main different approaches towards feedback analysis, one which looks at feedback processes in the broader process of "grounding" [Clark & Schaefer 1989, Clark & Brennan 1991], and one in which feedback is considered as a particular kind of speech act which aims at giving information about how the communication is proceeding [Allwood 1988].

The concept of two-way cooperative communication between interlocutors is at the basis of the idea of "grounding". In this approach discourse is described in terms of a joint activity in which participants in a conversation are committed to achieving maximally effective conversation [Grice 1975]. In Clark and Schaefer's approach the basic concept is that in order to successfully communicate, it is necessary for the interlocutors to share, beyond some basic conversational principles, some kind of common ground. Grounding can be defined as the management of the knowledge in the dialogue, keeping track of the changes in the common ground. In this framework back-channels are used to signal that the information has been integrated into the common ground shared by speaker and listener and that the listener understands that the speaker has not finished yet.

In the other approach, feedback is considered as one of the most important cohesion devices in human conversation and is analysed as a particular kind of speech act, aiming at signalling the failure or success of the listener's processing of a speaker's utterance. This last approach is best represented by Allwood [1988] and it is the one followed in this thesis,

Whether researchers speak of grounding, negotiation and whether they look at feedback, back-channels or responsive tokens, it is clear that beyond the differences in their formulation, they all seem to agree on the fact that in conversations some strategies are used as a "cooperative" way of exchanging information about how communication is proceeding.

2.2 Defining Feedback

The interpretation of feedback followed in this thesis is inspired by Allwood [1993], who considers feedback as a kind of speech act, and

defines it as: “linguistic mechanisms, which ensure that a set of basic requirements for communication, such as possibilities for continued contact, for mutual perception and understanding can be met” [cf. p.1]. In other words, feedback enables the participants in conversation to exchange information about “four basic communicative functions”. These functions are: “contact, perception, understanding and attitudinal reactions”. Thus an expression is considered feedback if its primary function serves one or more of the following purposes:

- show continuation of contact: when the interlocutor wishes to show that s/he is willing and able to continue the interaction;
- show perception: when the interlocutors show awareness and discernment of expression of the message;
- show understanding: when the interlocutors show that they have understood the message;
- show attitudinal reactions: by giving and eliciting feedback, interlocutors can show behavioral and attitudinal reactions towards the meaning conveyed, both speaker and listener can show emotions and attitudes, for instance they can agree enthusiastically, or signal lack of acceptance and disappointment.

The four basic communicative functions are related to basic requirements of human communication, in fact in order to obtain a successful communication it is necessary first of all that two participants establish a contact with each other. Once the contact is established it is possible to produce a message, which should be perceived by a receiver, who must be able and willing to understand it. It can be helpful for the sender to give and get attitudinal and behavioural reactions as indicators of how well he/she managed to send the intended message [Allwood, Nivre & Ahlsén 1992].

Feedback, together with turn management and sequencing (i.e. the structuring of a dialogue into sequences) are included in Allwood’s terms [Allwood 2001b] in the concept of “Interactive Communication Management”, which refer to all communicative phenomena dealing with the management of dialogue interaction.

The focus of this thesis is on feedback phenomena; however turn management is also taken into account since it is often interwoven with feedback.

The categorization proposed in Allwood, Nivre & Ahlsén [1992] has been modified in this thesis in order to boost the notion of explicit semantic-pragmatic function of feedback and include an indication of whether the speaker who gives feedback also explicitly aims at signalling the intention to gain the floor or not.

2.3 Feedback Types

Under normal circumstances, in face-to-face human-human communication, and even in human-machine communication, feedback involves the use of multi-modal expressions, which means that it can be expressed by means of verbal, and non-verbal expressions.

Several studies have been carried out to determine the basic ways of expressing linguistic communicative feedback during a conversation [Allwood, Nivre & Ahlsén 1992; Cerrato 2002b; Campbell 2003; Muller & Prevót 2003]; even if there appear to be cultural and speaker-dependent variables, there seems to be agreement on the fact that feedback can be expressed both by means of verbal and non-verbal kinesic actions. Verbal expressions involve not only the production of specific spoken words and definitive use of language, but also several prosodic phenomena, such as duration, pitch, tempo and intensity variations, which signal stress and emphasis. Non-verbal expressions comprise facial expressions, eye gaze, body and hand gestures and body posture.

Some general ways of expressing feedback by verbal means are:

- short words like: *yes, no, ah, ah ah, mm, mhm* together with some prosodic and phonological phenomena (like vocalic lengthening);
- short utterances such as: *I understand, oh really* and so on;
- repetition, either the last word uttered by the interlocutor, or of the core words of the last sentence with other types of reformulation of the meaning of the received message;
- anticipation or completion of the speaker contribution;
- short questions or request for clarifications.

Some general ways of expressing feedback by non-verbal means are:

- head movements, eyebrow rising, and/or specific hand and body movements.

2.3.1 Verbal Feedback Expressions

As indicated above, verbal feedback expressions can consist of words such as: *yes, ok*, *mm*, short utterances as: *I understand*, *I follow* and even longer expressions consisting of repetition or reformulation of what the speaker has just said.

The most common feedback words in Swedish, according to a list of frequency retrieved from 1.4 millions words automatically tagged in the Gothenburg Spoken Language Corpus, are shown in table 2.1 [Allwood et al. 2000]. *Ja* and its variants and m-like words are the most common feedback words in the whole corpus.

Table 2.1 The most common feedback words in the GSL Corpus [Allwood et al. 2000]

Swedish feedback words	**Number of occurrences in GSLC**
ja	37154
m	13405
nä	8651
va	6798
nej	2546
jo	2428
jaha,aha	2251
okej	1446
just	1404
visst	786
kanske	777
ehm	725
jaså	570
precis	553
nähä	339
javisst	269
nja	240
bra	221
jamen	151
nejmen	136
ah	135
ja-ja	132
absolut	108

However what expression is used to give or elicit feedback is not the whole story, also how a feedback expression is uttered is important, for this reason prosodic and phonological characteristics of feedback phenomena, in particular of echoic responses, have been investigated [Katagiri, Sugito & Nagano-Madsen 1999; Ward & Tsukahara 2000; Shimojima et al. 2002; Campbell 2004].

It is quite uncontroversial that acoustic cues can be used for the purpose of marking information structure at the discourse level [Ferrer, Shriberg & Stolke 2002]. Phonetic correlates for different functions of short expressions signalling feedback have been found for a variety of languages: English [Hirshberg & Nakatani 1996], Japanese [Ward & Tsukahara 2000], Dutch [Caspers 2000], Swedish, and Italian [Cerrato 2002b; 2002c].

Stubbe [1998] proposes a continuum of interactive feedback expressions, which ranges from low involvement and neutral affect (for instance in the

case of minimal responses such as mm, yeah), to high involvement and positive affect. The neutral minimal responses are "prosodically and lexically unmarked and are characterised by low mid pitch, fairly level intonation and relative low volume" [cf. p. 266]. Following this interesting proposal, in this thesis it has been hypothesized that short verbal feedback expressions such as *yes, mm, ah*, with continuation function, show shorter duration and lower energy than other more complex verbal expressions that have more complex feedback functions.

2.3.2 Non-Verbal Feedback Expressions

The studies of kinesic actions (head movement, eyebrow rising, and hand movements) related to feedback has been of interest to several researchers. Already Darwin [1872] at the end of the eighteenth century and other researchers one century later on [Eibl-Eibesfeldt 1970; Morris 1994] noticed that the affirmative head-nod is as a nearly universal indication of accord, agreement, and understanding.

Yngve [1970] and Duncan [1972] included head nods as a typical example of back-channels in their descriptions.

Maynard [1987] analysed head nods in dyadic conversation among Japanese speakers and noticed that the most common functions carried out by the many head nods produced during these conversation was that of back-channelling. Head nods have been studied more recently in face-to-face communication [Cassell et al. 1999] and it has been noticed that interlocutors use them to signal feedback, turn-taking and as indication of chunk processing.

Mc Clave [2000] looked at the several functions of head movements in dyad conversation between speakers of American English and noticed that most of the head nods produced by the listeners were responses to the speaker's non-verbal request for feedback. These requests were produced as "up-and-down nods", and listeners were able to recognize and respond to such requests in a fraction of a second.

An attempt to quantify the extent of head movements was made by Birdwhistell [1970], who assumed that all movements of the body, including head nods, are directly linked to linguistic structure and proposed a hierarchical system of units of movement in which lower-level units (kines) combined to form higher-level units (kinemes).

The results obtained in this thesis in chapters 6 and 7 show that head nods are the most frequent non-verbal behaviour produced to signal feedback in the materials investigated. For this reason in particular head nods have been further analysed with the aim of providing data that could be used as sources for their reproduction in synthetic agents.

2.4 Feedback across Languages

Cross-linguistic research has shown that speakers from different cultures exhibit different feedback behaviour. In some cultures feedback production is quite frequent, as for instance in Japanese compared to English [Ward & Tsukahara 2000]. In an intercultural communicative exchange the difference in the frequency of feedback production might have the effect that the interlocutors coming from different cultures and having different communicative feedback behaviour might have different expectation about feedback production and might interpret the same feedback expressions in different ways.

This might lead to misunderstanding and eventually to communication breakdown. For instance Berry [1994] interviewed Spanish and English speakers and found out that the Spanish speakers considered comments and questions that overlap with the speaker to be a positive part of conversation because they show that people are paying attention, having fun, or responding emotionally to the other speaker; whereas the English speakers commented that when two speakers talk at the same time, it means that they are not listening to each other. Moreover, although the English speakers consider back-channel comments such as “mmm” and “yeah” to be cooperative, the Spanish speakers generally agreed that a constant “uh-huh, uh-huh" makes a listener sound uninterested and pressures the speaker to hurry up and finish.

Moreover Berry found that although both the English and Spanish speakers who participated in her study used a variety of back-channels, the Spanish speakers tended to use longer and more explicit comments in their back-channel contributions ("Ay, sí, es verdad, sí" Oh, yes, that is so true), and they were more likely to repeat or rephrase what the speaker was saying as a way of showing understanding.

Similar results are shown in this thesis in chapter 5 as concerning the difference in feedback expressions in Italian and Swedish: in Italian longer feedback expressions, consisting of repetitions and reformulations of part or the entire previous utterance and anticipation of the end of the current utterance, are more common than in Swedish.

As concerning non-verbal feedback, Maynard [1986] observed that in general Americans nod less frequently than Japanese during conversations.

General observation of non-verbal feedback behaviour have also pointed out that the most common ways of expressing non-verbal feedback behaviour is by means of head movements, eyebrow rising, smiles and and/or by making specific hand movements [Knapp & Hall 2002].

2.5 Feedback in Human-Machine Interactions

In the past fifteen years, beside the descriptive studies of feedback aiming at gaining insight in the structures of human conversation, a series of other

studies, which might be described as applicative, have been carried out in order to look for regularities in human-human conversational behaviour that could be exploited in the attempt to improve the performance of spoken dialogue systems.

Spoken dialogue systems are increasingly becoming part of our everyday life, and the designers are striving to develop user interfaces that integrate a larger number of human discourse features. Perhaps because communication is so defining of humanness and human interaction (only humans communicate using language and carry out conversations with one another) the metaphor of face-to-face conversation has been applied to human computer interface design for quite a long time, the first attempt being in the late seventies [Nickerson 1976].

Even if there exist systems which allow users to accomplish useful tasks, current spoken dialogue systems, even the multi-modal ones, are still not error-free, mainly because of the deficiencies of the automatic speech recognition (ASR) engine [Lippmann 1997] and of the lack of appropriate use of some of the most important human communicative behaviour, such as feedback and turn management. As a consequence it is assumed that an appropriate use of feedback from the system to the user, as well as the possibility for the system to recognize the visual feedback provided by the user, could be optimal ways of not only compensating for the limitations of the speech recognizer, but also rendering the interaction more natural, thus improving human-machine interactions [Morency et al. 2005].

2.5.1 Verbal Feedback in Human-Machine Interactions

Prosodic and phonological characteristics of feedback phenomena have became a hot topic in the field of human-machine interaction studies, since it is believed that prosodic and phonological characteristics of feedback expressions can be the key to finding some constant behaviour that could be exploited in the implementation of spoken dialogue systems.

For instance [Swerts et al. 1998; Shimojima et al. 2002] have looked at acknowledgement and repair-request in Japanese dialogues, and have found out that the function of these parts of speech, which are usually repeated in spoken Japanese, is likely to be reflected in their prosodic characteristics (i.e. lower F0 mean, higher articulation rate).

Similarly, investigations of the Dutch word “nee”, which in the analysed corpus was either used as “go on” or a “go back” signal, showed close connections between the function of the word “nee” and its prosodic characteristics [Krahmer et al. 2002]. In other words it resulted that speakers use prosodically marked features when there is a communication problem, for instance lengthening phenomena, longer preceding pauses and more high-pitched accents in comparison to responses intended to signal to the interlocutor to go on. These prosodic characteristics could be exploited in a

spoken dialogue system as signals of non-comprehension and miscomprehension.

Since the results of the analysis of feedback phenomena in human-human communication had shown that human behaviour exhibits regularities, it started to be appealing to think that some of these regularities could be modelled and implemented in spoken dialogue systems.

When looking at the scheme provided by Clark and Schaefer's model of grounding [Clark & Schaefer 1989] it appears that the rules they propose resemble the type of grammatical building blocks and rules that computational linguists require for their implementations. In fact [Traum 1994] and Brennan and Hulteen [1995] provide an example of the application of this model in the design of human-computer interface, even though it is at a theoretical level only. Using the concept of grounding communication, they propose a collaborative model of feedback with a speech interface, which provides flexible feedback using those human conversational strategies that users bring with them into human-machine interactions.

They support the idea that feedback is important for coordinating the user and systems' knowledge states in a dialogue system and for facilitating problem solving.

However exporting a theory to a new field often presents some problems. In its original use the model of grounding was proposed to describe human-human dialogue and intended to passively describe some aspects of conversation, while in the context of the interface it is used as a practical theory in the active design of communication, and this causes some limitations that do not allow to completely formalise all the different aspects of human behaviour. As a consequence in most of the current spoken dialogue systems, grounding is often reduced to verification of the system's recognition of user utterances.

Current systems usually apply two strategies to give feedback:

- explicit, when they explicitly repeat the request of the user to get a confirmation, for instance if the user asks: "I would like to book a train ticket to Stockholm" and the system, to confirm the understating of this request, says "You would like to book a train ticket to Stockholm";
- implicit, when the system replies with another question which implicitly presupposes the understanding of the previous request of the user, for instance if the system says "at what time would you like to leave?".

In the framework of virtual conversational systems, feedback strategies are generally grouped in two main classes: positive and negative. Positive feedback is given when the user wishes to show that s/he understands and/or agrees with what the system says, while negative feedback is produced

when the user wishes to signal some problems of perception, understanding and/or disagreement.

Typically, conventional spoken dialogue systems wait until the end of the user turn to provide feedback. This feedback is usually quite elaborated, for instance under the form of verification questions that repeat the request of the user [San-Segundo et al. 2001]. The user is obliged to answer the question before being able to go on in the interaction, and this makes the interaction less efficient. In order to make human-machine interaction more efficient, some attempts have been made to provide dialogue systems with grounding strategies that are determined by the recognition score of the users' utterances [Larsson 2002]. This means that if the system has understood the utterance of the user, it does not need to produce an explicit feedback under the form of a verification question, but can provide feedback in an implicit way. Implicit feedback is thought to make the interaction proceed in a more efficient way, since the user does not need to give explicit answers to the verifications questions asked by the system.

In human-human communication, while the speaker is uttering his/her contribution, the other speaker can, and in fact does, produce short feedback to show that s/he is listening, following, understanding or not, accepting the information or not and so on. For this reason it could be advisable that the systems should also be able to produce short feedback expressions, such as "yes", "ah", or m-like words, even during the users' contribution, in order to signal continuation of contact, or in other words to show that the system is listening, but has not processed the message yet. This feedback production might enhance human-machine interaction by making it more effective and natural.

2.5.2 Non-Verbal Feedback in Human-Machine Interactions

Recently, technological and scientific development has favoured the advent of Embodied Conversational Agents (ECAs). These are animate anthropomorphic interface agents, able to engage in real-time multi-modal user interactions, by using speech and even non-verbal behaviour, to emulate the experience of human face-to-face interactions [Cassell 2000].

Since it has been shown that when speech is presented together with communicative non-verbal behaviour, it may result in a more robust, more natural and more efficient communication [Thorrisón & Cassell 1996], one of the major challenges in human-machine interaction has become that of equipping the embodied agents with the ability to produce appropriate non-verbal communicative behaviour and also perform visual communicative behaviour recognition in the same way people do [Morency & Darrel 2006]. Enabling this form of interaction in human-machine interfaces requires both advances in the understanding of visual correlates of feedback as naturally produced by humans when interacting with dialogue

systems and the development of efficient and robust algorithms to recognize these visual correlates.

The visual information carried out by some facial expressions and head movements, in particular by head nods in spoken communicative interactions, is without doubt extremely important. Facial expressions can carry out several communicative functions, such as showing attention, interest, disinterest [Harrison 1974], exchange relevant information about the state of communication [Allwood & Cerrato 2003], signalling focus and emphasis, and so on.

Thorrisón [1997] carried out some user testing of embodied humans, such as agents that exhibited face-to-face conversational behaviour, and he showed that the presence of non-verbal feedback behaviour increases believability and effectiveness in the interaction.

Analyses of human-machine interaction have shown that users of multi-modal dialogue systems produced quite a high number of verbal feedback expressions, despite the fact that the virtual agents of the systems never explicitly elicited nor gave feedback during the interactions [Bell & Gustafson 2000].

Similarly Sidner et al. [2004] carried out an experiment with people interacting with a humanoid robot and found that more than 50% of the subjects tended to naturally nod at the robot conversational contributions, even if the robot could not interpret head nods.

Rajan et al. [2001] report on an agent who is able to exhibit a variety of head nodding and head shaking behaviour, that can either accompany vocal feedback expressions or function as non-verbal feedback expressions. For instance, rapid head nods indicate agreement and therefore positive feedback, while slower, smoother head movements (either nods or shakes) co-occur with neutral or negative verbal feedback. Moreover head nods are used to provide back-channel feedback in order to smooth the progress of conversation.

In an experiment conducted by using a synthetic talking head inserted in an interactive situation in a simple travel agency scenario, it was investigated which parameters led external observers to judge the feedback produced by the talking head as affirmative or negative [Granström, House & Swerts 2002]. The results show that the parameters which had the most influence on subjects' judgements were, in rank order: smile, pitch contour, eyebrows and head movements. The conclusion of the study is that subjects are sensitive to both acoustic and visual parameters when they have to judge feedback as positive or negative.

Until lately, existing implementations of communicative (non-verbal) signals in talking heads or embodied conversational agents have often been based on prototypical descriptions of human-human communication found in psychology literature or on observations conducted in a non-systematic way, which means that reproduction of the behaviour in the talking head is

based on intuition rather than on observation. These implementations are unable to display the degree of variability and dynamics exhibited in human facial expressions in real communicative situations. Rather, they tend to appear as stereotypical and predictable.

One interesting first attempt to produce a model of non-verbal behaviour in embodied conversational agents, based on empirical data, has been carried out by Nakano et al. [2003]. They first carried out an investigation of eye gaze, attentional focus and head nods related to the process of grounding in human-human direction-giving tasks and then produced a model of the observed behaviour in an embodied conversational agent, able to use both verbal and non-verbal grounding acts to update the dialogue state.

This procedure is also followed in this thesis: verbal and non-verbal behaviour related to feedback is studied first on empirical data with the aim of finding regular behaviour that could be implemented in talking heads. However this thesis goes a step further and proposes that in order to emulate the degree of variability found in human-human facial displays, it is necessary to acquire dynamic data with motion capture systems, data which can be used to control facial displays in synthetic talking heads.

2.6 Feedback and other Communicative Phenomena

Feedback phenomena are often interwoven with other communicative phenomena, mainly with turn-management signals and expression of emotions and attitudes, and for this reason some marginal analyses of these phenomena have also been carried out in this thesis.

2.6.1 Turn Management and Feedback

With the term "turn-taking" [Goffman 1955] is indicated a kind of system that has the purpose of managing the flow of interaction, minimizing overlapping speech and pauses [Yngve 1970; Sacks, Schegloff & Jefferson 1974; Goodwin 1981].

Duncan [1972, page 283 and 284] suggested that the turn-taking mechanism is "mediated through signals composed of clear-cut behavioural cues, perceived as discrete"; moreover he pointed out that "the turn-taking signals are used and responded to according to rules".

However these phenomena often co-occur or are interwoven with feedback phenomena, and as a consequence it is difficult to describe and analyse them in a discrete way. For instance short verbal feedback expressions such as *sí* in Italian produced to show continuation, can be produced with a rising intonation and vocal lengthening, which is a typical turn-taking signal. Non-verbal feedback expressions, such as head nods

produced to give acceptance, can co-occur with a mutual gaze that signals the intention of not wanting to take the floor. According to Duncan [1972], in conversation turn-yielding cues, back-channel cues, and turn-maintaining cues are used. Turn maintaining cues are produced when the speaker does not wish to yield the turn. When the current speaker wishes to yield the turn s/he might produce turn-yielding cues to let the listener/s know that s/he has finished talking and that someone else may speak. At this point the listener may either take the turn as a response to the turn-yielding cue produced by the speaker, or respond by producing a back-channel, or even remain silent.

In Duncan's proposal, back-channel cues are considered as an alternative to turn-taking; this because in Duncan's perspective back-channels are coherently not viewed as speaker turns [Duncan 1974, Duncan & Fiske, 1977], but as optional utterances that occur during the turn of another speaker. Nevertheless, considering back-channels as optional is quite reductive, given the fact that they are so frequently produced in human communication and that participants in a conversation even expect to receive back-channels.

One way of solving this descriptive confusion between turn-management signals and back-channel or rather feedback production is to consider these two phenomena as non-mutually exclusive. This means that feedback expressions and turn-management signals can co-occur and in that case it is impossible to separate them. As a consequence their analysis and identification has to be performed by means of categories that consider both feedback and turn-management functions.

This is what has been done in the analyses carried out in this thesis. The semantic-pragmatic function of the expressions identified as feedback continuation have been further categorised in two sub-categories which include an indication of whether the speaker who gives feedback intends to get the floor or not.

Researchers in the field of human-machine interfaces, aware of the important role that turn-taking signals can play in communicative exchanges, have attempted to integrate some of them in the design and development of dialog systems.

Turn-management cues are multi-modal in nature; they include both expressions transmitted via the auditory channel (speech signals) and expressions transmitted via the visual channel (facial displays, hand and arm gestures, body postures).

Recently several attempts to reproduce non-verbal turn-management signals have been performed. For instance a series of intuitively designed and hand-tailored turn-taking gestures, consisting of eye-gaze, head tilts and eyebrow rise, were implemented in the AdApt dialogue system [Edlund & Nordstrand 2002]. The purpose of this implementation was to evaluate whether the presence of the turn-taking gestures would be noticed by the users, which was the case, even if the results of the evaluation test could not

clearly show that the presence of the turn-taking gestures, compared to their absence, resulted in a more efficient dialogue between the users and the system.

Thorrissón [2002] provides a good example of a computational turn-taking model in a humanoid agent. His model (YTTM-Ymir turn-taking model) includes multi-modal cues such as gestures and eye-gaze. The prototype system in which he implemented the computational model of turn-taking was able to perceive the user's behaviour and respond with real-time animation and speech output. He tested the system with human users and he found out that, when comparing the system with and without the turn-taking mechanism, the subjective scores for the system's language understanding and language expressions were significantly higher if the turn-taking system was implemented, and these scores were close to those obtained for human-human interactions.

Even if these represent good attempts to reproduce turn-taking behaviour in humanoid agents, as for feedback signals, the optimal analysis and reproduction of turn-management signals should be performed using dynamic data acquired with motion capture systems.

2.6.2 Emotions and Feedback

Participants in a conversation might colour their feedback with an emotional and attitudinal reaction. For instance they can agree showing enthusiasm, they can disagree showing disappointment, they can signal understanding with disinterest, or surprise and so on. This can be done by means of a subtle combination of features at the verbal and non-verbal level [Wallbott & Schrerer 1986]

A number of studies have tried to determine which acoustic cues signal the various emotions in the voice [Murray & Arnott 1993; Cowie et al. 2001; Laukka 2004]. The results of several experiments have suggested that emotions are signalled by prosody. Subjects can recognise the emotive content in a speech sample, also when all word meaning is filtered out [Scherer 1981; Banse & Scherer 1996; Mozziconacci 1988; Pereira 2000].

With the advent of modern speech technologies, in particular speech synthesis and automatic speaker recognition ASR, the studies of emotions have been driven by the new challenges of trying to add emotional effect to synthesized speech and to enable automatic speech recognizers to access and interpret emotive information in speech. The vision is the construction of sophisticated automatic spoken language systems, able to understand users' emotions and needs, and respond accordingly [Batliner et al. 2000; Picard 2000; Höök 2002].

Attempts to add emotional effect to synthesized speech have been carried out in the past 15 years; several prototypes and even operational systems

have been built based on different synthesis techniques [for an overview see Schröder 2001].

In the expression of emotion even the face plays an important, if not primary role in conveying emotions [Ekman 1979, Surakka & Hietanen 1998, Beskow & Cerrato 2004].

Recent studies have been carried out to show in what way expressiveness and emotions affect our facial displays, e.g. how we raise our eyebrows, move our eyes or blink, or nod and turn our head [Argyle & Cook 1976, Ekman 1993], and that even speech articulation is affected by expressiveness [Nordstrand et al. 2004; Magno Caldognetto et al. 2003; 2004].

These results lead to several attempts to reproduce the visual correlates of expressiveness and emotions in current talking heads [Pelachaud, Badler & Steedman 1996; Lundberg & Beskow 1999; De Carlo et al. 2002; Massaro et al. 2005; Cosi, Fusaro & Tisato 2003].

The study of facial movements related to emotions was rather impressionistic until Ekman and Friesen [1978] developed a method for measuring and describing facial behaviors based on muscle movement, a method called facial action coding system (FACS). By studying the faces of some subjects who had learned to control specific muscles, they identified the specific changes that occurred with muscular contractions and how best to differentiate one from another. They associated the appearance changes with the action of muscles that produced them by studying anatomy, reproducing the appearances, and palpating their faces. Their goal was to create a reliable means for skilled human scorers to determine the category or categories in which to fit each facial behaviour.

The FACS allows emotion researchers to describe objectively what movements have occurred on the face and also to categorise a face showing a given emotion based on extensive data relating those movements to other criteria, mainly observers' judgements of facial expressions; trained observers can in fact identify which muscles are moved. Even if accurate, the anatomical description of the face used in the FACS is quite time consuming both to learn and to use. As a consequence some alternative systems have been proposed.

One approach that eliminates human judgement is based on the fact that different emotions produce distinctive facial movements even when the movements are so slight that are impossible to be noted with the naked eye, "micromomentary facial expressions" [Haggard & Isaacs 1966], but can be registered by special recording systems, such as electromiografic responses (EMG) [Blairy, Herrera & Hess 1999; Lundqvist 1995] or optical motion capture system [Beskow et al. 2004b].

The expressions of emotions considered in this thesis are those related to the expression of communicative signals, in particular those that might signal feedback with some attitudinal reaction towards the interaction.

It is assumed that if these signals are correctly interpreted by the interlocutor, human or embodied agent, they could give useful information on how the interaction is proceeding. Both verbal and non-verbal cues related to feedback, turn management and expression of emotion could be used as “on-line" help for the system itself to evaluate how the interaction is going. This could help the system to consequently adapt its communicative strategy in order to make the interaction proceed in a smoother way.

2.7 Conclusions

The overview of previous and current research about conversational behaviour related to feedback presented in this chapter, though being not exhaustive in its intent, offers a picture of how spread the interest for this topic is.

However much research remains still to be done on verbal and non-verbal feedback behaviour, in particular to better understand the exquisitely complex set of their realization.

This thesis aims, therefore, at complementing our knowledge by presenting the results of systematic analysis of verbal and non-verbal feedback phenomena both in human-human and human-machine interactions.

Before describing the investigations in details, in the following chapter an overview of all the materials used to carry out the analysis in this thesis is presented.

3 Materials

3.1 Introduction

This chapter offers an overview of the materials used to carry out the investigations reported in this thesis. More details about the materials will be given in each of the specific chapters dealing with the studies.

The materials analysed in this thesis can be, for the sake of description, grouped in two main blocks:

- Audio recordings of human-human and human-machine interactions,
- audio-visual recordings of human-human and human-machine interactions.

Specifically the audio-recorded materials consisted of:

a) Digitalized audio recordings of four map-task dialogues, two in Italian and two in Swedish.
b) Digitalized audio recording of four interactions with the Swedish experimental dialogue system AdApt.

The audio-visual materials consisted of:

a) Audio-visual recordings of four spontaneous human-human dialogues recorded in a travel agency in Gothenburg.
b) Audio-visual recordings of six human-machine interactions with the Swedish experimental dialogue system AdApt.
c) Audio-visual and 3D data, consisting of prompted sentences and ten dialogues, recorded with an opto-electronic system.

The map-task dialogues were available both in Swedish and in Italian, while the rest of the materials analysed in this thesis are in Swedish and come from different sources. Some were selected from existing databases of human-human and human-machine interactions and were chosen considering that their degree of spontaneity would mirror the actual human communicative behaviour, in order to find regularities in human communicative behaviour that could be reproduced in talking heads.

Since the final goal of this thesis was not only to gain more insight in human verbal and non-verbal communicative behaviour related to feedback, but also to provide data that could be used to control the talking heads developed in our department at KTH, it was decided to systematically acquire and analyse more controlled high-precision materials.

An ideal plan would have been to retrieve available materials acquired in similar conditions both in Italian and Swedish, to allow for cross-linguistic and cross-modal investigations in different communicative situations. Apart from the map-task dialogues, it was not feasible to obtain materials acquired in similar circumstances in both languages, which were suitable for comparative analyses[2].

3.2 Audio-Recorded Materials

Audio recordings of human-human dialogues and human-machine interactions were used to carry out analysis of verbal feedback expressions. Human-human dialogues consist of map-task dialogues in Swedish and in Italian, human-machine interaction include interactions with the Swedish dialogue system AdApt.

3.2.1 Map-Task Dialogues

The map-task technique has become a sort of standard for the collection of task-oriented natural dialogues in a controlled situation [Anderson et al. 1991]. The dialogues do not have a script, but they are elicited according to a well defined task scheme, in which the two dialogue participants have a specific role, which is defined *a priori*: one is designated as "instruction giver", the other as "instruction follower". The instruction giver has a map with a route drawn on it and s/he has to instruct the follower to draw the route on his/her unmarked copy of the map. The participants cannot see each other and cannot see each other's map (see drawing in figure 3.1). Each map contains a number of reference points (e.g., "red point", "the seals" "the stars"). Some features are common to both maps, and some differences between the reference points are incorporated in the maps in order to make the dialogues more complex and elicit more turn changes.

Turn-taking is quite systematic in these dialogues, probably because there are only two participants. Given the map-task setting, cooperation is necessary for the accomplishment of the task, since the auditory channel of communication is maximised (due to the fact that dialogue participants cannot see each other), the production of verbal feedback expressions is also maximised. Feedback is in fact produced by both interlocutors to show attention, assure each other that the conversation can continue, that the message has been delivered and received correctly or not, and so on.

[2] An attempt to compare materials recorded in similar circumstances with two similar opto-electronic systems was carried out in the framework of the PF-Star project (Beskow et al. 2004a), however, besides this attempt, which showed several limitations, it was not feasible to perform further cross-linguistic studies.

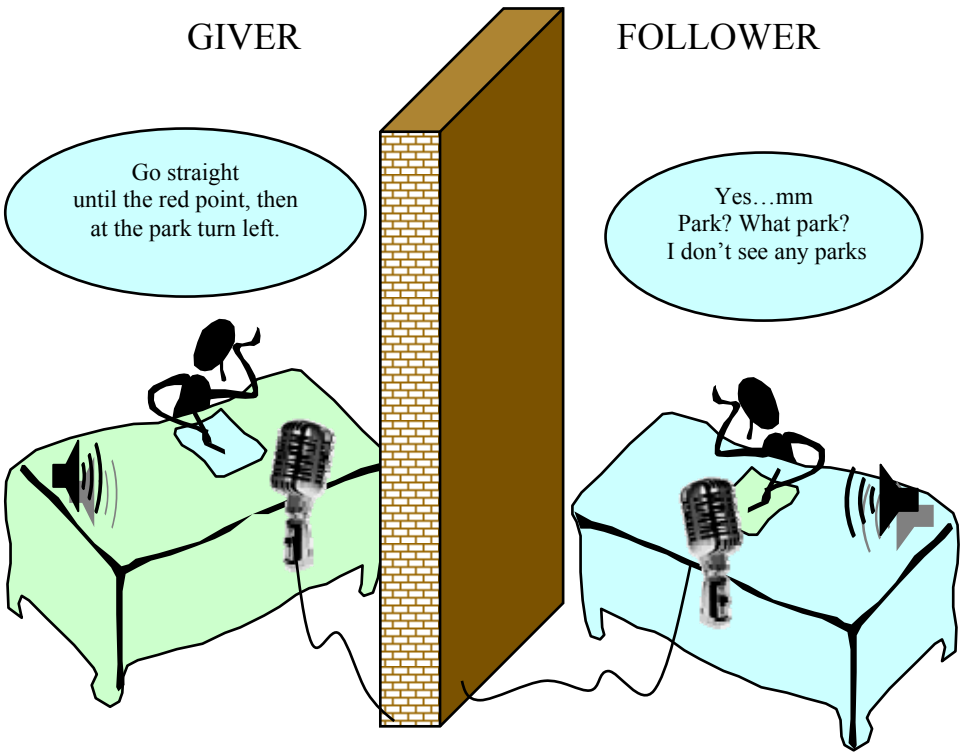

Figure 3.1 A schema of the map-task recording set-up.

Four map-task dialogues, two in Italian and two in Swedish, were selected for the study reported in this thesis. These map-task dialogues were recorded in similar circumstances and have similar characteristics, which allowed for comparative cross-linguistic studies.
The Italian map-task dialogues (referred to as MT-IT) used in this study are:

- MT-IT Dial 1: between 2 male speakers, lasts 5 minutes and counts 133 contributions[3].
- MT-IT Dial 2: between 2 female speakers, lasts 17 minutes, and counts 386 contributions. In this dialogue the two speakers were instructed to exchange their roles half way through the task. For this study only the first part of the dialogue has been used (171 contributions, for circa 8 minutes).

The Italian dialogues are part of the Italian corpus called CLIPS (Corpora and Lexicon of Written and Spoken Italian). CLIPS consists of 100 hours of spoken Italian of different types (dialogues, read speech, TV speech, telephone conversations and special corpora) collected in fifteen different Italian cities, considered as representative of Italian regional varieties.

The two dialogues were recorded in a sound-proof room at the University of Naples. The four dialogue participants were, at the time of the recordings,

[3] A contribution is whatever a speaker says or does: it can be a word, a vocalization, or a non-verbal behaviour.

university students, in their twenties and speakers of the Neapolitan variety of Italian[4].

The Swedish map-task dialogues Swedish (referred to as MT-SW) used in this study are:

- MT-SW Dial 1: between a female and a male speaker, lasts 6 minutes and contains 164 contributions.
- MT-SW Dial 2: between two female speakers, lasts 12 minutes and contains 244 contributions.

The two Swedish dialogues are not part of a large corpus as the Italian ones. They were recorded, together with other two dialogues, in a sound-treated room at Stockholm University [Helgason 2002].

The four dialogue participants, three females and one male were between 30 and 50 years and lived and worked in Stockholm when the recordings were made.

3.2.2 Human-Machine Interactions (AdApt I)

Four audio-recorded human-machine interactions were analysed in this thesis. The interactions were selected from the first AdApt database, which was collected at TMH-KTH in the framework of the AdApt project[5].

The Swedish conversational multi-modal dialogue system AdApt, was developed as a collaboration between KTH and Telia Research [Gustafson et al. 2000]. AdApt consists of a graphical interface containing a city map and an animated agent able to provide information about real estate in Stockholm. A screenshot of the interface is shown in figure 3.2.

The first data-collection with a prototype of the AdApt system was performed in 1999 by means of the Wizard of Oz (WOZ) technique[6]. In WOZ experiments a user interacts with what appears to be a dialogue

[4] The four interlocutors speak the "Neapolitan Southern Variety of Italian" which differs from "Standard Italian" in some phonetic realizations and in the use of some "regional" lexical items or short phrases, deriving from the Neapolitan dialect. In these dialogues however even if the speakers regularly follow the most common Neapolitan phonetic realizations, they very rarely use dialectal lexical items. Only twice did they use regional-dialectal feedback expressions, that is: *vabbé* which is the regional variant of *va bene* (it is fine), and *eh* used with a positive meaning instead of *sì* (yes).

[5] The AdApt project was run in collaboration between KTH/CTT and Telia Research during 1999-2002. One of the aims of the project was to investigate various aspects of human-machine interaction in a multi-modal conversational dialogue systems. Among the output of this project were the development of an experimental multi-modal system, AdApt, in which a user could collaborate with an animated agent to solve several tasks, and the collection of several corpora of user-machine interactions.

[6] This technique takes the name from the film "The Wizard of Oz". Everyone thought "the Wizard" was a tall imposing "living statue" when in fact there was a small man who controlled the "statue" from behind a curtain.

system but is in fact a simulation provided by either a human (referred to as the wizard) or the combination of a human and a computer.

The aim of this first collection was to obtain data for an evaluation of the system under development. This data include a total of 50 dialogues produced by 33 users. These dialogues were only audio recorded, however the agent, in this prototypical version of the system, did not produce any visual communicative gesture and did not give nor elicit any visual feedback.

The AdApt interactions can be described as "factual information seeking" since the users were given the task of finding apartments in Stockholm that fulfilled certain criteria.

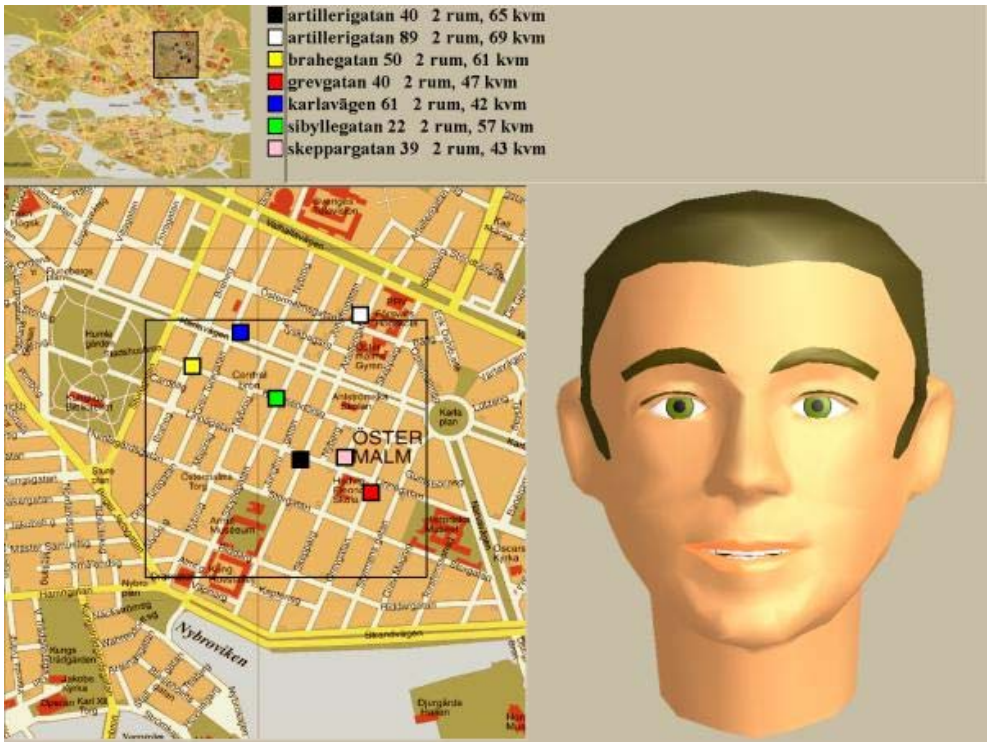

Figure 3.2 A screenshot of the AdApt interface as shown in Gustafson et al. 2000.

Several studies have previously been conducted on the AdApt database, for instance the results of the analysis of positive and negative users' feedback showed that 94% of the users used feedback at least once in their interaction with the system, even if large individual variation was noticed (i.e. some users gave more feedback than others) [Bell & Gustafson 2000].

For the study reported in this thesis, four user interactions (three male and one female user) were randomly selected from those dialogues in which the number of utterances labelled by Bell and Gustafson [2000] as containing feedback was at least 10%. Each dialogue consists of about 200

contributions. This corpus will be referred to as AdApt I and the interactions as AdAptI-Dial 1, AdAptI-Dial 2, AdAptI-Dial 3 and AdAptI-Dial 4.

3.3 Audio-Visual Materials

The audio-visual materials used in this thesis consist of audio-visual recordings of human-human and human-machine interactions which were either selected from existing databases (GSLC, AdApt databases), or collected for the specific purpose of the analysis (Qualisys acquisitions).

3.3.1 Spontaneous Human-Human Dialogues (GSLC)

Digitalized audio-visual recordings of four dialogues between four different customers and a travel agent were selected from the GSLC: Spoken Language Corpus of the Linguistics Department of Gothenburg University [Allwod et al. 2000]. The GSLC includes more than 171 hours of recordings of different activity types, such as telephone calls, role plays, formal meetings and travel agency interactions.

The dialogues used in this thesis will be referred to as Swedish GSLC-Dial 1, GSLC-Dial 2, GSLC-Dial 3, GSLC-Dial 4; they were chosen since they were recorded in the actual setting of a travel agency and can be described as real spontaneous "factual information seeking" exchanges, where the customer asks the travel agent for information about timetables, visas, hotels and so on, and the travel agent provides the information required.

The interactions were video recorded with a video camera and a microphone placed laterally on the desk. The customers did not know in advance that they would be video-filmed; however a written sign, placed on the desk, explained that the interaction was being filmed.

The recordings were made in a travel agency in Gothenburg and because of this a lot of background noise is present in the audio channel. As a consequence the audio quality of the recording does not allow for accurate acoustic analysis of the speech materials. Moreover since the video camera was placed in a steady position on the side of the interlocutors, it does not focus on the full face of the speakers, but rather on their profile. For this reason some of the facial displays are impossible to see. Notwithstanding their limitations, these materials were adequate to analyse and provide a categorization of head movements.

Table 3.1 is a summary of the information about these dialogues.

Table 3.1 Schema of the four dialogues selected from the GSLC corpus.

Dialogue name	**Customer**	**Short description**	**Contributions**	**Duration** (mins.)
GSLC-Dial 1	Female	Booking of a trip to Brazil	107	8,42
GSLC-Dial 2	Male	Request of info about Visa to Thailand	65	2,15
GSLC-Dial 3	Male	Booking of a trip to Thailand	150	16,42
GSLC-Dial 4	Female	Booking of a flight ticket to London	112	27,31

3.3.2 Human-Machine Interactions (AdApt II)

In the framework of the AdApt project a second data collection was made in 2002. Interactions between 24 users and the AdApt system were carried out [Edlund & Nordstrand 2002]. This time the users were also video recorded, both when listening to the instructions given by the test leader and when interacting with the system.

One of the main aims of this further collection was to obtain materials for an evaluation of three different set-ups of the system by using the PARADISE paradigm. The evaluation is presented in Hjalmarsson [2002]. The users were divided in three sub-groups; each sub-group interacted with a different set-up of the AdApt system for half an hour. The three different set-ups were characterized respectively by: a) presence of the agent turn-taking gestures, b) absence of the agent turn-taking gestures, c) absence of the agent turn-taking gestures and presence of an hourglass icon to signal when the system was busy.

The users were instructed to look for information related to apartments for sale in Stockholm that they would have an interest in. After half an hour the test leader interrupted them. Using the video recordings of the interactions it was possible to analyse the non-verbal behaviour of the users, also reported in Cerrato & Ekeklint [2004].

The users were sitting in front of the computer screen; behind the screen there was a digital video camera that filmed the user during the interactions. The users' voice was recorded by means of a microphone, which they could fasten to their clothes.

The original goal of the recordings and set-up of the AdApt system was not the study of non-verbal feedback phenomena, and for this reason it is possible to argue that the recording set-up might have constrained the acquisition and production of some non-verbal behaviour. For instance only

the upper part of the user's body was filmed due of the placement of the camera. This cuts out the possibility to look at the movements of the rest of the body; moreover most of the users ended up holding the microphone in one of their hands, instead than having it fast on their clothes, which might have limited their hand-movements.

Notwithstanding these constraints, the AdApt materials are nonetheless still a valuable source for the investigation of verbal and non-verbal feedback behaviour in human-machine interactions.

From the corpus resulting from the second data collection in the framework of the AdApt project, six users' interactions (three female and three male) were selected from the sub-group of recordings of the system set-up with presence of the agent turn-taking gestures.

In the set-up with the presence of the agent turn-taking gestures, the agent used gestures such as changing of gaze direction, eyebrow raising and head tilting to show when he was busy thinking and when signalling turn-taking.

This corpus will be referred to as AdApt II; table 3.2 shows information related to this corpus.

Table 3.2 Schema of the six interactions in AdApt Corpus II.

Interaction	Subj.	Gender	Total Contributions	System Contributions	Users Contributions
AdAptII-Dial 1	S11	M	383	108	275
AdAptII-Dial 2	S13	F	267	109	158
AdAptII-Dial 3	S22	F	250	81	169
AdAptII-Dial 4	S08	M	244	87	157
AdAptII-Dial 5	S12	F	183	78	105
AdAptII-Dial 6	S13	M	267	68	199

3.3.3 Qualisys Recordings I

Two data acquisitions have been performed with a recording set-up that allows the recording of audio-visual tri-dimensional data: audio data is recorded on a DAT-tape and visual data is recorded both by means of one or two digital video camera/s and with the optical motion tracking system Qualysis[7].

Thanks to a set of infrared reflecting markers attached on the subject's face, the tracking system is able to register the 3D coordinates for each marker at a frame-rate of 60Hz, that is every 17 ms. This allows for the high precision and quality of the data that captures the dynamics of facial displays.

[7] Qualisys MacReflex Motion Tracking System: http://www.qualisys.se (July 2006)

Figure 3.3 is a photo showing a reproduction of the recording session, with the participant with the marker on her face facing her interlocutor, the video camera and the four infra-red cameras[8].

Figure 3.3 Recording set-up in the first data acquisition.

The first data acquisition was carried out in 2002. Since it was foreseen that the results could be implemented in animated conversational agents, a communicative scenario similar to the one that might arise between a user and an embodied conversational agent in a dialogue system was reproduced. In this scenario, which can be also defined as “factual information seeking”, there are two dialogue participants: the “information seeker” and the “information giver” who interact with each other in a spontaneous way exchanging information relative to movies, actors, plots, schedules and so on.

Three native Swedish students at Linköping University of the age between 25 and 30 served as subjects for the first data acquisition: two males in the role of "information giver" and one female as “information seeker”. The information givers had the markers glued on their faces and were recorded by the digital video camera and the four infrared cameras.

Three dialogues were recorded in the first acquisition, but unfortunately one of them could not be used for the analysis since during most of the dialogue the subject sat in a constraining position, bent to one side and often hanging her head on her shoulder. This way it was not possible to measure the head movements. For this reason only two dialogues could be used (from now onward referred to as: 3D-Dial 1 and 3D-Dial 2).

[8] The people reported in figure 3.3 are reproducing the experimental set-up for the sake of documentation, but are not the subjects used in this study.

3.3.4 Qualisys Recordings II (PF-Star Corpus 2 and 3)

The second data acquisition was carried out by the speech group of KTH under the framework of the PF-Star project[9]. The two-year project's (2003-2004) aim was to establish future activities in the field of multi-sensorial and multi-lingual communication, by providing technological baselines, comparative evaluations, and assessment of prospects of core technologies, which future research and development efforts could build from.

One of the main activities of the first phase of the project was the collection of audio-visual speech corpora and the definition of annotation formats. The speech group at KTH collected three multi-modal corpora intended to provide materials for the analysis and modelling of human behaviour to be implemented in synthetic animated agents. [For more details see Beskow et al. 2004b]. For the studies reported in this thesis, data from the PF-Star Corpus 2 and 3 have been used.

PF-Star Corpus 2 includes:

- 180 non-sense words, such as *ADA, ADDA, DAD;*
- 75 short sentences, such as: *båten seglade förbi, grannen knackade på dörren* ("the boat sailed by, the neighbour knocked on the door").

A semi-professional actor was prompted with the non-sense words and the short sentences and was asked to produce them with the following different emotional expressions: *confident, confirming, questioning, insecure* and *happy*, plus *neutral*. These particular expressions were selected since they were expected to be appropriate in the context of dialogue systems. Some of these expressions can be interpreted pair-wise on a positive-negative scale: *confident* versus *insecure*, *confirming* versus *questioning*.

The actor had a total of 35 markers glued on his face and chin. These were used to record lip, eyebrow, cheek, chin, and eyelid movements. Five markers attached to a pair of spectacles served as reference to factor out head movements.

In PF-Star Corpus 3 the subject with the reflective markers is the same semi-professional actor as in PF-Star Corpus 2.

[9] PF-Star: www.pfstar.itc.it December 2006

PF-Star Corpus 3 includes:

- 10 short dialogues, which are schematised in table 3.3; to elicit these dialogues subject-S (the subject with the reflective markers on his face) interacted with one of the male experimenters, subject-M, who is also an amateur actor. They were asked to improvise the dialogues simulating a travel-agency scenario. They were provided with short scripts for each dialogues.
- 75 short sentences uttered with the six basic emotions[10] [Ekman 1982]: *happiness, sadness, surprise, disgust, fear* and *anger*, plus *neutral* thus yielding 7x75 = 525 recorded utterances.
- 15 content neutral sentences (a sub-set of those recorded in PF-Star Corpus 2) all with three content words which could each be focally accented. For the recordings the actor was instructed to utter the prompted sentences with the given emotional expressions and a varying position of the focus. The following expressions of emotion were chosen: *confident, confirming, questioning, insecure, happy,* and *angry*, plus *neutral*. These particular expressions were chosen since, in our opinion, they are likely to be more appropriate for a possible spoken dialogue system scenario. In order to elicit visual prosody in terms of prominence, the short sentences were recorded varying focal accent position first on the subject, then verb, then on the object as in the following example, thus making a total of 15x3x7= 315 sentences uttered with a varying position of the focus:

  Damen vattnade blommor
  Damen vattnade blommor
  Damen vattnade blommor[11]

For the recording of the ten short dialogues, the actor was instructed to interact in the most possible spontaneous way, with one of the experimenters (with whom he was well acquainted). The ten short dialogues intended to provide materials for the analysis of spontaneous non-verbal behaviour (from now onward PF-Star-Dial 1 to10). Each dialogue counts between 10 and 22 contributions per interlocutor. (Appendix C shows one of the PF-Star-Dialogues, with the relative annotation of feedback phenomena).

[10] Many emotion theories use the concept of "basic emotions" [Tomkins 1962; Izard 1977; Johnson-Laird & Oatley 1989], this means that certain emotions are more basic than others. However there is no complete agreement on which these basic emotions are. In the field of Multi-modal studies Ekman's "big six" (*happiness, sadness, surprise, disgust, fear* and *anger)* are often used as a starting point for the analysis and synthesis of emotions.

[11] The sentence in English would be: the lady watered the flowers.

Table 3.3 Schema of the ten dialogues in PF-Star Corpus 3

Dialogue name	**Role of Subject-S**	**Short description**
PF-Star-Dial 1	Agent	Booking a travel to Paris
PF-Star-Dial 2	Agent	Changing a train ticket to a flight ticket
PF-Star-Dial 3	Customer	Booking a flight to London
PF-Star-Dial 4	Customer	Buying an all-inclusive trip to Italy
PF-Star-Dial 5	Customer	Buying a flight to Australia at a student price
PF-Star-Dial 6	Customer	Trying to book a safari to Africa
PF-Star-Dial 7	Customer	Trying to book a fast-train to Lofoten Islands
PF-Star-Dial 8	Agent	Booking a trip to Scotland
PF-Star-Dial 9	Agent	Buying a train ticket to Stockholm
PF-Star-Dial 10	Agent	Asking for a trip information to Thailand

In PF Star Corpus 3 the actor had 29 IR-sensitive markers attached on his face, of which 4 markers were used as reference markers, instead of the glasses. The marker set-up in this data acquisition corresponds to MPEG-4 (Moving Picture Experts Group) Feature Point (FP) configuration (see figure 3.4).

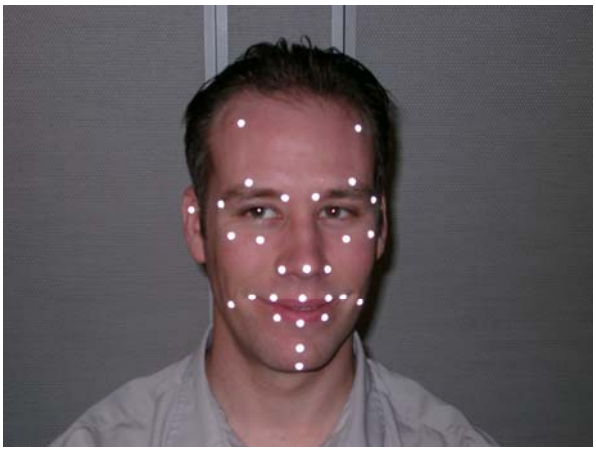

Figure 3.4 Marker placements on subject-S in PF-Star Corpus 3.

This configuration is a compact and standardized scheme for describing human facial displays (i.e. movements of the face and of the head) and it is therefore more suitable for the reproduction of human facial displays in talking heads [Beskow et al. 2004b].

The KTH talking head is based on the MPEG-4 facial animation standard [Pandzic & Forschheimer 2003]. It is a textured 3D-model of a male face consisting of approximately 15000 polygons (Figure 3.5).

The MPEG-4 standard allows the face to be controlled directly by a number of parameters (FAPs, Facial Animation Parameters). The FAPs specify the movements of a number of feature points in the face, and are

normalized with respect to face dimensions, to be independent of the specific face model. Thus it is possible to drive the face from points measured on a face that differs in geometry with respect to the model.

Figure 3.5 The talking head based on the MPEG-4 Facial Animation Standard.

The short sentences uttered with Ekman's basic emotions were especially collected to be used as materials for the data-driven synthesis [Beskow & Cerrato 2004, Beskow & Nordenberg 2005]. The short sentences uttered with topicalised words were recorded having in mind the aim of analysing visual prosody phenomena [Cerrato & Svanfeldt 2005].

The short dialogues were recorded in order to provide materials for the analysis of spontaneous communicative visual expressions [Cerrato 2004].

3.4 Data Transcription and Annotation

All the dialogues used for this study have been manually transcribed and annotated.

Each dialogue is described in terms of contributions. A contribution is whatever a speaker says or does: it can be a word, a vocalization, or a non-verbal behaviour.

The map-task dialogues are orthographically transcribed; the GSLC dialogues were provided with a transcription in GTS (Gothenburg Transcription Standard) [Nivre 1999] standard for transcription meant to be more faithful to spoken language than Swedish standard orthography, though less detailed than a phonetic or phonematic transcription. In GTS standard orthography is used unless there are several spoken language pronunciations of a word. When several variants occur, these are graphically

kept different. According to this principle, the Swedish word "jag" (I), which is mostly pronounced "ja" but occasionally as "jag" is transcribed in both these ways, depending on which form is actually used.

The annotation of all the interactions used for this thesis has been carried out by the author, using the coding scheme presented in chapter 4.

The short sentences in the PF-Star corpora which were used for the investigations presented in this thesis were provided with a phonetic transcription. The audio signal was used to label the data by first providing a phonetic transcription of the material; an automatic aligner [Sjölander & Heldner 2004] was then used to pair the phonetically transcribed speech with the sound signal, and retrieve the exact time for phoneme and word boundaries.

3.5 Summary

The materials analysed in this thesis can be grouped in two main blocks:

- audio recordings of human-human and human-machine interactions;
- audio-visual recordings of human-human and human-machine interactions.

Table 3.4 is a schema with all the materials used for the studies reported in this thesis.

Table 3.4 Schema of all the materials used for the studies reported in this thesis (A= audio, V= Visual, H-H= human-human, H-M=human-machine).

Modalities	Type of speech	Language	Description	Chapter
A	H-H	Italian, Swedish	4 map-task dialogues	5
A	H-M	Swedish	4 dialogues (AdApt Corpus I)	7
A-V	H-H	Swedish	4 dialogues (GSLC)	6
A-V	H-M	Swedish	6 dialogues (AdApt Corpus II)	7
A-V-3D	H-H, expressive	Swedish	2+10 dialogues, prompted sentences	8, 9

3.6 Conclusions

This chapter has shown an overview of all the materials used to carry out the investigations of feedback phenomena which will be presented in the following chapters of this thesis.

A great deal of the materials analysed in this thesis were chosen from existing speech databases and corpora in which the degree of spontaneity would mirror the actual human communicative behaviour, in order to find regularities in human communicative behaviour that could be observed empirically and reproduced in talking heads.

However since even the best of available corpus still suffer from the limitation of not having being collected for the specific purpose of the investigation of feedback phenomena, it was also necessary to collect specific data in a lab environment. The specific data collected for the purpose of analysing non-verbal communicative phenomena related to feedback represent a valuable potential source for the training and testing of a data-driven model of feedback non-verbal behaviour in talking heads.

As will be shown in the following chapters, the variety of materials selected for the analysis carried out in this thesis offers the possibility to apply the method developed for the purpose of analysing feedback phenomena in different speech styles (semi-spontaneous elicited speech vs. spontaneous natural speech), different communicative situations (human-human interactions vs. human-machine interactions), different modalities (audio vs. audio-visual and 3D audio-visual) and across different languages (Italian and Swedish).

The next chapter will introduce the method followed to analyse the data throughout the thesis.

4 Method

4.1 Introduction

The method followed to analyse the data is the same throughout the thesis: the materials were listened to or listened and looked at in order to first carry out an orthographic transcription, if a transcription was not already available. Then verbal and non-verbal feedback expressions were identified and annotated using the coding scheme designed for the purpose of annotating feedback.

The tags used for the annotation are then used to automatically retrieve several quantitative measures, such as the number of occurrences of feedback expressions, their type and position. This information provides an overall picture of the realization of feedback phenomena. A more detailed picture of the specific functions that feedback expressions can carry out in the given context is given by the coded explicit semantic-pragmatic function of each identified feedback expression. Moreover acoustic and visual characteristics of the identified feedback phenomena are analysed in order to get more insight into the use of feedback phenomena across languages, modalities and different communicative situations.

4.2 The Coding Scheme

The study of verbal and non-verbal human communicative behaviour in human-human and human-machine interactions is facilitated by a specific coding scheme for the annotation of the phenomena under observation.

Thanks to the coding scheme it is possible to categorise and label the phenomena under analysis. Since these phenomena can be multi-modal, the coding scheme should provide labels to describe the specific phenomena that are produced in different modalities, and how the different signals produced in different modalities are related.

The initial purpose of the coding scheme presented here was to provide an accurate instrument to annotate verbal feedback phenomena in conversational speech, according to their direction, type and semantic-pragmatic function in the given context [Cerrato 1999]. Then the coding scheme was further enriched to include categories for non-verbal feedback expressions [Allwood & Cerrato 2003] and for the relationship

between feedback expressions produced simultaneously in two different modalities, for instance an m-like word co-occurring with a head nod.

4.3 Coding Procedure

The coding procedure can vary depending on the kind of materials analysed and on the specific purpose of the analysis. First of all in this study it is assumed that in order to be able to code feedback phenomena it is necessary to identify them in the first place. To do so it is crucial to take contextual information into account, which means interpreting and categorising feedback in terms of reactions to the previous communicative act.

For this reason the annotation always starts with a coarse categorization of the identified phenomena and then proceeds, adding more detailed features, at greater degrees of complexity. This way the coding scheme is multi-level. The different levels can be freely added or removed from the annotation depending of the purpose of it and on the need for more or less fine-grained categorization.

The first step in the coding procedure is the identification of the speech act followed by the annotation of its type. A speech act can be verbal, non-verbal or a combination of the two (i.e. multi-modal). The analysis of multi-modal phenomena is of course possible only when the material analysed is multi-modal, which is when the annotator can listen to the audio and look at the video recordings.

Once a first coarse annotation has been performed, then for a specific phenomenon, in this case feedback, it is possible to annotate more features. Particular attention is paid in this coding scheme to the features of non-verbal expressions that describe facial displays, which are described at greater degrees of complexity with more detailed features.

4.3.1 Speech Act

The first step of the annotation consists in the coding of the speech act. To code the speech act a simplified version of the categories proposed by Strassel [2004] for the segmentation of discourse in semantic units has been used.

Strassel proposes to identify all units that function to express a complete thought or idea on the part of the speaker. Since these units do not always correspond to sentences, but they can be phrases, or words, she decides to call them Semantic Units SU.

Semantic Units include: STATEMENTS, QUESTIONS and FEEDBACK.

Beside Semantic Units, Strassel proposes other categories to annotate common phenomena of conversational speech such as fillers (e.g. filled pauses, discourse markers, explicit editing terms and side/parenthetical) and disfluencies. These phenomena are included by Allwood under the category OWN COMMUNICATION MANAGEMENT [Allwood 2001b]. In this thesis the

category HESITATION is included among the speech acts to annotate hesitation sounds such as *ehms,* self-corrections and non-verbal expression that might co-occur with hesitations and self-corrections.

STATEMENT, QUESTION, FEEDBACK and HESITATION are the four main speech acts considered in the coding scheme, and can be annotated in more details at greater degrees of complexity. In the coding scheme presented here higher degrees of complexity are specified only for FEEDBACK.

4.3.2 Feedback Types

Once the speech act is identified as FEEDBACK, its type is annotated. FEEDBACK can be VERBAL and NON-VERBAL.

VERBAL FEEDBACK, as well as other speech act, can be labelled according to their syntactic typology into: WORDS, PHRASES and SENTENCES. To these categories it is possible to add a more detailed level of description in terms of type of word, phrases and sentence, for instance a deeper grammatical analysis, or even an analysis of the phonological phenomena that might characterize the expressions in terms of lengthening, shortening, placement of focus, emphasis and so on.

NON-VERBAL FEEDBACK is first coded using the three following coarse categories: FACIAL DISPLAY, HAND MOVEMENT and OTHER. If higher degrees of complexity need to be taken into account it is then possible to add a more detailed set of features for each group.

In the coding scheme presented here greater degrees of complexity are considered only for facial displays.

4.3.2.1 Facial Displays

FACIAL DISPLAYS include phenomena such as changes in eyebrow position, expressions of the mouth, movement of the head and eyes [Cassell 2000]. With the coding scheme developed in this thesis FACIAL DISPLAYS can be further categorised and annotated by using more specific features concerning: GENERAL FACE, EYEBROWS MOVEMENTS, GAZE DIRECTIONS and HEAD MOVEMENTS. These categories are not mutually exclusive, since several facial displays can co-occur to signal one specific semantic-pragmatic function. For instance a head movement, such as a single nod, can co-occur with a smile and a specific gaze direction. The simultaneous display of different features to signal feedback can be annotated by writing the labels for the observed different features one after the other separated by the sign +. So for instance in the case of a single head nod co-occurring with a smile the annotation would be: S-Nod+smile.

At higher degrees of complexity even more detailed features can be introduced for each of the above mentioned categories, as illustrated below.

GENERAL FACE refers to the general impression that the coder gets from the facial expression of the subject under analysis. The general face is labelled in terms of:

- SMILE: when the facial expression shows pleasure, or amusement, but also derision or scorn. Smile is characterized by an upturning of the corners of the mouth and usually accompanied by a brightening of the face and eyes.
- SCOWL: when the facial expression shows displeasure, scowl, anger. Scowl can be characterized by drawing down or contracting the eyebrows (i.e. frown) in a sullen, displeased manner and may be accompanied by a down turning of the corners of the mouth.
- LAUGH: when the facial expression or appearance shows merriment or amusement, but also derision or nervousness and it is accompanied by an audible vocal expulsion of air from the lungs that can range from a loud burst of sound to a series of chuckles.
- OTHER

The features in these categories are intended as a general description of the facial expression of the speaker under analysis, and more detailed description could be given of the mouth position and movement if needed. In this study this has not been done, however a group of features to describe the position of the mouth related to facial displays other than "articulatory gestures" have been proposed in the MUMIN[12] coding scheme [Allwood et al. 2005].

EYEBROW MOVEMENTS can be annotated in terms of:

- FROWNING: when the eyebrows contract and move towards the nose.
- RAISING: when the eyebrows are lifted.
- OTHER

GAZE DIRECTION gaze refers to "an individual's looking behaviour, which may or not be at the other person" [Knapp & Hall 2002, p.349]. The categories in gaze direction do not consider the eye movements, which, if needed, can be coded using features such as: open, close, up, down, left and right. In this thesis the eye movements have not been coded, however specific features to code eye movements have been proposed in the MUMIN coding scheme [Allwood et al. 2005; 2006].

[12] MUMIN, the Nordic Network for Multi-modal Interfaces, was funded by the language technology programme under NorFA during 2004-2005. The network aimed at stimulating Nordic research in the area of multi-modal interfaces, and increase its visibility in the international research community. One of the outcomes of MUMIN was the development of a coding scheme intended as a general tool for the analysis of gestures (in particular hand gestures and facial displays) in interpersonal communication, focusing on the role played by multimodal expressions for feedback, turn management and sequencing. For more information see http://www.cst.dk/mumin/ (Jan.'07)

GAZE DIRECTION can be categorised as:

- TOWARDS INTERLOCUTOR: the person under observation appears to be looking towards the interlocutor. In a conversation, this corresponds to neutral or normal behaviour. In fact, normally the two interlocutors will be looking at each other. In practice, however, it is often impossible in videos to actually see a mutual gaze, since the camera focuses on one speaker at time.
- UP: when the person looks up.
- DOWN: when the person looks down.
- SIDEWAYS: when the person looks to the side.
- OTHER

HEAD MOVEMENTS are categorized by taking into account the visible deviation of the head position from an initial "default" position, defined as the position before the onset of the movement. So for instance if the most visible movement of the head is downward the movement is classified as a NOD, if the most visible movement of the head is upward the movement is classified as a JERK. Besides direction even the velocity is taken into account, for instance the distinction between a SINGLE JERK and a SINGLE SLOW BACKWARDS UP is based also on the velocity, since the term jerk implies quickness; while a SINGLE SLOW BACKWARD UP refers to a slow up-down movement. The categories provided in the coding scheme to code HEAD MOVEMENTS are the following:

- SINGLE NOD: a single head movement down-up.
- REPEATED NODS: multiple head movements down-up.
- SINGLE JERK: a single quick head movement up-down.
- REPEATED JERKS: multiple head movements up-down.
- SINGLE SLOW BACKWARDS UP: a single slow head movement backwards.
- MOVE FORWARD: is a movement of the head forward, this can either be a movement of the head only or can be a movement of the whole trunk. This movement occurs often as a turn elicit signal.
- MOVE BACKWARD: is a movement of the head backward, which can either be a movement of the head only or can be movement of the whole trunk. This movement occurs often as a turn accepting signal.
- SINGLE TILT (Sideways): a single movement of the head leaning on one side.
- REPEATED TILTS (Sideways): a multiple movement of the head leaning from side to side.
- SIDE-TURN: is a rotation of the head towards one side.
- SHAKE (repeated): is a repeated rotation of the head from one side to the other.

- WAGGLE: is a movement of the head back and forth, side to side, it is like a mixture of shake and move backward or forward it is usually produced to show uncertainty, doubtfulness.
- OTHER: either a different type of movement than those listed, or a combination of two or more of them.

The category OTHER is always present to code those phenomena that are not easily placed under one of the other designed categories.

Rather coarse-grained features are considered for the movements here. All features should be considered as dynamic features that refer to a movement as a whole or a protracted state, rather than categories referring to different stages of a movement. The duration of the movement or state is not indicated as an explicit attribute in the coding scheme. However in the concrete implementation of the coding scheme in a dedicated tool for analysis, it is usually possible to retrieve information about the start and end point of an annotated phenomenon, and even to ensure synchronization between the various modality tracks.

Internal gesture segmentation is not considered either, since it does not seem very relevant for the analysis of the pragmatic-semantic functions that FACIAL DISPLAYS can have. However, the features for shape and dynamics of the non-verbal expressions can be extended and made more detailed for specific purposes without changing the functional level of the annotation.

4.3.3 Multi-Modal Relationship

When information from two different modalities co-occurs it is possible to interpret the relationship between the expressions on the two modalities in terms of dependent or independent. When the expressions on the two modalities are dependent of each other they can either complement or contradict each other.

4.3.4 Feedback Direction

FEEDBACK can be categorised in terms of its direction: expressions can be produced to GIVE feedback, to ELICIT feedback and to GIVE-ELICIT.

This is what Allwood [2001a] defines as “directional function type” or “orientation”. Participants in a conversation give feedback when they wish to show their interlocutor that they are willing to continue the communication and that they are listening, paying attention, understanding/not understanding or agreeing/disagreeing with the message being conveyed. They elicit feedback when they wish to check whether the interlocutor is listening and paying attention, understanding, or agreeing/disagreeing with what they are saying or need more information.

4.3.5 Feedback Function

It is quite common to describe the production of non-verbal behaviour during speech as a complement to speech. Several researchers in fact, talk about accompanying gestures [Poggi & Magno Caldognetto 1996; Teston 1998]. However it is possible to apply an inverse perspective, where the non-verbal behaviour is considered as primary and speech subordinate [Kendon 1975]. According to this perspective non-verbal behaviour is accompanied by speech, in fact non-verbal communicative behaviour can even occur without any verbal behaviour occurring at the same time.

In this thesis neither modality is given primacy; verbal and non-verbal expressions are both considered to be part of the communicative intention of the speakers. For this reason in the coding scheme proposed here the annotation of the pragmatic-semantic function is applicable both to Verbal and Non-Verbal Feedback.

4.3.5.1 Feedback Give

In this thesis expressions identified as FEEDBACK GIVE can have one of the following explicit semantic-pragmatic functions: CONTINUATION I GO ON, CONTINUATION YOU GO ON, ACCEPTANCE, NON-ACCEPTANCE, EXPRESSIVE.

The category **CONTINUATION** has been designed for the situations in which the interlocutor has perceived that there is a message, but s/he explicitly shows only his/her willingness to go on in the communicative interaction, without showing whether s/he accepts the information. This can either be done by producing a short non-intrusive feedback expression (usually a short verbal expression and/or a head nod) that signals continuation of attention and the intention to let the interlocutor continuing to speak (**YOU GO ON**) or by showing the willingness to take the turn (**I GO ON**), and to ask a clarification question. In this last case the feedback expression undergoes lengthening phenomena that fills the pause and signals the intention to keep the turn. This last case is considered also a CONTINUATION because the interlocutor signals, with the lengthening phenomena, the intention to continue the interaction by taking the floor. The lengthening phenomena can also be a signal of some uncertainty and hesitation. It is assumed that in order to be able to produce a FEEDBACK CONTINUATION it is necessary to have perceived that there is a message, but it is not necessary to have understood it; a FEEDBACK CONTINUATION can in fact be given even without hearing what the interlocutor has said.

The category **ACCEPTANCE** takes account of those situations in which the interlocutor explicitly wishes to show that s/he has perceived, received and understood the message (or at least s/he believes so). It is assumed that the feedback expressions belonging to this category imply contact perception and understanding in Allwood's terms and include the concept of

acknowledgement [Clark & Schaefer 1986], which describes a hierarchy of methods used by interlocutors to signal that a contribution has been understood well enough to allow the conversation to proceed.

To show ACCEPTANCE, the interlocutor takes the floor and can either produce a short feedback expression at the beginning of his/her contribution or an expression such as: "I understand, I agree". The verbal expressions produced to show ACCEPTANCE can co-occur with some head movement, such as repeated nods or jerks. Repetition or reformulation of what the other interlocutor has just said can also be used as a way of showing ACCEPTANCE by confirming the correctness of the received information.

The category **NON-ACCEPTANCE** indicates that the interlocutor wishes to show non-acceptance, refusal of the information received. It is assumed that this does not always imply contact, perception and understanding, since the information can be non-accepted because of misperception and misunderstanding, not only because of disagreement. When the interlocutor wishes to show NON-ACCEPTANCE he might either use a negative short expression, such as "no" or an expressions such as: "I do not understand, I am not following, I do not agree" or even some repetition or reformulation of what the other interlocutor has just said in a questioning tone. These verbal feedback expressions might co-occur with head movements such as shakes, waggles, and facial expressions such as eyebrow rising.

The category **EXPRESSIVE** accounts for those cases, other than CONTINUATION, ACCEPTANCE and NON-ACCEPTANCE in which the interlocutor wishes to explicitly colour his feedback expression with some attitudinal/emotional reactions towards the meaning conveyed; this includes for instance surprise, disappointment, frustration, enthusiasm and so on, and implies contact, perception and understanding.

In Allwood's scheme [2001b] the attitudinal reaction is placed on another level, which means that feedback having whatever function can be expressed with a given attitudinal reaction towards the meaning conveyed. In this scheme the expression of an attitudinal reaction is one of the possible primary functions of feedback expressions.

A feedback expression can be multifunctional, since it can signal at the same time understanding and acceptance or understanding and non-acceptance. However in this scheme feedback categories are defined by looking at the explicit function that the feedback phenomena carries out in the given communicative situation, which means that only the most explicit function is annotated.

In the case of continuation feedback two categories are proposed: CONTINUATION I GO ON and CONTINUATION YOU GO ON. These two categories are multifunctional since they indicate that the expression carries out both the function of signalling feedback and turn management. This

choice was dictated by the observation that often feedback and turn-management functions are interwoven, therefore they cannot be mutually exclusive: a communicative signal, whether uni- or multi-modal, may carry several communicative functions at the same time: it might signal ACCEPTANCE and at the same time the intention to gain the floor.

The other categories used to annotate the explicit function of feedback do not include an explicit indication of whether the speaker who gives feedback explicitly intends to signal the intention to take the turn or not, and because it is assumed that when a feedback expression other than CONTINUATION YOU GO ON is produced there is a turn-taking phenomenon taking place. However for an accurate annotation of TURN MANAGEMENT specific categories can be used; these are shown in chapter 7 in table 7.1.

4.3.5.2 Feedback Elicit

In this thesis an expression identified as FEEDBACK ELICIT can have one of the following explicit semantic-pragmatic function: CHECK ATTENTION**,** REQUIRE ACCEPTANCE, MORE INFORMATION**.**

CHECK ATTENTION, when the interlocutor wants to make sure that the other interlocutor is paying attention; this is usually done by asking short question such as: “are you following?”

REQUIRE ACCEPTANCE, when the interlocutor wants to ensure that the other interlocutor has understood what is being said; this is usually done by asking short question such as: “do you understand? Ok?” and by producing repeated had nods.

MORE INFORMATION, when the interlocutor explicitly asks for more information, by using expressions such as: “and then?”

4.4 Labels

The labels used for each category tend to follow a consistent system that aims at being as transparent and logical as possible. The idea is that each label consists of the initial letter or of the two initial letters of the name of the category, so for instance the labels for the speech act categories **STATEMENT, QUESTION, HESITATION** and **FEEDBACK** are respectively **St, Q, H** and **FB**. The label for feedback takes the initial letter of the two words that form the word feed-back.

The label for the semantic pragmatic function of the feedback expression **CONTINUATION, ACCEPTANCE, EXPRESSIVE**, are **C, A, Ex.**

Sometimes the chosen labels seem not to follow a consistent system, for instance in the case of the labels for NON-ACCEPTANCE and POSITIVE ANSWER which are respectively R and RP, originally from the name of the categories: Refusal and Reply, Positive.

In other cases the consistent system is not possible to completely realize because of ambiguity, for instance in the case of the categories for facial display, being so many and so detailed, the labels have more extended names, so for instance a single nod is labelled as S-NOD and a repeated nod is labelled as R-NOD.

The labels for all the categories in the coding scheme are shown in appendix A. In this chapter only the labels for the explicit semantic-pragmatic functions of feedback expressions are illustrated in detail, since these are the ones which will mostly appear throughout the thesis.

Table 4.1 illustrates the use of the labels for the annotation of the FEEDBACK WORD "ok" produced to GIVE feedback with the semantic-pragmatic function ACCEPTANCE, as in the following exchange consisting of two contributions: one instruction to which follows the production of "ok" as a reaction:

Giver: It starts on the left of the red cross
Follower: **ok**

Table 4.1 Example of the labels for the annotation of a FEEDBACK WORD.

Speech act	Type	Direction	Semantic-Pragmatic Function
FB	W	Gi	A

The functions and the relative labels, for expressions produced to GIVE and ELICIT feedback, are summarised in table 4.2a and 4.2b.

Table 4.2a Labels used to code the explicit semantic-pragmatic function of expressions that give feedback.

FEEDBACK GIVE	
Category	**Labels**
CONTINUATION **I** GO ON	**FB**Gi**CI**
CONTINUATION **Y**OU GO ON	**FB**Gi**CY**
ACCEPTANCE	**FB**Gi**A**
NON-**A**CCEPTANCE (**R**EFUSAL)	**FB**Gi**R**
EXPRESSIVE	**FB**Gi**Ex**

Table 4.2b Labels used to code the communicative function of expressions that elicit feedback.

FEEDBACK ELICIT	
Category	**Labels**
CHECK **A**TTENTION	**FB** El**ChA**
REQUIRE **ACCEPTANCE**	**FB** El**RA**
MORE INFORMATION	**FB** El**M**

4.5 Software

The analysis and annotation is more easily performed if the coding scheme is implemented in a dedicated tool for the analysis.

The implementation in the different available tools follows the "partiture" system. This means that the coding is performed in a multi-tier structure, which allows describing different aspects of the same phenomena under analysis [Poggi & Magno Caldognetto 1996]. The different tiers are either automatically synchronised with each other, or can be manually synchronised.

The user interface represents the annotation on parallel lines/tiers on the "annotation board", displayed under the video window. The annotator chooses the tiers that are most appropriate for the given materials and performs the annotation by marking the start and the end point of an element and by inserting the appropriate label.

The coding scheme has been implemented in three tools for audio-visual analysis: Anvil [Kipp 2001], Multitool [Allwood et al. 2002] and WaveSurfer [Sjölander & Beskow 2000] with a video plug-in. The implementation in Anvil has been tested under the framework of two EU projects: PF-STAR and MUMIN [Allwood et al. 2005; 2006]. However all the analyses reported in this thesis have been performed with the support of Multitool and WaveSurfer. A short presentation of the three tools follows, however a formal evaluation of the tools is beyond the scope of this thesis.

Anvil is a video annotation tool which allows annotating and coding human behaviour, in terms of visual accessible information, in temporal alignment with speech. The videos should be in AVI format to be displayed in ANVIL. Before performing the coding the user needs to define a coding scheme, representing the range of behaviour that can occur. The user can divide the audio-video material under analysis in behavioural units.

Anvil represents behavioural units depicted as boxes whose borders represent the initial (left) and end (right) point. During coding and subsequent display of the data, the coding is shown in layers, which are displayed one below the other as in a musical partiture. Coding takes place on a time-aligned annotation board that can be customized with colour-coding to allow efficient and intuitive annotation.

Anvil is a quite flexible tool, able to read data from the widely used, public domain phonetic tools such as PRAAT and XWaves. Anvil can display waveform and pitch contour; the latter must be imported from PRAAT.

Multitool is a video annotation tool, that allows annotating and coding human behaviour. The video format read by Multitool is MPEG. Coding is not automatically time-aligned with the video and the relative transcription in Multitool, for this reason the user needs to insert manual synchronisation points.

Before coding communicative behaviour the user needs to define an annotation-coding scheme, representing the range of behaviour that can occur. The annotation with Multitool is represented in a multi-tier manner on a free-definable number of tiers. Transcriptions and codings are shown in layers, which are displayed one below the other in a partiture-like display. On the partiture display different layers are reported with different colours to allow efficient and intuitive annotation and interpretation. Multitool does not support display of waveforms and pitch contours.

WaveSurfer is a free-downloadable toolkit for the analysis of speech[13]. By adding a video-plug-in that allows the display of videos in .mpeg formats in synchronization with the speech analysis panels, it is also possible to analyse multi-modal materials. In particular the advantage of WaveSurfer lies in the possibility to display- on a panel synchronised to the video and to the waveform- the chosen location dimension of the 3D data acquired with the motion caption system Qualisys (see sections 8.2.2, 8.2.3, 8.3.2).

A multi-layer coding scheme can be easily implemented in WaveSurfer in order to annotate the type and function of the verbal and non-verbal phenomena analysed.

4.6 Conclusions

The method developed in this chapter has been followed to analyse data throughout the thesis. As summarized in chapter 3, different materials have been used to carry out the investigation of feedback behaviour. Some of the materials have been selected from available sources (existing corpora and databases) some others have been collected specifically for the purpose of the investigation.

The advantage of using a variety of materials for the analysis of feedback phenomena offers the possibility to get a more varied picture of the production of feedback phenomena in different communicative situations. The spectrum of varying communicative situations, recording set-ups and speaking styles offered by the materials analysed in this thesis, represent a challenging testing ground for the method of analysis described in this chapter.

In the following chapter, the method has been followed to analyse verbal feedback phenomena in semi-spontaneous dialogues elicited with the map-task technique, in Swedish and Italian. Having available similar materials in different languages has offered the possibility to test the method even across languages.

[13] http://www.speech.kth.se/wavesurfer/

5 Feedback Phenomena in Map-Task Dialogues

5.1 Introduction

The coding scheme described in the previous chapter is the key to all the investigations carried out in this thesis. It works as a classification tool which regulates how to categorise the phenomena under observation with respect to pre-defined categories. For this reason it is important that the classifications done with the coding scheme respond to the criteria of reproducibility.

Since the categories designed to code the semantic-pragmatic functions of feedback expressions are the core of the coding scheme designed for the analysis of feedback phenomena, it is important to assess their quality in order to understand whether it is feasible to use them to annotate the materials throughout the thesis. For this reason, in the first part of this chapter the categories designed to code the semantic-pragmatic functions of feedback expressions are assessed following a strategy which employs three different tests to evaluate the reliability of a coding scheme.

The materials used for the assessment consist of map-task dialogues in Swedish and Italian. Having available similar materials in two different languages offers the possibility to assess the appropriateness of the coding scheme across languages.

Besides the assessment of the appropriateness of the categories in the coding scheme, this chapter deals with the investigation of verbal feedback phenomena in map-task dialogues in Swedish and Italian.

The map-task dialogues used for the investigations reported in this chapter represent a valuable source for the study of different aspects of verbal feedback phenomena across languages. The map-task setting has been originally developed with the intention to elicit unscripted dialogues in such a way as to boost the likelihood of occurrence of certain linguistic phenomena [Anderson et al. 1991]. Given the fact that the participants were interacting with each other in a particular setting, where they could not see each other, it is possible to expect that in these dialogues the non-availability of the visual channel maximizes the use of the verbal channel of communication, and this, together with the actual cooperative task, may results in a large production of verbal feedback phenomena.

The first hypothesis tested in the investigations reported in this chapter concerns the categorization of the semantic-pragmatic function of feedback

expressions. It is hypothesised that the specific semantic-pragmatic function of the identified feedback expression can be categorised by using the pre-defined categories provided in the coding scheme. It is assumed that these categories are independent of the language in which the feedback expression occurs. To test this hypothesis, the semantic-pragmatic function of feedback expressions are assessed using similar materials in two different languages: Italian and Swedish.

The validity of the categories provided in the coding scheme is a necessary precondition to be able to proceed, throughout the thesis, with the analysis of feedback phenomena using different kinds of materials.

The second hypothesis tested in this chapter concerns the distribution of the specific semantic-pragmatic categories of the feedback expressions encountered in the data. It is hypothesised that the most frequent functions that feedback expression carry out are those that signal continuation of attention (FEEDBACK GIVE CONTINUATION YOU GO ON) and acceptance of the received information (FEEDBACK GIVE ACCEPTANCE).

This hypothesis is based on the assumption that feedback expressions serving these two functions are those that most effectively contribute to the smooth and effective unfolding of task-oriented interactions. In map-task dialogues the participants have the specific task to succeed in drawing the correct route on the map of the follower. To do this in an effective way they have to cooperate, and one important means of cooperation is the production of effective feedback signals that facilitate the accomplishment of the task and ensure the smooth unfolding of the interaction. This second hypothesis is tested by quantifying the phenomena of interest in terms of number and percentages of occurring instances.

The third hypothesis concerns the relation between the semantic-pragmatic function of feedback expressions and their acoustic characteristics. It is hypothesised that the semantic-pragmatic function of feedback expressions is reflected in their acoustic characteristics, such as duration, pitch contour and intensity.

Since FEEDBACK WORDS were shown to be the most common feedback expressions in the analysed data, an acoustic analysis of feedback words was carried out to test the third hypothesis, and the acoustic characteristics of feedback words were related to the specific function that they carry out in the given context.

Then a perceptual test was run with the aim of finding evidence that duration and pitch contour can be regarded as perceptual cues to the interpretation of the semantic-pragmatic function that FEEDBACK WORDS carry out, even when they are taken out of the original context.

The investigation of the acoustic characteristics of FEEDBACK WORDS in relation to their communicative function is of fundamental interest given the fact that the same word can be used in different contexts to convey different

feedback functions, and that the different functions can be expressed by means of F0 and duration variation.

Given the fact that FEEDBACK WORDS are frequently used in natural conversational interaction to assure a smooth progression of the communication process, a deeper knowledge of their acoustic realization is of great importance when it comes to technological applications, in that it may help disambiguate ambiguous utterances produced by humans talking to a computer and therefore enhance human-computer interactions.

5.1.1 Materials

Audio recordings of four dialogues, two in Italian and two in Swedish, have been used for the investigations reported in this chapter. The four dialogues were elicited with the map-task technique and acquired in similar circumstances. The two Italian dialogues used for this study are part of the Italian corpus called CLIPS (Corpora and Lexicon of written and spoken Italian). The two Swedish dialogues used for this study were recorded at Stockholm University.

A schema with the information about the dialogues used in this study is shown in table 5.1, for more details see section 3.2.1.

Table 5.1 Schema of the information about the 4 dialogues used for this study

Dialogue name	**Language**	**Speakers**	**Original duration (minutes)**	**Number of contributions**
MT-IT Dial 1	Italian	2 male	5	133
MT-ITDial 2	Italian	2 female	17	386
MT-SW-Dial 1	Swedish	1 female 1 male	6	164
MT-SW Dial 2	Swedish	2 female	12	243

The four map-task dialogues were digitalized under the form of .wav audio files and were provided with an orthographic transcription.

Each dialogue is described in terms of contributions. A contribution is whatever a speaker says or does: it can be a word, a vocalization, or a non-verbal behaviour (not in these dialogues since they are only available in audio format).

The notion of contribution is preferred here since the notion of turn or utterance can be sometimes misleading for feedback analysis. A great deal of short feedback expressions are in fact produced as an insertion in the turn of the current speaker and not as a real turn of their own.

5.1.2 Method

Feedback expressions were identified and coded using the coding scheme developed in chapter 4. The coding procedure starts with the identification of the feedback phenomena. In order to be able to identify and categorise feedback expressions it is necessary to take contextual information into account. In this study feedback expressions are interpreted and categorised in terms of reactions to the previous communicative act [Allwood, Nivre & Ahlsén 1992].

When a dialogue participant receives a message, s/he has to evaluate whether s/he is able and willing to continue the communication; this evaluation is done by means of a response or reaction, given to explicitly exchange information about the state of communication, in other words to make communication efficient. For instance if one interlocutor asks: *do you have a red mark on the top left of your map?* and the other interlocutor reacts by saying: *mhm,* this is interpreted as a feedback reaction that expresses doubt, hesitation, possible misunderstanding, and it is therefore labelled as FEEDBACK GIVE EXPRESSIVE. While if the response of the interlocutor is: *yes* then this is not interpreted as FEEDBACK, but as a POSITIVE ANSWER to a polar question.

Once an explicit feedback expression is identified, its type and direction are annotated. The types of verbal feedback expression can be WORD, PHRASE or SENTENCE. For instance if the identified feedback is: *mhm* this is coded as FEEDBACK WORD, if the identified feedback expression is: *straight on* this is coded as FEEDBACK PHRASE and if the expression is: *I agree* this is coded as FEEDBACK SENTENCE.

The direction of a feedback expression depends on whether a participant to a conversation explicitly wishes to give or elicit feedback. A feedback expression with GIVE direction is usually produced when an interlocutor wishes to signal that s/he is willing to continue the communication and that s/he is listening, paying attention, and understanding/not understanding or agreeing/disagreeing with what the other interlocutor is saying. Feedback with ELICIT direction, on the other hand, is produced when one of the interlocutors wishes to check whether the other interlocutor is listening and paying attention, understanding, or agreeing/disagreeing with what they are saying or need more information.

Thanks to the tags in the coding scheme, several quantitative measures that provide an overall picture of the distribution, type and semantic-pragmatic function of feedback expressions have been retrieved. For instance by counting all the instances of FEEDBACK annotated for a given dialogue participant, it has been possible to obtain a number that indicates the distribution of feedback expressions for this dialogue participant. The distribution of feedback expressions can also be calculated as the percentage of the total number of contributions per each dialogue participant.

A series of acoustic measures, such as duration and F0 contour, have been made in order to find systematic relationships between the phonetic realization of the feedback expressions and their semantic-pragmatic function. In particular acoustic measures have been performed on FEEDBACK WORDS, especially with the aim of comparing the realization of similar feedback words across languages, as for the case of *m-like* words and *sì* and *ja* in Italian and Swedish. Finally a perceptual test has been run, using part of the identified feedback words in Swedish and Italian, with the aim of verifying the hypothesis that duration and pitch contour can be regarded as perceptual cues to the interpretation of the semantic pragmatic function that feedback words carry out, even when they are taken out of their original context.

5.1.3 Feedback Function Annotation

It is assumed that each feedback expression carries out a specific explicit semantic-pragmatic function in the given context. This function is annotated by means of semantic-pragmatic function categories, which are highly dependent on the context.

For instance if the giver says: *passa tra le due macchine* (go between the two cars) and to this the reaction of the follower is: *sí* (yes), this contribution is identified as FEEDBACK WORD. The direction type in this example is GIVE and the semantic-pragmatic function is CONTINUATION YOU GO ON. The labels for this contribution would be: FB;W;Gi;CY.

A schema of the functional categories and the relative labels for FEEDBACK GIVE and FEEDBACK ELICIT expression is re-proposed in tables 5.2a and 5.2b.

Table 5.2a Labels used to code the explicit semantic-pragmatic function of expressions that give feedback.

FEEDBACK GIVE	
Category	**Labels**
CONTINUATION **I** GO ON	**FB**Gi**CI**
CONTINUATION **Y**OU GO ON	**FB**Gi**CY**
ACCEPTANCE	**FB**Gi**A**
NON-**A**CCEPTANCE (**R**EFUSAL)	**FB**Gi**R**
EXPRESSIVE	**FB**Gi**Ex**

Table 5.2b Labels used to code the communicative function of expressions that elicit feedback.

FEEDBACK ELICIT	
Category	**Labels**
CHECK **A**TTENTION	**FB**E|**ChA**
REQUIRE **A**CCEPTANCE	**FB**E|**RA**
MORE INFORMATION	**FB**E|**M**

5.2 Reliability of the Coding Scheme

The categories for the annotation of the semantic-pragmatic functions of feedback expressions have been assessed following the strategy described in [Krippendorff 1980; Carletta 1996; Carletta et al. 1997; Gustafson-Čapková 2005]. The strategy uses *Kappa* statistics to measure inter-annotator agreement on the assignment of pre-defined tag-sets. Three tests are proposed in Krippendorff [1980] to evaluate the reliability of a coding scheme:

1. **Stability test**, or inter-variance test, which checks whether the same coder varies his/her judgments over time.
2. **Reproducibility test**, or inter-coder-variance, which checks the agreement in the codings of two coders.
3. **Accuracy test**, which compares the codings produced by two coders to a standard, if a standard is available.

In this thesis the reliability of the categories designed to annotate the semantic-pragmatic functions of feedback expressions has been tested running the above mentioned three tests.

MT-IT Dial 1 and MP–SW Dial 1 were used to run the tests. These are audio recordings of map-task dialogues acquired in similar circumstances in Italian and Swedish. The advantage of using these dialogues as test materials lies in the fact that they consist of similar materials in two languages, which allows assessing the quality of the annotation cross-linguistically.

The stability and reproducibility test consisted of two main tasks:

- Identification and segmentation of feedback phenomena
- Annotation of their semantic-pragmatic functions using a pre-defined set of categories.

First of all, a method for comparing the segmentations between coders had to be established. It was decided to accept a difference in time coding of under one fourth of a second per segmentation. In other words, if in both annotations a phenomenon was coded within the same time span apart from a possible difference in start and/or end of it of under ¼ of a second, it was

assumed that the two segments described the same unit. This criterion was adopted during MUMIN workshop on multi-modal annotation, which was held at KTH, Stockholm, in June 2004.

The inter-coder agreement was first ensured on the segmentation by checking that the two annotators, or the same annotator who repeats the annotation twice or more, do agree on the identification of the feedback phenomena. Once the agreement on the identification and segmentation of the feedback phenomena is ensured, it is possible to proceed with the calculation of the agreement of the assignment of the categories.

The overall agreement on all the categories used to code the semantic-pragmatic function of the identified feedback phenomena is calculated using the *Kappa* coefficient of agreement. According to Carletta [1996] the *Kappa* coefficient should be used to measure the reliability for category classification, since the amount of agreement one would expect by chance depends on the number and relative frequencies of the categories under test. The *Kappa* coefficient is calculated as follows:

$$K= (P(A)-P(E))/(1-P(E))$$

where *P(A)* is the proportion of times the coders agree and *P(E)* is the proportion of times one can expect them to agree by chance. *P(E)* varies depending on the number of available values that can be assigned to a single feature. For instance, if the annotators can choose between two values, *P(E)* will be 0.50; if the values from which to choose are 4, *P(E)* will be 0.25, and so on.

The value of *Kappa* is 1 in case of total agreement and 0 in case of total disagreement. There is variation among researchers on how to interpret the *Kappa* coefficients in the range between 0 and 1. Some researchers as Carletta et al. [1997] working in the field of natural language processing, propose a strict interpretation of the *Kappa* coefficient in terms of indication of the reliability of the annotations. According to this interpretation, values below 0.67 are not indicating any acceptable agreement and therefore they should not be used to draw tentative conclusions about the reliability of the annotations, while values above 0.67 can be considered as indication of reliability of the annotation.

Other researchers, as for instance Landis and Koch [1977], consider values between 0.41 and 0.60 as indicating a moderate strength of agreement, values between 0.61 and 0.80 as substantial strength of agreement and above 0.80 as almost perfect. El Emam [1999] also considers *Kappa* coefficients between 0.45 and 0.62 as an indication of moderate strength of agreement.

Gustafson-Čapková [2005, p. 98] after an extensive survey of the different thresholds for the interpretation of the *Kappa* values, concludes that:

- values over 0.80 indicate a high degree of inter-annotator agreement;
- values from 0.60-0.65 to around 0.80 a fair degree of agreement;
- values between 0.40 and 0.60 indicate a lower degree of agreement;
- values below 0.40 indicate a poor degree of agreement.

This range is followed for the interpretation of the results of the reliability test.

5.2.1 Stability Test

The stability test for the semantic-pragmatic functions of verbal feedback was performed on the annotations made by an expert coder (i.e. author), who first coded all the materials once and after about six months repeated the coding. For the stability test the first 22 feedback phenomena identified in the map-task dialogues were taken into account to measure the agreement of identification. These 22 feedback phenomena were identified by the expert coder with complete agreement across the two successive annotations in each dialogue. In all the cases there was agreement of segmentation.

The first 22 feedback phenomena represent one third of all the feedback phenomena in MT-IT Dial 1 and ca. one fourth of the feedback phenomena in MP-SW Dial 1. These 22 feedback phenomena in each dialogue have been used for the reproducibility and accuracy test (see Appendix B for the test materials).

The stability test aimed at assessing the appropriateness of the categories designed to annotate the semantic-pragmatic functions of verbal feedback phenomena. In the materials used for the test, most of the identified feedback phenomena have a GIVE direction, and for this reason only the semantic-pragmatic functions for FEEDBACK GIVE expressions were assessed. These functions are summarized, with their relative labels, in table 5.2.

The overall agreement on the assignment of semantic-pragmatic categories on the identified feedback phenomena in the two successive annotations of the same expert annotator has been calculated using the *Kappa* coefficient of agreement.

The results of the stability test show that the same coder had 82% of agreement in the assignment of the semantic-pragmatic categories for verbal feedback in the Italian materials and 95% in the Swedish materials.

For the calculation of the *Kappa* coefficient $P(E) = 0.20$, which gives the results shown in table 5.3.

Table 5.3 Result of the stability test, consistency among successive coding of the same coder.

Materials	***Kappa* coefficient**
MT-IT Dial 1	0.77
MT-SW Dial 1	0.94

The *Kappa* coefficient indicates that the coding is stable over time and the agreement for categories assignment is substantial, especially for the Swedish materials.

For the Italian materials, in four cases the coding showed some disagreement. In particular in the first coding, two items were assigned to the category CONTINUATION I GO ON while in the second coding to the category ACCEPTANCE, and two other items were first assigned to the category ACCEPTANCE and in the second coding to the category CONTINUATION YOU GO ON. The other categories showed instead complete inter-coder agreement.

For the annotation of the 22 identified feedback expressions the expert annotator used some of the categories more frequently than others. This means for instance that both in the Swedish and in the Italian materials the category feedback EXPRESSIVE was assigned only two times and the category NON-ACCEPTANCE only once, while the other categories had an even distribution.

5.2.2 Reproducibility Test

The reproducibility test aimed at assessing the reproducibility of the assignment of the semantic-pragmatic functions of verbal feedback.

Two linguists, one native speaker of Swedish with good fluency in Italian, and one native Italian speaker with good fluency in Swedish participated in the test. The two linguists ran the test in two different locations, but under similar circumstances: the native speaker of Swedish ran the test at the department of Speech, Music and Hearing in Stockholm, the native speaker of Italian ran the test at the department of Linguistics of the University of Naples. Both the coders were asked to identify verbal feedback expressions in the same materials used by the expert coder to perform the stability test, that is MT-IT Dial 1 and MP-SW Dial 1 up to the 22nd feedback phenomenon identified by the expert coder in her second annotation.

The two linguists received both an oral explanation and written instructions about the tasks to perform, which is identification of the feedback phenomena and annotation of their semantic-pragmatic functions. They also received written definitions of feedback and an explanation of the categories to assign. The two linguists were instructed to code the

semantic-pragmatic function by using the five categories presented in table 5.2.

Before starting their task they listened to some examples of verbal feedback expressions to get accustomed to their tasks. The test took about five hours for each linguist. The two coders carried out their annotation in different places and at different times, however, they followed the same procedure which included two tasks.

The first task of the test consisted in the identification of the verbal feedback phenomena. The two coders agreed in 90% of cases in identifying and segmenting verbal feedback phenomena.

The second task consisted in the assignment of the pre-defined semantic-pragmatic functions to the identified feedback expressions. The overall agreement on the assignment of semantic-pragmatic categories on the identified verbal feedback phenomena has been calculated using the *Kappa* coefficient, and the results are shown in table 5.4. The *Kappa* coefficient for the Italian materials is 0.6, for the Swedish materials is 0.69.

Both in the Italian and in the Swedish materials the disagreements in the assignment of the categories did not concern one category in particular.

Table 5.4 Results of the reproducibility test: inter-coder agreement for the semantic-pragmatic categories of verbal feedback.

Materials	***Kappa* coefficient**
MT-IT Dial 1	0.60
MT-SW Dial 1	0.69

5.2.3 Accuracy Test

For the accuracy test the annotations of the two coders were compared with the second annotation made by the expert coder, which was considered as the "golden standard". The results of the inter-coding agreement and the *Kappa* values are showed in table 5.5.

Table 5.5 Results of the accuracy test: inter-coding agreement for the semantic-pragmatic categories of verbal feedback among the codings of the two coders and the golden standard.

	Percentage of agreement		***Kappa* coefficient**	
Materials	**Swedish Coder**	**Italian Coder**	**Swedish Coder**	**Italian Coder**
MT-IT Dial 1	72%	77%	0.65	0.71
MT-SW Dial 1	82%	77%	0.77	0.71

The *Kappa* values for the accuracy test indicate a fair degree of agreement.

Not surprisingly the Swedish coder obtained better values on the Swedish materials and the Italian coder obtained better values on the Italian materials.

The comparison of the annotation showed some differences in the assignment of the semantic-pragmatic categories, but no disagreement concerned any category in particular.

One case of disagreement from the Italian MT dial 1 is here translated into English for exemplification: the giver in giving the instruction about the route he has on his map, says: *then it follows, so to speak, a route towards the left, which is horizontal*, to this the follower reacts by saying: *yes*. This reaction has been interpreted as FEEDBACK GIVE ACCEPTANCE by the expert coder, while both the coders who participated in the test assigned to this item the category FEEDBACK GIVE CONTINUATION YOU GO ON.

5.2.4 Conclusions on Reliability

The results of the reliability test run on a sub-set of the corpus used in this chapter, with the aim of assessing whether the categories designed for the semantic-pragmatic functions of verbal feedback expressions are appropriate to code feedback phenomena, can be considered as positive.

The reliability and ease of use of the categories in the coding scheme and feasibility across languages is indicated by the scores of the *Kappa* coefficients, which range between 0.6 and 0.94, thus indicating a fair degree of agreement for the assignment of the pre-defined semantic-pragmatic categories for feedback functions.

Considering the fact that assigning pre-defined theoretical categories always implies a dose of subjectivity, the results obtained in the reproducibility test can indeed be considered as positive.

It could be argued that the observed consistency in the stability test is not due to the validity of the coding scheme, but rather to the fact that the expert coder is also the developer of the coding scheme. Unfortunately this is a difficult shortcoming to avoid, since often empirical work in linguistics builds on the subjective judgements of the researchers themselves.

Nonetheless the results of the reproducibility and accuracy test showed also consistency across coders, which can be interpreted in favour of the validity of the scheme.

The audio recordings of the map-task dialogues used to assess the reliability of the categories designed for the semantic-pragmatic functions of verbal feedback expressions allowed for the assessment of the feasibility of the categories across languages, but did not allow for the assessment of the categories across modalities, since it was not possible to use them to annotate the semantic-pragmatic functions of non-verbal behaviour related to feedback.

An attempt to assess the degree of agreement in the identification of non-verbal feedback–phenomena, and the assignment of the semantic-pragmatic categories for feedback function to non-verbal feedback phenomena was run during the MUMIN workshop on "Multi-modal Annotation" held in 2004 at KTH, Stockholm [Allwood et al. 2005; 2006].

The results of the reproducibility test run by two non-expert coders who independently coded the semantic-pragmatic functions of facial displays related to feedback in a one-minute clip extracted from a TV talk-show in Danish, showed *Kappa* scores ranging between 0.68 and 0.9. This result can be taken as a positive indication of the validity of the categories, not only across languages, but also across modalities. As a consequence of these positive results it was assumed that the coding scheme could be considered a useful tool for the investigation of feedback phenomena in different materials.

5.3 Cross-Linguistic Analysis

Having evaluated the reliability of the semantic-pragmatic categories designed to annotate the specific functions of feedback expressions, it has been possible to proceed with the analyses. The tags provided by the annotation help to automatically retrieve several quantitative measures that provide an overall picture of the distribution, type and function of feedback expressions across languages.

5.3.1 Feedback Distribution

The first step in the investigation of feedback consists in providing a quantification of the phenomenon in terms of number of occurring instances. After all there is no guarantee that feedback phenomena will appear with the needed frequency in the selected speech materials. Even big corpora may fail to provide sufficient instances of feedback phenomena, in particular when the corpora are not originally intended for the purpose of analysing feedback phenomena.

The occurrences of feedback expressions can be measured in two ways:

1. by counting the occurrences of contributions in the dialogues that include at least one feedback expression;
2. by counting all the feedback expressions produced in the dialogues.

Measure 1, counting the occurrences of contributions in the dialogues containing at least one feedback expression, does not take into account all the occurrences of feedback expressions. In example 1 from MT-IT Dial 2, the contribution \$G382[14] is an utterance containing two feedback expressions: one initial: *eh*, produced to give a counter-feedback with the

[14]The convention used to number and indicate whose contribution it is in the dialogue is the following: \$=turn, G=giver or F=Follower followed by the contribution number in the dialogue. This way \$G382 indicates the 382nd contribution in the dialogue, and that it was produced by the giver; this does not mean the 382nd contribution of the giver, but the 382nd contribution in the dialogue.

function ACCEPTANCE (FB;Gi;A) to the preceding feedback produced in $F381: *sí*, and one final feedback expression: the short question *ci sei?* (are you following?) produced to elicit feedback with the function CHECK ATTENTION (FB;El;Ch;A).

By applying measure 1, only one utterance containing feedback is counted in contribution $G382. By applying measure 2 two feedback expressions are counted. In conclusion: by applying measure 1 to MT-IT Dial 2 we obtain that 42% of the produced utterances include at least one feedback expression. By applying measure 2 we count a total of 167 feedback expressions in MT-IT Dial 2.

$G380:	*ora devi entrare in mezzo a questo gruppo di palline* (now you have to go through this group of small balls)
$F381:	***sí* <FB;W;Gi;A>** *(yes)*
$G382:	***eh*< FB;W;Gi;A>** *e quindi continui // come se fosse una linea retta//* (eh and then you go on as if it was a straight line) *fino // all'ultima pallina /in basso /che sta spostata leggermente* *verso<+>destra* (until the lowermost ball which is slightly moved towards the right) // ***ci sei ?* <FB;S;El;ChA >** (are you following?)

Example 1 from MT-IT Dial 2 of one contribution containing an initial FEEDBACK GIVE and a final FEEDBACK ELICIT[15] (the English translation is shown under each contribution in parenthesis).

The cases in which a contribution includes more than one feedback expression are not so frequent in the four MT-dialogues, as shown in table 5.6, which reports the distribution of utterances containing feedback expressions and the total number of feedback expressions per dialogue.

The distribution of feedback expressions has been calculated also per speaker, as shown in the bar chart in figure 5.1. The percentage of the occurrences of feedback expression per each speaker has been calculated relative to the total number of feedback occurrences. So for instance in MT-IT Dial 2 where the total number of identified feedback expressions is

[15] In the transcription convention short pauses are indicated by / and long pauses by //. Lengthening of final vowels are indicated by a + in angle brackets. In example 1 *verso*<+> means that the final vowel of this word undergoes lengthening. Overlapped productions are transcribed in square brackets and cross-referred with a number, so in the example 4 the transcription of the words *cerchio* and *aha* in square brackets and with the number 1 in front means that these two words are produced in overlap with each other.

167 (see table 5.6), 65 of these feedback expressions have been produced by the giver, which makes 39% of the total number of feedback, and 97 feedback expressions have been produced by the follower, which makes 61% of the total number of feedback (as shown in figure 5.1).

Table 5.6 Distribution of contributions containing feedback and number of feedback expressions in each MT dialogue.

Dialogue	**Number of contributions**	**Contributions containing at least one Feedback**		**Feedback occurrences**	
		Number	**Percentage**	**Number**	**Percentage**
MT-IT Dial 1	133	68	51%	85	64%
MT-IT Dial 2	386	162	42%	167	43%
MT-SW Dial 1	164	92	56%	96	58%
MT-SW Dial 2	243	121	50%	129	53%

In all the dialogues the follower produces more contributions containing feedback expressions. This is quite natural, since her/his role is to follow and understand the instructions of the giver, therefore s/he produces a great number of feedback to ensure the giver that s/he is following, understanding, agreeing, and so on.

The dialogue participants were interacting with each other in a particular setting, where they could not see each other. This way even if non-verbal behaviour were produced by the participants in the dialogues, they were not intended to be communicative, since the interlocutors were aware of the fact that they could not see each other. For this reason the verbal channel of communication may have been maximised, which can be one explanation for the great deal of verbal feedback phenomena encountered in these data.

In support to this supposition are the results of a previous investigation using the map-task setting, which showed evidence that face-to-face communication were shown to be more efficient than audio-only interactions [Boyle, Anderson & Newlands 1994]. For the same level of task performance, participants who could see each other produced shorter dialogues and interrupted each other significantly less than in audio-only conditions. This means that when the map-task participants communicate without seeing each other, a number of adaptations might take place to compensate for the loss of information from the visual channel. For example they may make the most of the auditory channel by producing more verbal feedback.

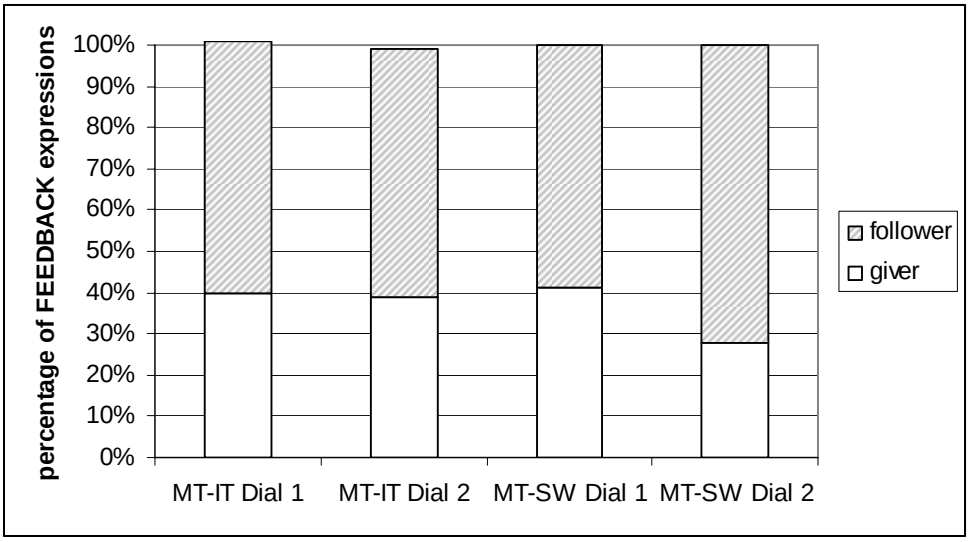

Figure 5.1 Distribution of contributions containing feedback expressions per speaker and per dialogue.

In order to give a more complete picture of the amount and the type of feedback expressions produced in the four dialogues, detailed lists of all the encountered feedback expressions are reported in table 5.7a for Italian and in table 5.7b for Swedish.

Feedback words are the most frequent type of feedback encountered in the four dialogues. *Ja* and *sí* are the most frequent items. This is not surprising since the equivalent of "yes" or a variant of it is one of the most used feedback expressions in many languages, as reported in previous result obtained studying feedback expressions in Dutch [Caspers 2000], Italian, Danish, Swedish and English [Bergström et al. 2002].

The second most frequent feedback words are m-like words. These are very common in the Swedish MT dialogues: 24% of all the feedback expressions produced are m-like words. They can serve several semantic-pragmatic functions, which are conveyed by means of different prosodic cues (as shown in section 5.4.3), for instance by means of different pitch contour and/or different duration.

In the Italian MT dialogues the negation *no* is used more often than the equivalent *nej* in the Swedish MT dialogues with the function of NON-ACCEPTANCE. In the MT-Italian dialogues of 18 produced *no*, 13 were used with the function FEEDBACK GIVE NON-ACCEPTANCE (FBR). In the Swedish MT dialogues *nej* is produced only 8 times of which 5 with the FUNCTION FEEDBACK GIVE NON-ACCEPTANCE. Some variant of *nej,* for instance *nja, nä* or m-like words are instead used to show NON-ACCEPTANCE in the MT-Swedish dialogues. However it seems like these variants are coloured with an expression of doubt, hesitation rather than definite NON-ACCEPTANCE.

Table 5.7a Occurrence of feedback expressions in the Italian MT-dialogues

Feedback Expressions	Number of occurrences
sí	69
sí sí / sí sí sí	8
m-like words	27
repetition, reformulation	23
ah	18
no	18
ok	15
poi	12
eh	12
vabbe	8
anticipation	8
esatto	6
ci sei?	6
capito? ho capito	6
laugh	3
ottimo	3
ecco	3
aspetta un momento	3
vai	2
perfetto	1
giusto	1
Total number of FEEDBACK	252

Table 5.7b Occurrence of feedback expressions in the Swedish MT-dialogues.

Feedback Expressions	Number of occurrences
ja	84
ja ja	3
m-like words	53
okej	22
jaha	13
repetitions, reformulations	9
nej	8
ja just det	7
ja okey	6
ja precis	6
eller	4
nja	4
nä	3
just precis	2
visst	1
Total number of FEEDBACK	225

5.3.2 Feedback Types and Direction

As already mentioned in the previous section, most of the identified feedback expressions in the four MT-dialogues consist of WORDS. These short expressions are used both to GIVE and ELICIT FEEDBACK, however the percentages of feedback with direction ELICIT are much lower compared to those for feedback GIVE in all the four dialogues The percentages of FEEDBACK expressions with GIVE and ELICIT direction for the Swedish and Italian MT dialogue are shown in table 5.8. These percentages are calculated relative to the total number of occurrences of feedback in the two MT dialogue in each languages.

Table 5.8 Percentages of FEEDBACK expressions with GIVE and ELICIT direction for the Italian and Swedish MT dialogues.

	FB GIVE	**FB ELICIT**
MT-IT dialogues	78%	12%
MT-SW dialogues	93%	7%

Table 5.9 shows the distribution of feedback types per direction for the two Italian and the two Swedish MT-dialogues. Both for feedback with give and elicit direction, most of the feedback expressions consist of words. This might depend on the fact that in task-oriented interactions feedback needs to be as effective as possible, therefore it is expressed by means of short and concise expressions such as “yes, no”, m-like words and other short words, that convey important information about the state of communication.

Table 5.9 distribution of feedback types per direction in the Italian and Swedish MT-dialogues.

FEEDBACK type	**FEEDBACK Direction**			
	GIVE		**ELICIT**	
	MT-IT dialogues	MT-SW dialogues	MT-IT dialogues	MT-SW dialogues
WORD	81%	87%	54%	60%
PHRASE	13%	10%	21%	20%
SENTENCE	6%	3%	25%	20%

The most common way of expressing verbal feedback is by means of short words; however feedback expressions can also consist of repetitions, reformulations of parts or of the entire previous utterance, and anticipations of the end of the current utterance.

By repetition is meant the repetition by one of the interlocutors of key words, phrases or sentences uttered in the previous contribution of the other interlocutor. By reformulations it is meant a paraphrase by one interlocutor of the previous contribution (or of a part of it) of the other interlocutor.

In the analysed dialogues, feedback under the form of repetition or reformulation is produced when interlocutors are dealing with complex information, such as when trying to draw the route or looking for some specific markers on the map. The role of repetitions and reformulations seems to be twofold:

- give, or elicit feedback in a marked way;
- help in the cognitive process of acquisition of information, which means that repetitions and reformulation of the received instructions help listeners (in this case the followers in the map task) to "think aloud" while trying to accomplish their task of drawing the route on the map.

An instance of a repetition is shown in example 2, from MT-IT Dial 1. The giver is saying that the starting point is *a sinistra di una televisione* (on the left of the TV set), and the follower repeats exactly the same words: *a sinistra della televisione* and then he says *vai* (go on).

$G7:	*parte da<+> / alla <+> sinistra di una televisione /*
$F8:	***alla sinistra della televisione*< FB;Ph;Gi;A>** / *vai*

Example 2 from MT-IT Dial 1: Feedback under the form of repetition.

A similar example is shown in example 3 from MT-SW Dial 1. The follower says: *då går ja nära strandkanten* (then I go near the shore) and the giver repeats almost the same utterance to signal that the indication/instruction has been understood, accepted and it is going to be followed (FBA).

$F67:	*då går jag nära stranden*
$G68:	***då går du nära strandkanten* < FB;Ph;Gi;A>**

Example 3 from MT-SW Dial 1: Feedback under the form of repetition.

Both repetitions and reformulation can be produced with an F0 contour typical for questions, when the dialogue participant needs to ask for a clarification related to the correct understanding of the message. When repetitions and reformulations are produced as FEEDBACK GIVE they mainly have the function ACCEPTANCE, when they are produced as FEEDBACK ELICIT they are formulated or uttered as a question with the function ASKING FOR MORE INFORMATION.

Another phenomenon, which can be interpreted as feedback, is the anticipation of the end of the contribution of the interlocutor. Examples of anticipations are found only in the Italian MT dialogues, eight in total. These occur when the follower wishes to show that s/he has understood the instruction which is being given by the giver and is ready to move on to some new information/instruction. As a consequence s/he shows her/his "impatience" by completing the instruction, which is being given by the

giver. One instance of anticipation from MT-IT Dial 2 is shown in example 4.

The giver in $G111 is trying to explain, with some hesitations, where the follower has to start: *allora / dalla<+> dalla macchina dalla quale insomma<+> insomma partiamo*, (well, from the car from which well, we start) while doing this the follower anticipates the end of her utterance by saying: *dalla macchina rossa* (from the red car) which overlaps with the last word uttered by the giver. This is an example of feedback under the form of anticipation of the end of the contribution of the giver by the follower. To this anticipation the giver, in contribution 113, reacts by giving a feedback: *esatto* (exactly) with the function of showing ACCEPTANCE.

$G111:	*allora / dalla<+> dalla macchina dalla quale insomma<+> partiamo*
$F112:	***dalla macchina rossa* <FB;Ph;Gi;A>**
$G113:	***esatto* <FB;W;Gi;A>**

Example 4 instance of feedback under the form of anticipation.

5.3.3 Semantic-Pragmatic Functions (FEEDBACK GIVE)

Participants in the four dialogues mainly produce feedback with GIVE direction. Giving feedback can be accomplished in Italian by using short words, such as *sì, mm, mhm, ok, ah, eh,* and in Swedish by using *ja, mm, mhm, ok, ja visst, jaha.*

The distribution of FEEDBACK GIVE expressions per semantic-pragmatic category is shown in figure 5.2a for the two Swedish MT-dialogues and in figure 5.2b for the two Italian MT dialogues. The semantic-pragmatic categories for feedback give expressions are: CONTINUATION YOU GO ON (FBGiCY), ACCEPTANCE, (FBGiA), NON-ACCEPTANCE (FBGiR), EXPRESSIVE (FBGiEx).

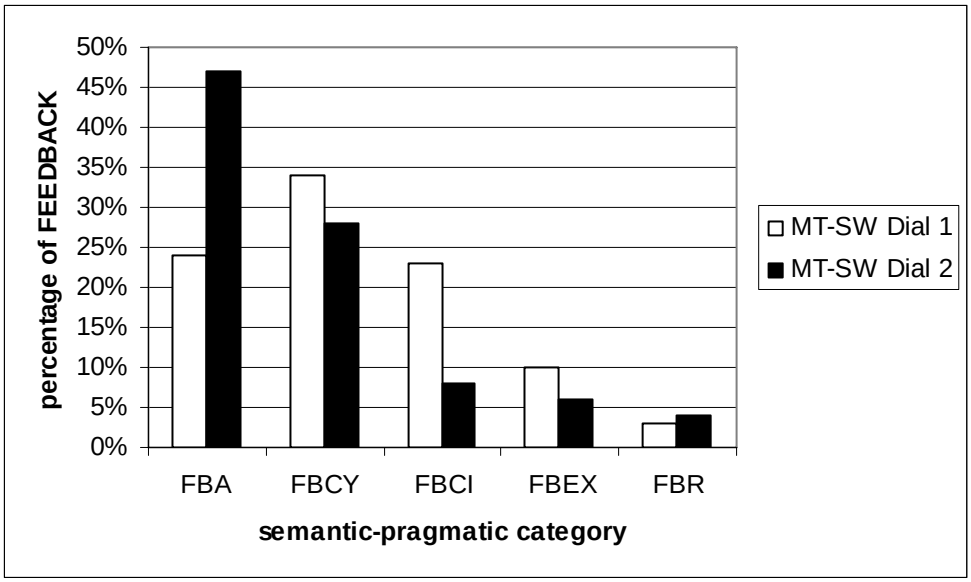

Figure 5.2a Distribution of FEEDBACK GIVE expressions per semantic-pragmatic category in the two Swedish MT-dialogues.

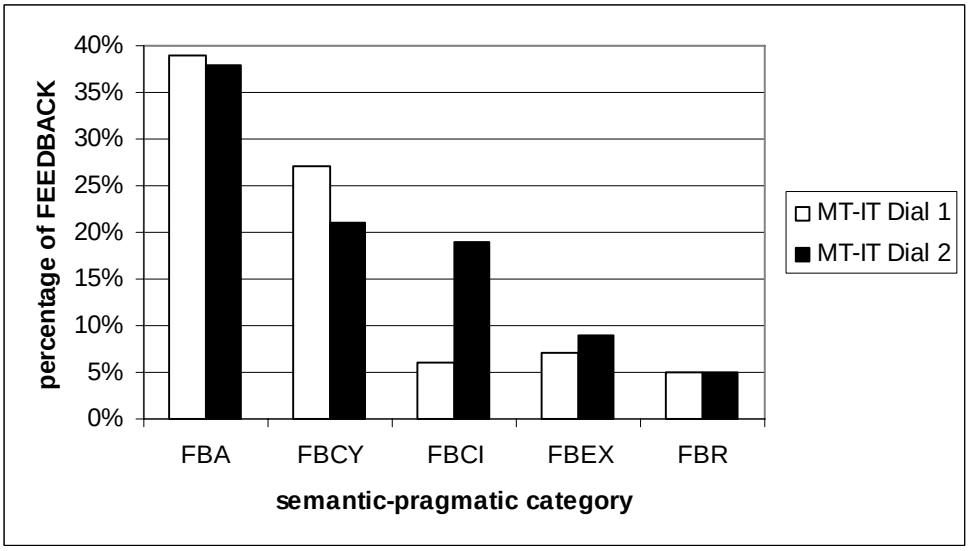

Figure 5.2b Distribution of FEEDBACK GIVE expressions per semantic-pragmatic category in the two Italian MT-dialogues.

The most common semantic-pragmatic functions across languages are FEEDBACK ACCEPTANCE (FBA) and FEEDBACK CONTINUATION YOU GO ON (FBCY). This result confirms the initial hypothesis that these two functions are the most frequent since they are those that most effectively contribute to the smooth and successful unfolding of task-oriented interactions.

Another point in favour of the assumption of effectiveness of the functions GIVE FEEDBACK ACCEPTANCE and GIVE FEEDBACK CONTINUATION YOU GO ON is the fact that most of the feedback expressions having these two functions consist of FEEDBACK WORDS. In particular in the Swedish MT dialogues, m-like words are often produced in a minimally intrusive way to show CONTINUATION YOU GO ON.

Non-intrusiveness can be considered as another aspect of effectiveness of feedback with GIVE FEEDBACK CONTINUATION YOU GO ON FUNCTION. In order to quantify non-intrusiveness, the distribution of the number of times when feedback words are produced in the four dialogues in a non-intrusive way has been calculated, relative to the total number of feedback expression in the dialogues. This percentage is shown in table 5.10 for the Italian and Swedish dialogues.

Table 5.10 Percentage of occurrence of short non intrusive feedback per dialogue.

Dialogue	**Percentage of non-intrusive feedback words**
MT-IT Dial 1	9%
MT-IT Dial 2	6%
MT-SW Dial 1	21%
MT-SW Dial 2	8%

By using a non-intrusive feedback the listener pops in during the contribution of the other speaker to produce a short verbal expression that shows that s/he is paying attention to what is being said and s/he is willing to continue the communication without the intention of taking the floor and interrupting the speaker. One instance of this behaviour is illustrated in example 5, which shows an occurrence of *sì* (yes) produced by the follower in MT-IT Dial 1, in minimally intrusive way.

$G67:	*passa tra le due macchine /*
$F68:	***sì* <FB;W;Gi;CY>**
$G69:	*e poi c'è il puntino nero*

Example 5 from MT-IT Dial 1: sì used as in a minimally intrusive way.

In this example the giver is giving the instruction: *passa tra le due macchine (go between the two cars).* After this first utterance the giver makes a short pause (marked by: /) in which the follower appropriately inserts the feedback word *sì* with FEEDBACK GIVE CONTINUATION YOU GO ON function.

Example 6 is a similar case from MT-SW Dial 2. Here *ja* is used to give feedback after an instruction. The giver says: *då ska du börja gå lite mer söderut (you have to start to go more towards south)* followed by a short pause before going on with the instruction in $G169*;* during the production

of the giver's short pause the follower ($F168) "inserts" the short feedback expression *ja* with the function FEEDBACK GIVE CONTINUATION YOU GO ON.

$G167:	*då ska du börja gå lite mer söderut /*
$F168:	***ja* <FB;W;Gi;CY>**
$G169:	*men du får hålla dig en aning i sydöstlig riktning*

Example 6 from MT-SW Dial 2: ja used in a minimally intrusive way.

In these last two examples feedback is produced in a minimally intrusive way, without overlapping with the giver's contribution almost as if the follower knew exactly the appropriate timing for the production of this short feedback expression. The result of a study on feedback production in Japanese and English [Ward & Tsukahara 2000] has shown that feedback is in fact produced in "appropriate" points of the conversation (which correspond to a region of low pitch late in the utterance) so that it does not interrupt the contribution of the main speaker. According to this study Japanese subjects were even able to predict, (on the basis of prosodic cues like F0 contour) where the feedback was going to occur. This means that participants in a conversation not only seem to know when it is appropriate to produce feedback, but they even expect their interlocutor to produce feedback and it is for this reason that when feedback is not produced communication can break down.

5.3.4 Semantic-Pragmatic Functions (FEEDBACK ELICIT)

Eliciting feedback does not occur as often in the analysed MT dialogues; only 3% of the identified feedback expressions have been assigned to the direction ELICIT in the two MT-Swedish dialogues and 12% in the two MT-Italian dialogues.

Participants in a dialogue can elicit feedback when they wish to know whether the interlocutor is listening and paying attention (**FB**El**ChA**), when they wish to require acceptance, agreement from the other interlocutor (**FB E**l **RA**) and when they wish to require more information from the interlocutor (**FB** El **M**).

In the Italian MT dialogues short questions, such as *ci sei?* (are you following?), are asked to check the interlocutor's attention, short questions such as *vero? no?* (really? isn't it?) *capito?* (do you understand?) are asked to require acceptance and short questions such as *poi?* (then?) are used to ask for more information.

Figure 5.3 shows the distribution of FEEDBACK ELICIT expressions per specific semantic category in the two MT-Italian dialogues.

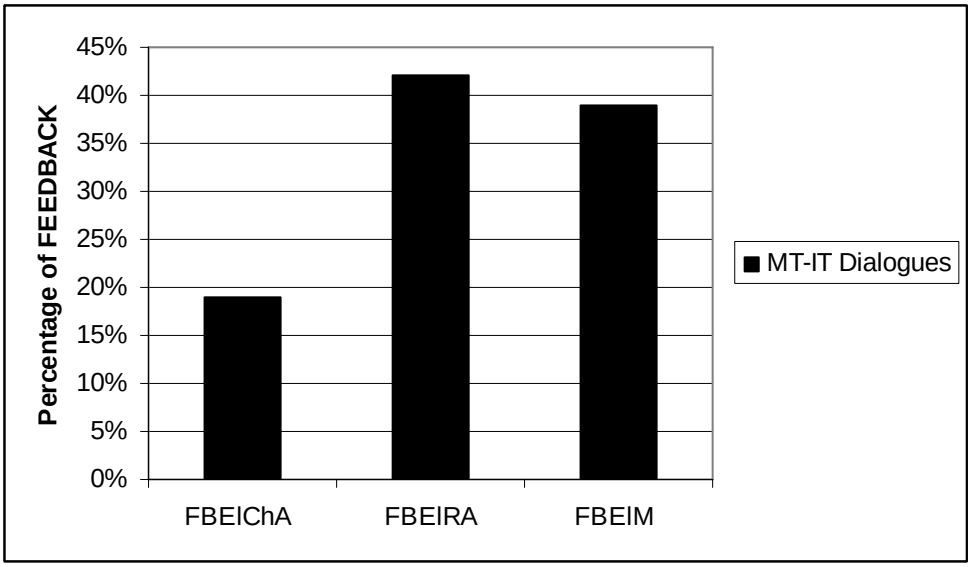

Figure 5.3 Distribution of FEEDBACK ELICIT expressions per specific semantic-pragmatic category in the two MT-Italian dialogues

In the Swedish MT-dialogues, feedback has been assigned the direction ELICIT only eight times, and in all these cases with the specific function of REQUIRING ACCEPTANCE (**FB** El **RA**)**.** REQUIRING ACCEPTANCE is done either by using short expression such as *eller?* (or?) *vad?* (what?) or by means of prosodic cues, producing utterances with an interrogative intonation.

One instance of the FEEDBACK WORD *no,* in Italian, produced with an interrogative intonation at the end of a contribution in order to elicit feedback, is shown in example 7, form MT-IT Dial 2.

$F52:	*in basso o in alto?*
$G53:	*quasi a chiudere il [1cerchio]/*
$F54:	***[1aha]*** **<FB;W;Gi;Ex>**
$G55:	***partendo dal basso/ no?*** **<FB;W;El; RA>**
$F56:	***mhm*** **/<FB;W;Gi;A>**

Example 7 from MT-IT Dial 2: no used to elicit FEEDBACK.

In example 4 in the contribution $F52 the follower asks a clarification question about a direction: *in basso o in alto*? (upward or downward)? The giver, as a reaction, explains: *quasi a chiudere il cerchio / partendo dal basso* (almost as if closing the circle starting from below) and ends his instruction with a *no* produced with a F0 contour typical for questions. This *no* has the aim of eliciting a feedback, which comes in $F54 with the production of *mhm*.

The first feedback expression produced by the follower in contribution $F54 is *aha*, overlapped with *cerchio*. This feedback expresses surprise, while the feedback expression *mhm* in $F56 signals understanding,

acceptance of the instruction received and therefore it has been labelled as FEEDBACK GIVE ACCEPTANCE.

5.4 Acoustic Characteristics of feedback

The same feedback expression can be used in different contexts to convey different functions; the different functions are expressed by means of F0 and duration variation and different intensity. To investigate how different functions are expressed by means of variation in the acoustic characteristics, some acoustic analyses have been performed on a sub-set of the short expressions identified in the map-task dialogues.

Previous studies have shown that short feedback expressions can undergo some intentional variation of F0 [Cerrato 1999, 2002c, Gardner 2001, Lindahl 2001, Caspers 2003]. Therefore it seems plausible to hypothesize that the specific communicative function of short expressions is reflected by their prosodic characteristics.

In order to test this hypothesis, F0 contour, intensity and some duration measurement were performed on the most frequent and comparable feedback expressions encountered in the four map-task dialogues: *sì* and m-like words in Italian and *ja* and m-like words in Swedish.

Segmentation and measurement of the duration and F0 contour of the short expressions and observation of the intensity of the feedback words were carried out with the help of WaveSurfer [Sjölander & Beskow 2000]. The measurement of temporal values was performed both from spectrograms and waveforms adopting the following criteria for segmentation: for the feedback expressions produced in a contribution of their own, which means that they were produced between pauses, the onset was set at the appearance of energy, and the offset was marked at the disappearance of energy.

When the short expression was coarticulated with preceding or following items, it was decided to include the transitions in the segmentation, in order to follow the same criterion for the segmentation of the items and the measurement of the duration.

The acoustic analysis also comprised pitch contour, which was analysed in terms of rising, flat and falling contour. The intensity was compared relatively intra speaker, per semantic-pragmatic category.

5.4.1 Italian sì

A total of 55 *sì* selected in the e Italian map-task dialogues were analysed. The instances of *sì* were uttered by the two female speakers in the dialogues. According to the second annotation made by the expert coder (i.e. the author) s*ì* is usually produced as FEEDBACK GIVE CONTINUATION and FEEDBACK GIVE ACCEPTANCE, however it can be also used as a POSITIVE

ANSWER to a polar question (this occurs 22 times in the two MT Italian dialogues, and of these 22 times it was possible to analyse 14 instances). When *sì* is produced on its own, in a non-intrusive way, to show CONTINUATION YOU GO ON without the intention to take the floor, it typically shows a rising F0 contour, a low intensity, and an average duration of 300 msec. When produced in a minimal non-intrusive way, it stands usually on its own, either during a short pause, or in overlap with the contribution of the interlocutor.

When *sì* is used with a CONTINUATION I GO ON function it usually shows a rising F0, but differently from the non-intrusive *sì* with a CONTINUATION YOU GO ON function, it is produced at the beginning of a longer utterance.

When *sì* is produced with the function of giving ACCEPTANCE it is usually produced on its own and shows a falling pitch contour and a high intensity. This falling pitch contour in the Neapolitan variant of Italian is typical at the end of a statement [D'Imperio 2002].

Sometimes *sì* can be lengthened (up to 40% longer), and produced with a falling-rising contour. In this last case the explicit function of *sì* is FEEDBACK ELICIT MORE INFORMATION.

In the Italian dialogues *sì* is sometimes reduplicated. Reduplication strengthens the explicit function of ACCEPTANCE, almost as to convey the feeling of satisfaction with the information received.

When *sì* is produced as a POSITIVE ANSWER to a yes/no question it shows a falling pitch contour.

Table 5.11 shows the results of duration analysis for all the occurrences of *sí* produced by the two female speakers in the Italian MT-IT Dial 2.

Table 5.11 Average and standard deviation of the duration of "sì" produced by the two Italian female speakers in MT-IT Dial 2.

Semantic- Pragmatic Function	**Number of analysed items**	**Duration (msec)**	**Standard Deviation**
FEEDBACK GIVE CONTINUATION YOU GO ON	20	320	59
FEEDBACK GIVE CONTINUATION I GO ON	6	185	38
FEEDBACK GIVE ACCEPTANCE	15	250	46
POSITIVE ANSWER	14	260	59

The results show that when *sì* is produced as a FEEDBACK CONTINUATION YOU GO ON it has a longer duration than when it is produced with any other function. The results of a one-factor analysis of variance, with function as independent variable, show that the duration difference was significant for

the three feedback categories [$F(3, 55 = 8.99; p<0.05)$]. A post-hoc analysis (Tukey, confidence interval = 0.05) revealed that the only significant differences were indeed between FEEDBACK GIVE CONTINUATION YOU GO ON and the other two feedback categories. This longer duration can be explained given the fact that the typical F0 contour of FEEDBACK GIVE CONTINUATION YOU GO ON items in Italian is that of a continuation rise, with a rising F0, which is usually coupled to a longer duration [Bolinger 1989].

The most evident difference in duration is between *sì* as FEEDBACK GIVE CONTINUATION YOU GO ON (FBGiCY) and FEEDBACK GIVE CONTINUATION I GO ON (FBGiCI). The latter show short duration and this might be due to the fact that *sì* with FEEDBACK GIVE CONTINUATION I GO ON in the analysed Italian map-task dialogues is always produced at the beginning of a longer utterance, with the intention to obtain the floor as quickly as possible.

5.4.2 Swedish ja

A total of 40 *ja* selected in the e Swedish map-task dialogues were analysed. They were uttered by the two female speakers in the dialogues. According to the second annotation made by the expert coder, Swedish *ja* is also usually produced as FEEDBACK GIVE CONTINUATION and FEEDBACK GIVE ACCEPTANCE, and it can also be used as a POSITIVE ANSWER to a polar question, even if this happens seldom in the Swedish MT dialogues (only five times).

When *ja* has a FEEDBACK GIVE CONTINUATION YOU GO ON function, without the intention to take the floor, it is mostly produced on its own, with low intensity, with a rising F0 contour and an average duration of 338 msec. When it is used as FEEDBACK GIVE CONTINUATION I GO ON**,** that is when listener's wish to show continuation of attention and interest, but also take the floor, it is usually produced at the beginning of a longer utterance, with a rising F0 contour.

Table 5.12 shows the average duration and standard deviation for all the *ja* produced by the two female speakers in the Swedish MT dialogues. The results show that *ja* with a FEEDBACK GIVE CONTINUATION I GO ON show longer durations in the Swedish map-task dialogues compared to Italian *sì* with the same function.

The high value of the duration and standard deviation for the category FEEDBACK GIVE CONTINUATION I GO ON is due to the fact that one of the female speakers in the Swedish MT dialogues produces a lengthening of the vowel [ɑ] to hold her turn. This results in an average duration of 458 msec for her production of *ja* with FEEDBACK GIVE CONTINUATION I GO ON.

Swedish *ja* has several variants, and is often produced in a reduced form with a dropping of the initial sound, and/or a lengthening of the final vocalic sound (up to 40% longer). It can be realised as a disyllabic word [jɑhɑ]

produced with a pitch contour which is flat at the beginning, then rising-falling. In this last case the explicit function is EXPRESSIVE, mainly to express surprise.

In the analysed Swedish map task dialogues *ja* is rarely reduplicated, but it is reinforced by adverbs such as: *precis, visst, just.*

Ja as a positive answer to a yes/no question is produced partly with a steady intonation, shifting into a slight rise.

*Table 5.12 Average and standard deviation of the duration of "ja" produced by the two Swedish female speakers in the MT dialogue*s.

Semantic- Pragmatic Function	**Number of analysed items**	**Duration (msec)**	**Standard Deviation**
FEEDBACK GIVE CONTINUATION YOU GO ON	15	338	48
FEEDBACK GIVE CONTINUATION I GO ON	9	397	132
FEEDBACK GIVE ACCEPTANCE	8	327	49
POSITIVE ANSWER	8	303	45

5.4.3 Comparative Results

Table 5.13 shows a schema of the similarities and the differences between F0 contour for Swedish *ja* and Italian *sì* for the different functions. The function CONTINUATION YOU GO ON is characterized, both in Italian and Swedish, by a rising pitch contour, which is a typical continuation contour. A raised F0 is considered a marker of non assertiveness [Ohala 1983] and in fact the feedback category CONTINUATION YOU GO ON signals continuation of attention, but not acceptance or agreement, which is instead signalled by the feedback expressions assigned to the category ACCEPTANCE.

In the analysed dialogues Swedish *ja* with a CONTINUATION I GO ON function is characterized by a rising F0 and Italian *sí* by a falling F0. The falling contour is typical for assertiveness and categoricalness [Kohler 2004] which is quite consistent with the realization of Italian *sí* having a rising F0 even as a POSITIVE ANSWER.

In Swedish instead the categories ACCEPTANCE and POSITIVE ANSWER are realized with a slightly rising F0. Rising F0 can signal non-assertiveness and uncertainty, therefore it is typical of question intonation.

Moreover Kohler [2004] and House [2005], analysing respectively German and Swedish material, found that final rises can pragmatically signal intended social interaction and friendliness. This pragmatical-attitudinal explanation might be the key to understand the different contours

shown by Italian and Swedish for the categories ACCEPTANCE and POSITIVE ANSWER. This difference might also be linked to the cultural difference, which depicts Italian people as being more assertive, categorical and self-confident in expressing their points of view (acceptance, agreement) and in giving their responses, while Swedish people as being oriented to seek consensus, by not showing self-confidence and categoricalness, hence by using a rising pitch contour which denotes uncertainty, openness towards the addressee and friendliness.

When *ja* is produced with the function of FEEDBACK GIVE ACCEPTANCE it shows a pitch contour which is in average rising (+80 Hz), but in fact the F0 rises at first (+100Hz) and then towards the end slightly falls (-60 Hz). This final fall in Stockholm Swedish is an uncontroversial completion signal, which is appended to the sentence accent rise [Grönnum 1991].

An interesting observation is that when FEEDBACK WORDS are produced with the function GIVE CONTINUATION YOU GO ON they are produced with the lowest relative intensity by all the speakers, while the expressions produced as POSITIVE ANSWER, in isolation, are produced with the highest intensity.

The prosodic phenomena observed in this investigation show that F0 contour seems to play an important role for assigning a specific meaning to the expressions used. However, since some of the functions are characterized by the same F0 contour, it is important to also take duration into account when interpreting the different semantic-pragmatic functions of feedback expressions. Duration differences resulted to be significant for Italian *sí* uttered with the function FEEDBACK GIVE CONTINUATION YOU GO ON, while the duration of Swedish *ja* did not show any relevant difference across semantic-pragmatic categories.

Table 5.13 Schema of the comparison between the F0 contours of "sì" in Italian and "ja" in Swedish according to the different functions.

Function	SWEDISH F0 contour	ITALIAN F0 contour
CONTINUATION YOU GO ON (CY)	rising + lengthening	rising + lengthening
CONTINUATION I GO ON (CI)	flat or rising	falling
ACCEPTANCE (A)	flat, slightly rising, or rising-falling	falling
POSITIVE ANSWER (RP)	flat or slightly rising	falling
EXPRESSIVE (EX)	varying +lengthening	varying + lengthening

5.4.4 Swedish m-like words

In the Swedish MT dialogues several occurrences of m-like words appear. The total number of m-like words identified in the Swedish MT dialogues is 58, of which 24 items in MT-SW Dial 1 and 34 in MT-SW Dial 2. These represent 30% of the total production of short expressions. Of these m-like words, fewer than 8% are used for communicative function other than FEEDBACK**,** mostly HESITATION.

All the m-like words were labelled as FEEDBACK GIVE. There are no instances of m-like words produced to elicit feedback. Table 5.14 shows the distribution of m-like words per communicative function in the Swedish map-task dialogues.

Table 5.14 Distribution of m-like words per function in the Swedish dialogues.

Dialogue	CONTINUATION		ACCEPTANCE/ NON ACCEPTANCE	EXPRESSIVE	OTHER
	YOU GO ON	I GO ON			
MT-SW Dial 1	8	5	6	2	3
MT-SW Dial 2	18	2	6	6	2

The acoustic analysis of F0 contour of the m-like words shows that it is possible to relate a prototypical F0 contour and other acoustic characteristics to specific communicative functions. Table 5.15 shows the prototypical F0 contour related to the different functions carried out by m-like words.

Table 5.15 Prototypical F0 contour related to different functions of the m-like words.

Category	**F0 contour**
CONTINUATION YOU GO ON, ACCEPTANCE	slightly rising (+50 Hz)
CONTINUATION I GO ON	rising (or flat)
EXPRESSIVE{SURPRISE}	falling-rising
EXPRESSIVE {DOUBT}	flat

These results are in agreement with those obtained in a comparable analysis performed on similar materials in Swedish [Lindahl 2001], where "m-feedback morpheme", which is the equivalent of the m-like words here presented, is reported to have six different F0 contours depending on the different communicative intentions of the speaker.

The main difference in the F0 contour seems to be between m-like words with flat contour and m-like words with falling-rising or rising contour. Flat contour is typical when the function of the m-like is EXPRESSIVE, with an attitudinal reaction showing hesitation. Falling-rising contours are typical for EXPRESSIVE function with an attitudinal reaction of surprise.

A rising contour is typical for m-like words produced as FEEDBACK GIVE CONTINUATION I GO ON and FEEDBACK GIVE ACCEPTANCE. Most of the m-like words encountered in the Swedish MT dialogues have been labelled as FEEDBACK GIVE CONTINUATION YOU GO ON (CY). In only seven cases do speakers start their contribution by producing an m-like word, which was labelled as FEEDBACK GIVE CONTINUATION I GO ON (CI).

One instance of an m-like word produced at the beginning of a contribution with a FEEDBACK GIVE CONTINUATION I GO ON function is shown in example 8 for Swedish. This m-like word is characterized by a flat F0 contour and duration of 215 msec.

A flat F0 contour is quite atypical for a continuer. Usually continuers show rising contours, while flat contours are more typical for expressions produced to show doubt or hesitation. In this case the m-like word might in fact be interpreted both as HESITATION and FEEDBACK GIVE CONTINUATION I GO ON. After the initial short m-like word, the speaker holds the turn and asks for a clarification: *hur gör jag med krabborna där då? (*how do I do with the crabs there then?)

$G36:	*så du håller dig emellan den å eeh konturen och buktens kontur*
$F37:	***mm/*** **<FB;W;Gi;CI>** *hur gör jag med krabborna där då?*

Example 8 from MP SW Dial 2: m-like word labelled as FEEDBACK GIVE CONTINUATION *I GO ON*

Figure 5.4 shows a bar diagram of the average duration, in milliseconds, of m-like words per semantic-pragmatic function.

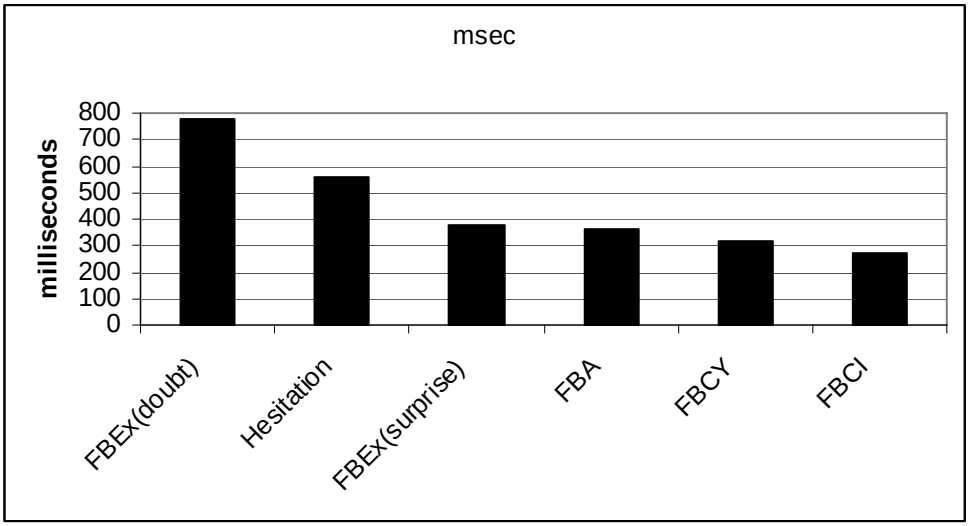

*Figure 5.4 Average duration in msec of m-like words per function in the two Swedish map-task dialogue*s.

There is an evident difference in these results between short and long m-like words. Short m-like words are usually produced with FEEDBACK GIVE CONTINUATION and ACCEPTANCE functions. Longer m-like words are produced when the function is FEEDBACK GIVE EXPRESSIVE, which means that the speaker intentionally adds some extra information to feedback, for instance, doubt/hesitation, surprise. In the case of expression of surprise, the m-like word shows a disyllabic structure with longer duration and a falling-rising F0 contour.

Figure 5.5 shows a disyllabic m-like word labelled as FEEDBACK GIVE EXPRESSIVE where the oscillating F0 contour conveys the expression of surprise. Figure 5.6 shows a typical monosyllabic m-like word in Swedish, uttered by a female speaker and labelled as FEEDBACK GIVE CONTINUATION YOU GO ON.

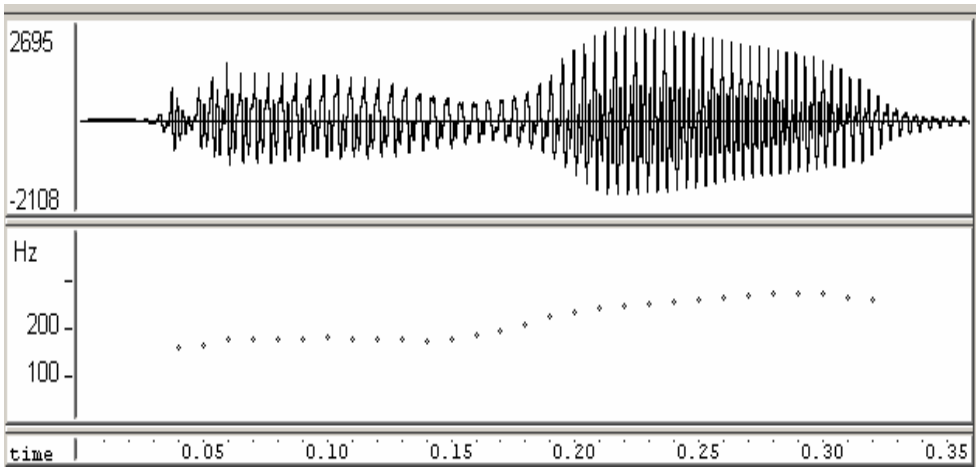

Figure 5.5 Waveforms and F0 contour of a disyllabic m-like word in Swedish with oscillating F0 contour, uttered by a female speaker.

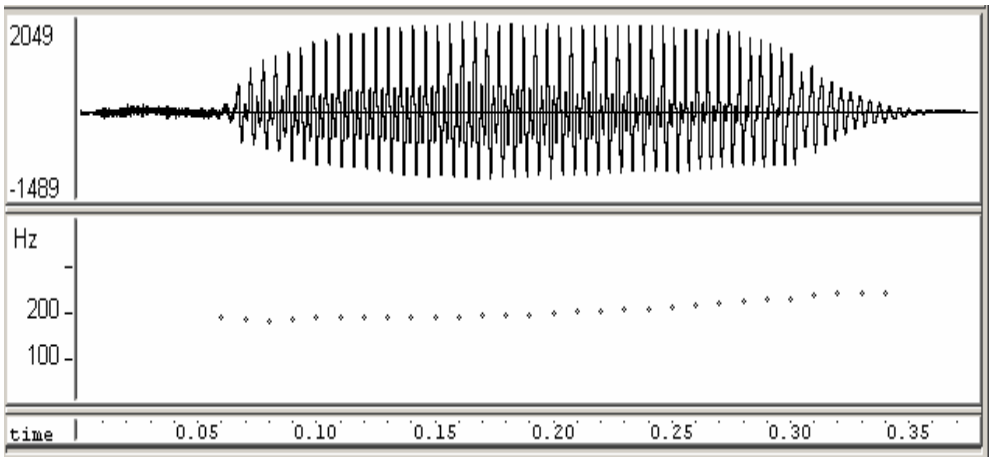

Figure 5.6 Waveform and F0 contour of a typical monosyllabic m-like word in Swedish, uttered by a female speaker with FEEDBACK CONTINUATION YOU GO ON *function.*

M-like words can also be produced when speakers wish to express some doubt, scepticism about the information received. In these cases m-like words show longer duration and flat pitch contours. However they differentiate from those words produced as HESITATION, which also show longer duration and flat pitch contour, because hesitations are produced as nasalized vowels, and could be orthographically transcribed rather as *em, ehm* than *mm*.

5.4.5 Italian m-like words

In the Italian MT dialogues, the m-like words are less frequent compared with the Swedish MT dialogues. The total number of m-like words

identified is 27, of which 20 in MT-IT Dial 2, and 7 in MT-IT Dial 1. This represents 13% of the total production of FEEDBACK WORDS.

M-like words were produced either as FEEDBACK GIVE or as HESITATIONS; there are no instances of m-like words produced to elicit feedback in the Italian MT Dialogues.

Table 5.16 shows the distribution of m-like words per communicative function.

Table 5.16 Distribution of m-like words per function in the Italian dialogues.

Dialogue	CONTINUATION		ACCEPTANCE	EXPRESSIVE	OTHER
	YOU GO ON	I GO ON			
MT-IT Dial 1	1	-	1	3	2
MT-IT Dial 2	7	1	10	1	1

The acoustic analysis of F0 contour of the m-like words shows that it is possible to relate a prototypical F0 contour to specific communicative functions. The main trends are reported in table 5.17.

Table 5.17 Prototypical F0 contour related to different functions of m-like words in Italian.

Category	F0 contour
CONTINUATION YOU GO ON	rising
CONTINUATION I GO ON	falling (just one item)
ACCEPTANCE	falling
EXPRESSIVE{SURPRISE}	no instances in the data
EXPRESSIVE{DOUBT}	flat

The main difference in the F0 contour seems to be between m-like words with rising contour, used as FEEDBACK GIVE CONTINUATION YOU GO ON and those with falling contour used mainly to signal FEEDBACK GIVE ACCEPTANCE.

The difference between feedback expressions with CONTINUATION YOU GO ON and ACCEPTANCE function is quite marked in Italian, while in Swedish these two functions can show similar pitch contours. However, both in Italian and Swedish, feedback expressions serving the function CONTINUATION YOU GO ON are characterized, by a rising pitch contour, which is a typical continuation contour.

In general the m-like words in the Italian MT dialogues show shorter durations compared to the Swedish m-like words, however the relative durations per category seem to follow the same trend. In Italian items labelled as FEEDBACK GIVE CONTINUATION YOU GO ON have a longer average duration than those labelled as FEEDBACK GIVE ACCEPTANCE. This is due to the fact that CONTINUATION is signalled in Italian by a lengthening of the final vowel and rise at the same time. Also in Italian the m-like words

produced as HESITATION are much longer compared to the other functions as shown in figure 5.7.

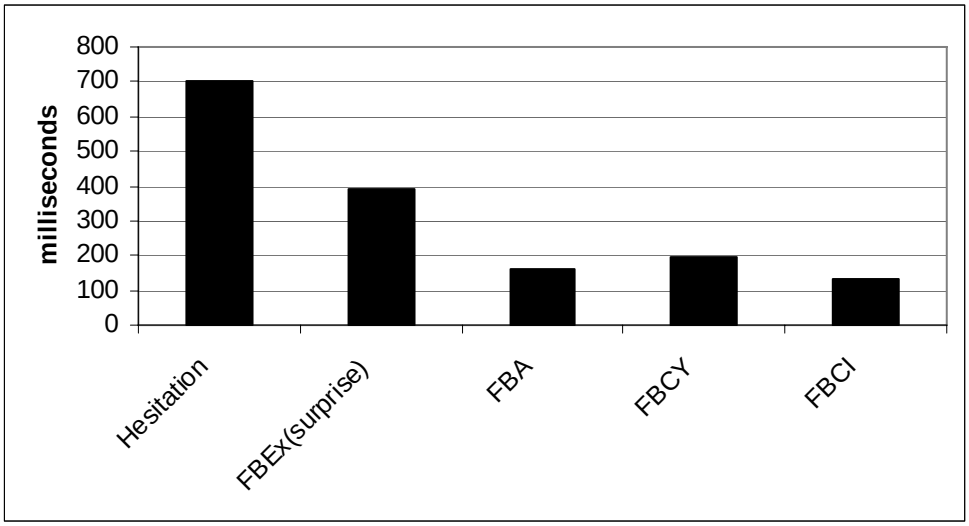

Figure 5.7 Average duration of m-like words per category in Italian.

5.4.6 Overlaps and Pauses

One of the most common prejudices about Italians is that "they all talk at the same time". This prejudice seems to be confirmed by the fact that in the Italian MT dialogues, feedback expressions are often produced in overlap with the main speaker contribution. In order to quantify this phenomenon, the frequency of overlapping of the feedback expressions with the contribution of the other speaker has been calculated.

The results show that the percentage of times when the feedback expression is produced in overlap with the current speaker's contribution is much higher in the Italian dialogues than in the Swedish ones, as reported in table 5.18.

Table 5.18 Number and percentage of overlapped feedback expressions per dialogue.

Dialogue	**Overlapped feedback** (instances)	**Overlapped feedback** (percentage)
MT-IT Dial 1	18 of 85	21%
MT-IT Dial 2	73 of 167	44%
MT-SW Dial 1	10 of 96	10%
MT-SW Dial 2	14 of 129	11%

The average duration of the pauses between a speaker's contribution and the following feedback (when not overlapped) has also been measured. The silent pause has been measured starting at the end of the contribution of one speaker and finishing at the beginning of the feedback expression of the

other speaker. The results show that the Italian pauses are shorter than the Swedish pauses and the difference is as high as 30%, as reported in table 5.19.

Table 5.19 Average duration in milliseconds of the duration of the silent pause before the feedback expression in each dialogue.

Dialogue	**Average duration** (msec)
MT-IT Dial 1	280
MT-IT Dial 2	255
MT-SW Dial 1	360
MT-SW Dial 2	385

The fact that Italian dialogue participants continuously overlap, interrupt and anticipate each other should not be interpreted as a sign of impoliteness, but rather a sign of co-operation. In fact overlapping sentences and anticipation are seen as markers of involvement [Jefferson 1973; Tannen 1984].

In particular the Neapolitan-Italian dialogues used in this study have been defined, in another investigation, [Castagneto & Ferrari 2003], as "high-involvement" dialogues. "High-involvement" style is considered as the opposite of "high-considerateness" style, which is characterized by the interlocutors focusing on the task and on the message they receive or deliver, rather on the relationship between them, which instead is the focus of "high-involvement" style. In "high-considerateness" style the interlocutors show more distance and tend to follow turn-management rules. In "high-involvement" style the distance between the interlocutors is shortened due to the fact that the focus is on their relationship rather than on the actual task. This might have as a consequence the production of overlapping contributions and anticipation of the contribution of the other speaker.

5.4.7 Perceptual Test: Design

The results of the acoustic analysis of feedback words have shown that it is possible to relate a prototypical F0 contour and duration characteristics to a specific communicative function. On the basis of these results it is possible to hypothesize that duration and pitch contour can be regarded as cues to the interpretation of the semantic-pragmatic function that feedback words carry out, even when they are excerpted from their original context.

To test this hypothesis a perceptual test was designed and run using a sub-set of the data analysed in this chapter [Cerrato 2005a]. The aim of the test was to verify whether the prosodic cues alone can be used to distinguish different feedback functions of the same feedback word, namely *sì* in Italian and *ja* in Swedish, when this is presented out of its context.

All the items used for the test were selected from those that had been identified as FEEDBACK and annotated taking contextual information into account. The selected items/stimuli were segmented from their context and no manipulations were performed on them, in order to preserve their naturalness.

Throughout the whole thesis it has been underlined that in order to identify feedback phenomena it is crucial to take contextual information into account, since feedback is interpreted and categorised in terms of reactions to the previous communicative act. In the perceptual test, it was decided instead, that the subjects should not have access to contextual information, since the aim of the test was to verify whether the prosodic cues alone can be used to distinguish different feedback functions carried out by the same feedback word. For this reason the subjects could only listen to the feedback word excerpted from the original context.

The stimuli consisted of 32 Italian *sì* and 32 Swedish *ja* uttered by two speakers per each language.

The stimuli were divided per categories, so that for each of the following category, eight stimuli were played:

- FEEDBACK GIVE CONTINUATION YOU GO ON (**FB**Gi**CY**),
- FEEDBACK GIVE CONTINUATION I GO ON (**FB**Gi**CI**),
- FEEDBACK GIVE ACCEPTANCE (**FB**Gi**A**),
- POSITIVE ANSWER (**R**EPLY **P**OSITIVE) (**RP**).

For the categories POSITIVE ANSWER and FEEDBACK GIVE CONTINUATION I GO ON there were not enough instances of stimuli per speaker in the dialogues, hence some of them were played twice.

The stimuli were organised in two blocks of 34, the first two stimuli in each block being dummies.

The test consisted of two sub-tests, one with the Italian stimuli presented to 8 Italian listeners and one with the Swedish stimuli presented to ten Swedish listeners.

Before the experimental session the participants were given written instructions and were involved in a short training session to familiarise with the task. Their task was to listen to the stimuli, which were presented individually over headphones, in randomized order, and after each presentation choose, on the answering sheet, which function they believed the stimulus carried out in the original conversation.

The possible functions with relative labels were:

- FEEDBACK GIVE CONTINUATION YOU GO ON (**FB**Gi**CY**),
- FEEDBACK GIVE CONTINUATION I GO ON (**FB**Gi**CI**),
- FEEDBACK GIVE ACCEPTANCE (**FB**Gi**A**),
- POSITIVE ANSWER (**R**EPLY **P**OSITIVE) (**RP**).

5.4.7.1 Results for Swedish Stimuli

The results for the ten Swedish listeners judging the *ja* of the two Swedish speakers are shown in table 5.20a and 5.20b. For the Swedish stimuli, not all the recognition rates appear to be above chance level.

FEEDBACK GIVE ACCEPTANCE (**FB**Gi**A**) and FEEDBACK GIVE CONTINUATION YOU GO ON (**FB**Gi**CY**) are confused with each other. However the items of the category FEEDBACK GIVE CONTINUATION YOU GO ON (**FB**Gi**CY**) get better recognition rates than those in the category ACCEPTANCE (**FB**Gi**A)**. This result seems to be consistent with the results of the acoustic analysis of Swedish *ja* (in section 5.4.2), which showed that both the functions FEEDBACK GIVE CONTINUATION YOU GO ON and FEEDBACK GIVE ACCEPTANCE can be characterized by a rising pitch contour. The category POSITIVE ANSWER (**RP**) is also much confused and this can also be due to the fact that in Swedish this category shows the same pitch contour as FEEDBACK GIVE ACCEPTANCE.

The only category which gets high recognition rates is CONTINUATION I GO ON (**FB**Gi**CI**), in particular for speaker 1, and this might due to the longer duration of the items belonging to this category for this speaker.

Table 5.20a Confusion matrix for the identification test for Swedish speaker 1.

STIMULI	RESPONSES (%)			
	FBGi**A**	**FB**Gi**CI**	**FB**Gi**CY**	**RP**
FBGi**A**	**48**	2	30	20
FBGi**CI**	2	**83**	14	0
FBGi**CY**	33	2	**55**	11
RP	38	0	38	**23**

Table 5.20b Confusion matrix for the identification test for Swedish speaker 2.

STIMULI	RESPONSES (%)			
	FBGi**A**	**FB**Gi**CI**	**FB**Gi**CY**	**RP**
FBGi**A**	**41**	9	40	9
FBGi**CI**	29	**52**	2	17
FBGi**CY**	33	5	**59**	3
RP	23	31	19	**27**

5.4.7.2 Results for Italian Stimuli

The results, in the form of confusion matrices, for the Italian listeners judging the *sì* of the two Italian speakers, are shown in table 5.21a and 5.21b. For the Italian stimuli all the recognition rates appear to be above chance level.

The stimuli coded as CONTINUATION YOU GO ON (**FB**Gi**CY**) for Italian speaker 1 and 2 get high recognition rates, and this maybe due to the typical lengthening phenomenon that characterises the items assigned to this category.

The category ACCEPTANCE (**FB**Gi**A**) and POSITIVE ANSWER (**RP**) are confused with each other. This might depend on the fact that they have similar acoustic characteristics, in particular similar pitch contour and duration and even similar tonal contour [Cerrato & D' Imperio 2003]. The only difference consisting in the higher intensity of POSITIVE ANSWER (**RP**) stimuli.

The feedback categories CONTINUATION YOU GO ON (**FB**Gi**CY**) and ACCEPTANCE (**FB**Gi**A**) are very seldom confused with each other in the Italian stimuli, and this result is consistent with the outcome of the acoustic analysis in section 5.4.1 which showed different pitch contours and different durations for the items belonging to these two categories.

Table 5.21a Confusion matrix for the identification test for Italian speaker 1

	RESPONSES (%)			
STIMULI	**FB**Gi**A**	**FB**Gi**CI**	**FB**Gi**CY**	**RP**
FBGi**A**	**48**	8	2	42
FBGi**CI**	16	**59**	5	20
FBGi**CY**	2	5	**90**	3
RP	38	11	6	**45**

Table 5.21b Confusion matrix for the identification test for Italian speaker 2

	RESPONSES (%)			
STIMULI	**FB**Gi**A**	**FB**Gi**CI**	**FB**Gi**CY**	**RP**
FBGi**A**	**45**	13	5	37
FBGi**CI**	14	**59**	5	22
FBGi**CY**	9	8	**69**	14
RP	37	14	11	**38**

5.4.8 Perceptual Test: Comparative Results

The results of the perceptual test show that the acoustic characteristics of *ja* in Swedish and in particular of *sì* in Italian, extracted from their context, reflect the semantic-pragmatic functions that the short expressions carried out in the given context. Duration cues together with pitch contour

characteristics seems to be very helpful in the perceptual test for the recognition of the specific semantic-pragmatic function carried out in particular by the Italian *sí* excerpted from their original context. The good recognition scores in the perceptual test with the Italian stimuli is quite remarkable considering the difficulty of the perceptual task, in which the subjects had to recognize which semantic-pragmatic function the feedback word carried out in the original context, by only listening to the feedback word.

The confusions showed in the recognition scores for the Swedish stimuli belonging to the categories FEEDBACK GIVE CONTINUATION YOU GO on and FEEDBACK GIVE ACCEPTANCE can also be interpreted as an indication of the fact that duration cues together with pitch contour characteristics can be helpful to recognize the function carried out by the feedback word. Indeed the Swedish *ja* with FEEDBACK GIVE CONTINUATION YOU GO on and FEEDBACK GIVE ACCEPTANCE function, used as stimuli in the perceptual test, were not characterized by the marked acoustic characteristics differences, as were the Italian stimuli belonging to the same categories.

5.5 Conclusions and Discussion

The results of this comparative study of verbal feedback expressions in Italian and Swedish dialogues elicited with the map-task technique show that it is possible to categorise feedback according to their type, direction, and the semantic-pragmatic function they carry out in the given context, using the coding scheme developed for the purpose of annotating feedback phenomena.

The coding scheme used in this study has shown satisfactory results in the inter-coder reliability test (reported in section 5.2), which means that the feedback categories can be considered appropriate both for Italian and Swedish verbal feedback expressions.

That the results of the stability test might have been biased by the fact that the expert coder is also the developer of the coding scheme is a difficult shortcoming to avoid, since often empirical work in linguistics builds on the subjective judgements of the researchers themselves. According to Carletta [1996], in subjective coding tasks no coder can be considered as an expert, since subjectivity cannot be avoided anyway.

Undeniably when dealing with annotations and application of theoretical concepts such as categories, it is important to understand the subjective nature of the phenomena that are being coded, and also accept the fact that it might not be possible to obtain substantial or perfect scores at all. For this reason the results obtained in the reproducibility test can be considered as positive results indeed.

The results of the distributional analysis show, as expected, that the most common functions conveyed by feedback expressions in the analysed MT dialogues are GIVE CONTINUATION YOU GO ON and GIVE ACCEPTANCE.

These two functions are considered to be those that most effectively contribute to the smooth and effective unfolding of task-oriented interactions. In map-task dialogues the participants have the specific task to succeed in drawing the correct route on the map of the follower. To do this in an effective way they have to cooperate, and one important means of cooperation is the production of effective feedback signals that facilitate the accomplishment of the task and ensure the smooth unfolding of the interaction. The feedback functions GIVE CONTINUATION YOU GO ON and GIVE ACCEPTANCE are often signalled by means of words such as *yes, no, mm, ah* and so on, which are short and effective.

However besides feedback words, feedback phenomena under the form of repetitions and reformulations have also been observed in the analysed data. The role of repetitions and reformulations is to give, or elicit feedback in a marked way and at the same time to help in the cognitive process of acquisition of information. This means that repetitions and reformulations of the received instructions help the followers in the map-task to "think aloud" while trying to accomplish their task of drawing the route on the map.

As for the acoustic analysis, the results suggest that acoustic characteristics of feedback words, such as duration and F0 contour, reflect the semantic-pragmatic function they carry out and the communicative intention they convey.

Different F0 contour and durations are produced in the two languages to express the same function, in other words Italian and Swedish dialogue participants seem to have different conventions for the use of feedback expressions and turn-management rules. This depends of course on the fact that people belonging to different cultural communities have different norms, expectations and procedures which affect the way they behave when taking part in a conversation.

One of the most evident differences across the two languages concerns the realization of F0 contour for the feedback words having a GIVE ACCEPTANCE function. These are characterized by a flat or slightly rising F0 in the Swedish dialogues and by a falling F0 in the Italian ones.

This dissimilarity could be interpreted by taking cultural differences into account. In fact Italian people are thought to be more assertive, categorical and self-confident in expressing their points of view and in giving their responses, compared to Swedish people who are instead stereotypically depicted as being oriented to seek consensus, and not showing self-confidence and categoricalness.

In this perspective the categoricalness and assertiveness of the Italian feedback words with GIVE ACCEPTANCE function is signalled by a falling F0, which is typical for assertiveness and categoricalness [Kohler 2004] while

the non-assertiveness and non-categoricalness of the feedback words with GIVE ACCEPTANCE function in Swedish is mostly signalled by a slightly rising F0, which typically indicates non-assertiveness and uncertainty [Kohler 2004; House 2005].

Another common prejudice about Italians is that "they all talk at the same time", which means that they do not seem to respect the turn-management rules. This prejudice seems to be confirmed by the fact that feedback expressions are produced in overlap with the main speaker contribution more often in Italian than in Swedish and that the pause before the feedback expression is much longer in Swedish than in Italian.

Swedish participants seem to be more "considerate" and seem to respect the conventional turn-management rules, while Italian participants seem to be more "involved" in the conversation [Castagneto & Ferrari 2003]. This involvement seems to make them "impatient", since some times they even finish the contribution of their interlocutor, by producing an anticipation feedback.

Even if it might be argued that the analyses were limited to a particular kind of communicative situation, namely map-task setting, and to only four speakers in two specific varieties of each language, it is evident from the results that acoustic characteristics of feedback words can be considered as cues to the interpretation of the dialogue functions they serve. This is confirmed by the results of the perceptual test that show that the acoustic characteristics of *ja* in Swedish and in particular of *sì* in Italian, extracted from their context, reflect the semantic-pragmatic functions that the short expressions carried out in the given context. The good recognition scores in the perceptual test with the Italian stimuli is quite remarkable considering the difficulty of the perceptual task, in which the subjects had to recognize which semantic-pragmatic function the feedback word carried out in the original context, by only listening to the feedback word.

Duration cues together with pitch contour characteristics can be therefore considered as helpful cues for the e recognition of the function carried out by the Italian *sí* excerpted from their original context.

It is therefore possible to conclude that the investigation of the prosodic marking of short expressions in relation to their communicative function is of fundamental interest when it comes to technological applications, in that it may help to interpret the communicative function intended by humans in interaction with computer dialogue systems, and therefore contribute to make human-machine interactions smoother.

Another important aspect of human communication that can be exploited to make human-machine interaction smoother is the production of non-verbal communicative behaviour. In the next chapters the production of non-verbal behaviour related to communicative feedback becomes therefore the focus of the investigations.

6 Feedback Phenomena in Spontaneous Human-Human Dialogues

6.1 Introduction

In chapter 5 verbal feedback phenomena have been categorised using the coding scheme developed to annotate feedback phenomena, which allows analysing feedback in terms of form and function.

The study presented in this chapter[16] shows evidence that it is also possible to categorise and analyse non-verbal feedback in terms of form and function. It is hypothesised that the specific categories provided in the coding scheme to code the semantic-pragmatic functions of feedback are independent of the modality in which feedback is expressed. Moreover, the investigation carried out in this chapter aims at exploring the realization of co-occurring verbal and non-verbal behaviour signalling feedback.

The materials analysed in this chapter consist of video recordings of four spontaneous dialogues between a travel agent and four customers recorded in a travel agency in Sweden. These materials not only represent a good testing ground for the appropriateness of the pre-defined categories for semantic-pragmatic functions to non-verbal feedback phenomena, but they also allow for the investigation of feedback as a multi-modal phenomenon.

Among non-verbal feedback phenomena, particular attention is paid here to facial displays, even if hand movements and other gestures can be produced to signal feedback. Facial displays include phenomena such as changes in eyebrow position, expressions of the mouth, movement of the head and eyes [Chovil 1992; Cassell 2000].

6.2 Materials

Video recordings of four real spontaneous dialogues between four different customers (two females and two males) and a travel agent (always the same woman), in a travel agency in Gothenburg, Sweden, were selected from the Spoken Language Corpus of the Linguistics Department of Gothenburg University –GSLC-- [Allwood et al. 2000].

[16] Part of this study was conducted together with Jens Allwood [Allwood & Cerrato 2003]

These dialogues were selected for several reasons: in the first place they consist of genuine spontaneous interactions recorded in a real environment: a travel agency in Sweden. They belong to the activity type of "factual information seeking", where the customer asks the travel agent for information about timetables, visas, hotels and so on, and the travel agent provides the information required. They are video recorded from a close distance from the dialogue participants, which makes them quite suitable for the analysis of facial displays related to feedback, in particular for the analysis of head movements, which is one of the main foci of this thesis.

In the video recordings it is possible to see the agent standing or sitting behind a desk, and the customer standing on the other side. The microphone and the video camera were placed on the desk, in a position which allowed the recording of both dialogue participants on one side. The customers were informed of the presence of the recording apparatus and of the purpose of the recording by means of a sign placed on the desk.

Table 6.1 shows some information related to the dialogues, and a more detailed table is reported in section 3.3.1.

In GSLC-Dial 1 and 3 the customers not only get several pieces of information from the travel agent, but they also book their trips. GSLC-Dial 2 is very short because the customer only asks for a specific piece of information, while GSLC-Dial 4 is the longest because of some problems occurring with the terminal during the booking of a flight.

Table 6.1 Schema of the information related to four dialogues used in this study.

Dialogue	**Customer**	**Number of contributions**	**Duration (minutes)**
GSLC-Dial 1	female	107	8.42
GSLC-Dial 2	male	65	2.15
GSLC-Dial 3	male	150	16.42
GSLC-Dial 4	female	112	27.31

6.3 Method

Feedback expressions were identified as reactions to the previous speech act and coded using the categories provided in the coding scheme developed in chapter 4.

Feedback expressions have been identified and coded with the support of Multitool [Allwood et al. 2002], a tool for audio-visual analysis, which simultaneously displays the video and the relative orthographic transcription and annotation of the dialogues, as shown in the screenshot reproduced in figure 6.1.

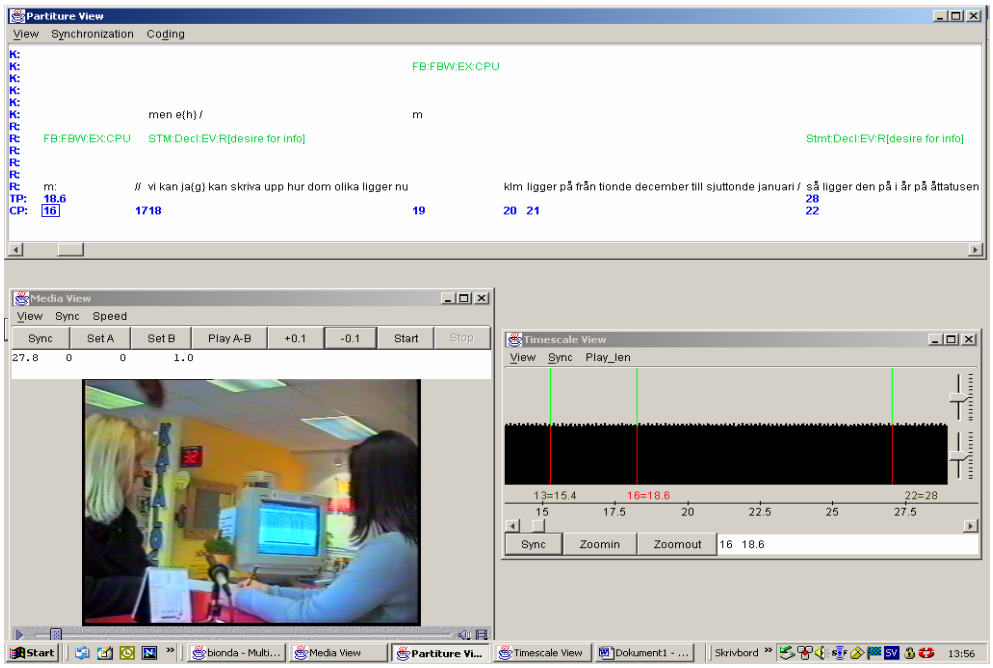

Figure 6.1 Screen-shot of Multitool, with the simultaneous display of the video recordings of the dialogue and the relative orthographic transcription and multi-tier annotation on the partiture.

The coding procedure starts from a first monitoring of the material under analysis and proceeds with the multi-tier annotation.

The annotator chooses the tiers that are most appropriate for the given materials. For this study the following tiers were displayed on the score lines of Multitool:

- **Text**: reports the orthographic transcription of the contributions for each speaker.
- **Comments**: reports different kinds of comments of the transcribers, which are not part of the coding scheme described below.
- **Speech Act**: reports the annotation of the speech act and for feedback also the direction.
- **Verbal FB**: reports the annotation of the type of verbal feedback expression
- **Non-verbal FB**: reports the annotation of the type of visible non-verbal feedback behaviour.
- **Function**: reports the semantic-pragmatic function of the feedback phenomenon under analysis.

- **Multi-modal relation**: reports the specific relationship between the verbal and the non-verbal expressions produced to express feedback.
- **Gaze**: reports the direction of the speakers' gaze using only two values: EYECONTACT, NON-EYECONTACT.

6.3.1 Coding Procedure and Coding Scheme

The coding scheme presented in chapter 4 was used to code the feedback expression. The annotation starts with the identification and annotation of the speech act (in the tier called Speech act). The speech act can be a STATEMENT, a QUESTION, a HESITATION or a FEEDBACK.

When a Speech Act is identified as FEEDBACK, it is then coded in terms of type. Feedback types can be verbal (WORDS, PHRASE and SENTENCE) and non-verbal (FACIAL DISPLAY, HAND MOVEMENT and OTHER).

For FACIAL DISPLAYS, higher degrees of complexity are taken into account, which means that a more detailed set of features is considered. FACIAL DISPLAYS can be further categorised and annotated by using more specific features concerning: GENERAL FACE, EYEBROWS MOVEMENTS, and HEAD MOVEMENTS and GAZE DIRECTION. (The categories are illustrated in chapter 4 section 4.3.2.1).

Since the results of the investigation reported in this chapter will show that the most frequent FACIAL DISPLAYS related to feedback are HEAD MOVEMENTS, the specific features for HEAD MOVEMENTS and their labels are here presented again in table 6.2.

HEAD MOVEMENTS can be categorised as:

- SINGLE NOD: a single head movement down-up.
- REPEATED NODS: multiple head movements down-up.
- SINGLE JERK: a single quick head movement up-down.
- REPEATED JERKS: multiple head movements up-down.
- SINGLE SLOW BACKWARDS UP: a single slow head movement backwards. (This movement is differentiated from single jerk on the basis of the velocity. The term jerk implies quickness; while a single slow backward up refers to a slow up-down movement.)
- MOVE FORWARD: is a movement of the head forward, this can either be a movement of the head only or can be a movement of the whole trunk.
- MOVE BACKWARD: is a movement of the head backward, this can either be a movement of the head only or can be movement of the whole trunk.
- SINGLE TILT (Sideways): a single movement of the head leaning on one side.

- REPEATED TILTS (Sideways): a multiple movement of the head leaning from side to side.
- SIDE-TURN: is a rotation of the head towards one side.
- SHAKE (repeated): is a repeated rotation of the head from one side to the other.
- WAGGLE: is a movement of the head back and forth, side to side, it is like a mixture of shake and move backward or forward it is usually produced to show uncertainty, doubtfulness.
- OTHER: either a different type of movement than the three mentioned, or a combination of two or more of them.

Table 6.2 Labels used to code the specific HEAD MOVEMENTS.

Type of non-verbal expression	
Category	**Labels**
SINGLE NOD (DOWN)	S-NOD
REPEATED NODS (DOWN)	R-NOD
SINGLE JERK (BACKWARDS UP)	S-JERK
REPEATED JERKS (BACKWARDS UP)	R-JERK
SINGLE SLOW BACKWARDS UP	BACKUP
MOVE FORWARD	FORWARD
MOVE BACKWARD	BACK
SINGLE TILT (SIDEWAYS)	S-TILT
REPEATED TILTS (SIDEWAYS)	R-TILT
SIDE-TURN	SIDE-TURN
SHAKE (REPEATED)	SHAKE
WAGGLE	WAGGLE
OTHER	OTHER

Since the video recordings were shot from the side, it was not always possible to analyse eyebrow movements and gaze direction; however it was often possible to perceive whether the interlocutors were looking at each other or not while expressing feedback. For this reason the annotation of gaze used just two categories to mark whether the interlocutors were exchanging eye contact or not. This is done on the tier called Gaze by using the categories EYECONTACT and NON EYECONTACT.

On the tier called Function, the semantic-pragmatic function of the identified feedback phenomenon is coded. On this tier there are two fine-grained sub-groups of categories: one for speech act annotated as FEEDBACK ELICIT and one for those annotated as FEEDBACK GIVE.

The functions and the relative labels that expressions produced to give feedback can have are those shown in tables 4.3a and 4.3b in chapter 4.

If both verbal and non-verbal phenomena are annotated, then it is possible to interpret the multi-modal relationship between them in terms of

DEPENDENCY or INDEPENDENCY and annotate it in the tier called: Multi-Modal Relation. When the verbal and non-verbal expression are dependent on each other, they can either complement or contradict each other.

6.4 Results

The tags used for the annotation of verbal and non-verbal phenomena related to feedback allow the automatic retrieval of several quantitative measures, such as the number of occurrences of feedback expressions, their direction and type, their distribution and their specific semantic-pragmatic function.

Since FEEDBACK is investigated in this chapter as a multi-modal phenomenon, in showing the results the focus is placed on the co-occurrence of feedback signals in different modalities, that is auditory (verbal feedback) versus visual (non-verbal feedback).

6.4.1 Feedback Distribution

First of all, the distribution of feedback expressions was calculated counting all the identified FEEDBACK in each dialogue and for each dialogue participant. This result is shown in figure 6.2.

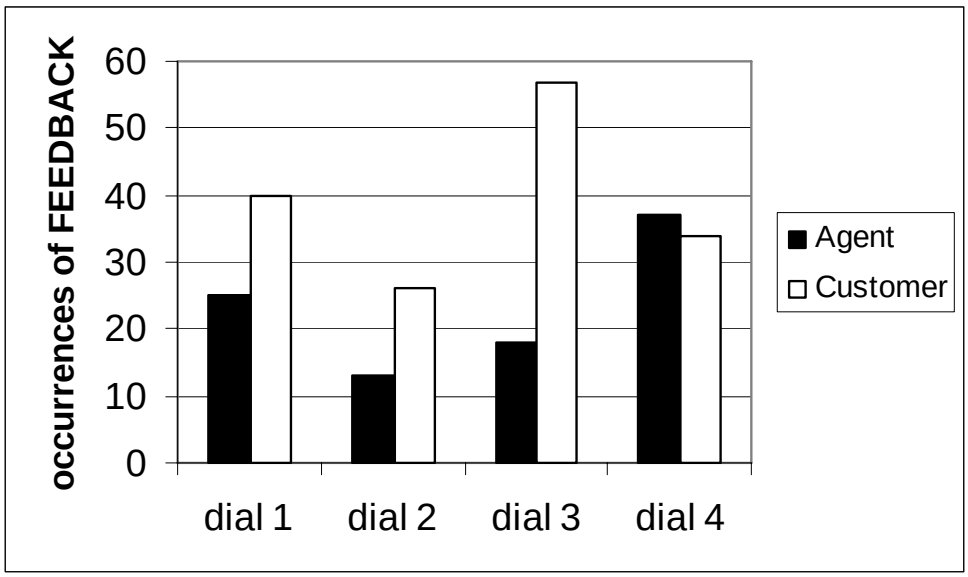

Figure 6.2 Number of occurrences of feedback expressions per dialogue and dialogue participant.

The customers tend to produce a higher number of feedback expressions than the agent does. Given the particular communicative situation, it is mostly the customer that has the role of "listener" while the travel agent has

the role of "speaker". The agent has, in fact, the right to maintain the turn until she supplies the information appropriate to the needs of the customer. As a consequence, the agent produces longer contributions, while the customers produce a great number of short contributions containing or consisting of FEEDBACK. The number of feedback expression shown in figure 6.2 includes verbal feedback expressions, non-verbal feedback expressions and co-occurring verbal and non-verbal feedback expressions. The pie chart in figure 6.3 shows the general distribution of FEEDBACK.

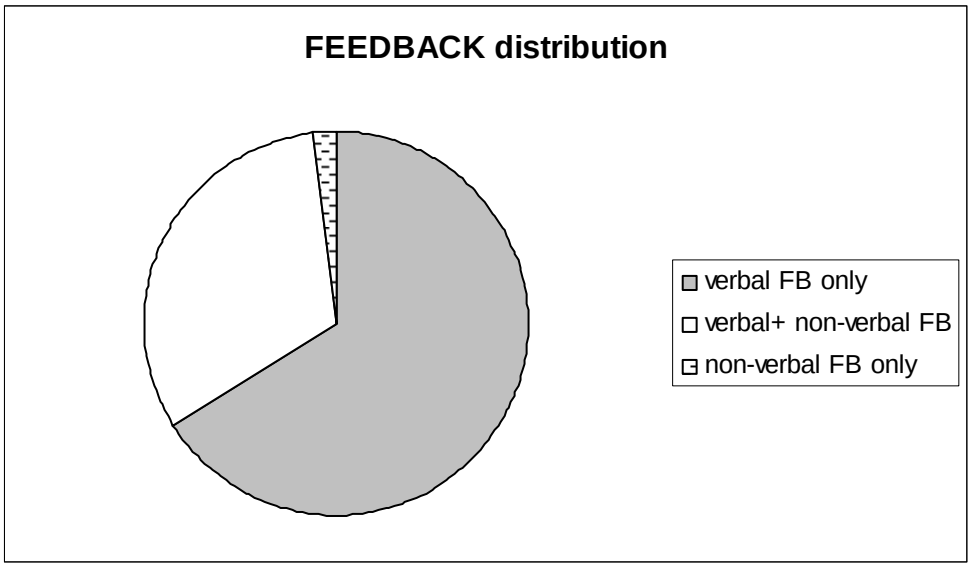

Figure 6.3 Distribution of FEEDBACK (FB) in all four dialogues.

FEEDBACK is distributed as follow: 65% consists of verbal expressions only, 31 % of co-occurring verbal and non-verbal expressions and 2% of non-verbal expression only.

The bar chart in figure 6.4 shows the distribution of FEEDBACK in a more detailed way: each column displays the percentage of verbal feedback occurring without non-verbal feedback (the black area), the verbal feedback co-occurring with non-verbal feedback (the white area) and non-verbal feedback occurring without verbal feedback (the patterned area, only four examples in GSLC-Dial 1 and 4). The percentages are calculated relative to the total number of feedback produced in each dialogue.

In GSLC-Dial 2 the production of non-verbal feedback is higher compared to the other dialogues, depending on the fact that the customer in this dialogue is a very high producer of feedback, in fact more than 70% of his contributions include feedback, which is often signalled by the co-occurrence of short feedback words and head nods. (This will be shown in more detail in chapter 7, section 7.2.3.1).

In GSLC-Dial 3 the production of feedback expressions co-occurring with non-verbal expressions is very low. This might be interpreted as a

peculiarity of this customer, who might be considered as a "low producer" of non-verbal feedback or can be explained by the fact that the customer in this dialogue was standing in a position which is likely to have prevented him from moving his head and hands freely, namely he was bent towards the agent with his elbow on the desk and his hand under his chin.

The fact that the production of non-verbal feedback only is so low in the four analysed dialogues can probably be explained by assuming that in dyadic conversations, dialogue participants feel like they can produce verbal feedback without the risk of interrupting the other interlocutor and without disturbing the smooth unfolding of conversation.

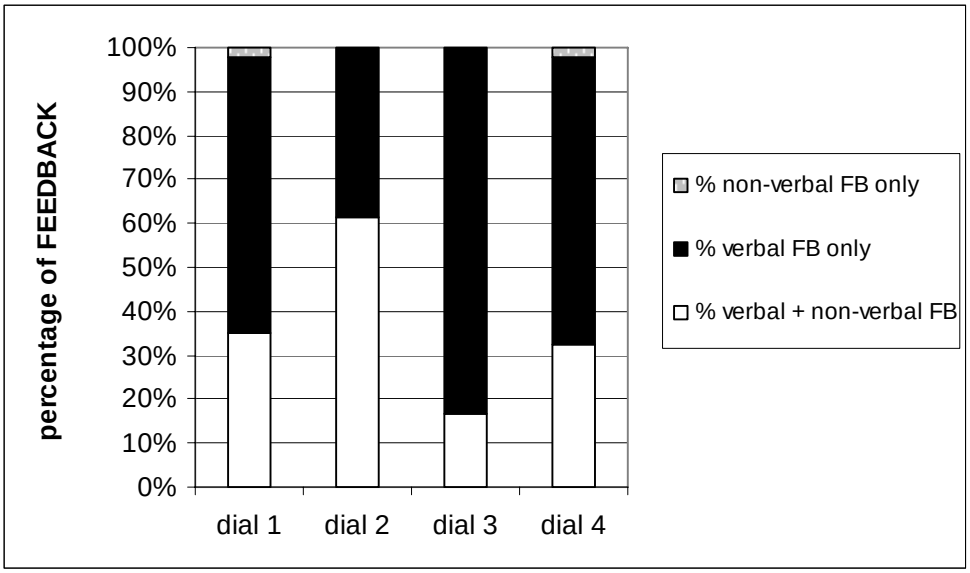

Figure 6.4 Distribution of feedback expressions (FB) per dialogue.

Figure 6.5 shows how non-verbal feedback expressions co-occurring with verbal feedback expressions are distributed across speakers in each dialogue. This is calculated as percentage of the total number of verbal feedback co-occurring with non-verbal feedback.

No relevant differences appear in the total distribution of non-verbal feedback expressions co-occurring with verbal feedback across speakers.

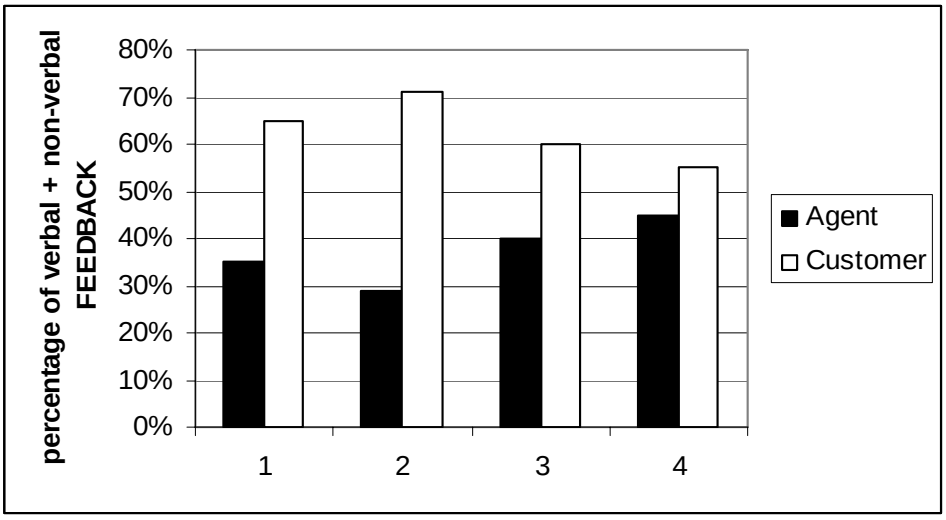

Figure 6.5 Percentage of verbal feedback expressions (FB) co-occurring with non-verbal feedback per dialogue and dialogue participant.

6.4.2 Feedback Type and Direction

The analysis of the direction the feedback phenomena have in the four analysed dialogues shows that the travel agent elicits FEEDBACK more often than the customers: 12% of the feedback expressions produced by the agent have the direction type ELICIT, while only 2% of the customers' FEEDBACK expressions have the direction type ELICIT. The agent elicits FEEDBACK mostly to REQUIRE ACCEPTANCE and does so by producing HEAD MOVEMENTS consisting of REPEATED NODS, SHAKES and WAGGLES. Figure 6.6 shows the distribution of feedback per specific semantic-pragmatic category.

The semantic-pragmatic functions of feedback expression with give direction were coded using the following categories:

- CONTINUATION YOU GO ON (FBGiCY),
- CONTINUATION I GO ON (FBGiCI),
- ACCEPTANCE (FBGiA),
- NON-ACCEPTANCE (FBGiR),
- EXPRESSIVE (FBGiEx);

The semantic-pragmatic functions of feedback expressions with ELICIT direction were coded using the following categories:

- CHECK ATTENTION (FBElChA)
- REQUIRE ACCEPTANCE (FBElRA)
- More information (FBElM)

The most common functions that feedback carries out in the four analysed dialogues are GIVE CONTINUATION YOU GO ON and GIVE ACCEPTANCE.

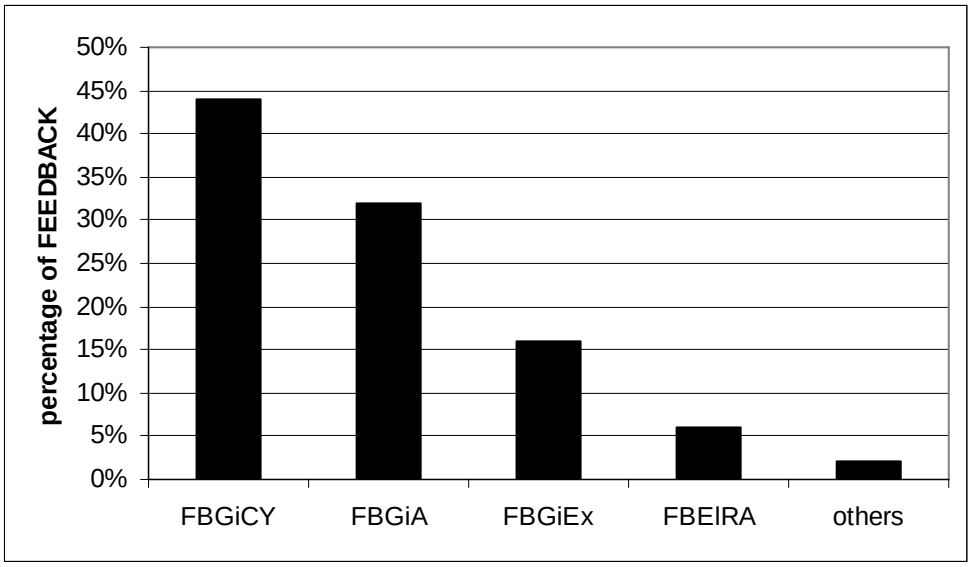

Figure 6.6 Distribution of feedback per specific semantic-pragmatic category.

A more detailed analysis of the distribution of feedback per category, dialogue participant and dialogue is shown in chapter 7, section 7.2.3.2.

A total of 91 facial displays signalling FEEDBACK were identified in the four dialogues. Table 6.3 shows the distribution of FACIAL DISPLAYS per specific FEEDBACK GIVE and FEEDBACK ELICIT function. The category FACIAL DISPLAYS refers to phenomena such as changes in eyebrow position, expressions of the mouth, and movement of the head and eyes [Cassell 2000]. Of the identified FACIAL DISPLAYS with FEEDBACK function, 90% consist of HEAD MOVEMENTS.

The low number of eyebrows features depends on the fact that it was not always possible to observe the eyebrow behaviour on the videos, since the dialogue participants were recorded from the side.

SINGLE NOD, REPEATED NODS and SINGLE JERK are the most frequent HEAD MOVEMENTS identified in these materials. REPEATED NODS seem to mostly serve the function of FEEDBACK GIVE ACCEPTANCE, while SINGLE NOD is mostly produced with a FEEDBACK GIVE CONTINUATION YOU GO ON function. SHAKES are produced either to GIVE NON ACCEPTANCE or to ELICIT REQUIRE ACCEPTANCE.

Table 6.3 Number of occurrences of FACIAL DISPLAYS per FEEDBACK GIVE and FEEDBACK ELICIT function[17].

FACIAL DISPLAYS		FB GIVE					FB ELICIT		Total number of Occurrences
		CY	CI	A	EX	R	RA	M	
HEAD MOVEMENTS	R- NODS	4		22	3		6	2	36
	S-NOD	14		2					16
	S-JERK	3		5	4				12
	SHAKE					2	3		5
	FORWARD	4							4
	SIDE TURN		2						2
	BACK	2	1						3
	WAGGLE						2		2
GENERAL FACE	LAUGH				3				3
	SMILE			2	2				4
	EYEBROW RAISE				2	2			4

SINGLE NOD and REPEATED NODS are mostly related to one specific function, however it is not possible to conclude that there is a one-to-one correspondence between a specific head movement and a specific feedback function: the same non-verbal expression can in fact be used to convey different specific feedback functions, as in the case of REPEATED NODS which are used to GIVE CONTINUATION YOU GO ON, GIVE ACCEPTANCE, GIVE EXPRESSIVE, ELICIT REQUIRE ACCEPTANCE and ELICIT MORE INFORMATION.

The co-occurrence between verbal and non-verbal expressions was also analysed, and the results for the most common HEAD MOVEMENTS are shown in table 6.4. SINGLE NOD, and JERKS co-occur with FEEDBACK WORDS, while REPEATED NODS and SHAKES can co-occur with all types of verbal feedback.

The most common words co-occurring with SINGLE NOD and JERK are *ja* and m-like words, however, it is not possible to establish a one-to-one correspondence between a specific head movement and a specific word. In fact, in the analysed dialogues SINGLE NOD can co-occur even with verbal expressions other than m-like words and *ja,* for instance with sentences such as: *ja men det går ju bra* (yes it is ok) and even longer sentences.

[17] No FACIAL DISPLAYS having the function FEEDBACK ELICIT CHECK ATTENTION were found in the data; as a consequence, this category is not listed on the table.

Table 6.4 Number of occurrences of HEAD MOVEMENTS per verbal FEEDBACK TYPE.

HEAD MOVEMENT	Verbal FEEDBACK TYPE		
	WORD	PHRASE	SENTENCE
R- NODS	8	5	23
S-NOD	14		2
S-JERK	12		
SHAKE	1	2	2

GAZE DIRECTION is not listed among FACIAL DISPLAYS in table 6.3 since it was analysed only in terms of the two categories EYECONTACT versus NON-EYE CONTACT. The percentage of times when non-verbal feedback was produced at the same time as EYECONTACT has been calculated. The results show that when non-verbal feedback is produced, in 71% of cases it is produced simultaneously with the interlocutors having eyecontact.

Looking at the videos, it is easy to notice a kind of "mimicking" phenomenon of the interlocutors in which they tend to adapt to each other and produce head movements in response to the production of a head movement. This behaviour is quite evident especially when the interlocutors look at each other. McClave [2000] shows that a higher number of head movements is produced when the interlocutors establish eyecontact.

However in the four dialogues analysed for this study often small head movements, such as SINGLE NOD and JERK, are produced even when there is no eyecontact between the speakers. This observation seems to be consistent with previous investigations that have shown that head movements co-occurring with verbal feedback expressions are produced even when the interlocutors do not or cannot have eye contact, as for instance in telephone conversations [Nivre & Richthoff 1988].

Besides facial displays, it was possible to observe a few instances of other non-verbal behaviour related to FEEDBACK, such as shoulder shrugs, and hand movements, but these were not common (a maximum of 2 or 3 instances per dialogue).

It was possible to observe even FACIAL DISPLAYS and other non-verbal behaviour not related to FEEDBACK, mostly head movements such as batonic[18] gestures, head shakes for negative answers, hand movements as pointing, iconic and batonic gestures. However, a detailed analysis of this non-verbal behaviour was beyond the scope of this investigation.

[18] Batonic gestures [Bull, 1987], also called "beats" [McNeill, 1992] are small movements the shape of which does not change with the content of the accompanying speech; they have no meaning, but relate to phrasal stress to emphasise some parts of speech. According to Bavelas et al. [1992], they serve the function of keeping the listener attentive.

6.4.3 Multi-Modal Relationship

The multi-modal relationship between co-occurring verbal and non-verbal expressions is summarised in table 6.5. For each FEEDBACK GIVE FUNCTION the most common non-verbal expressions and multi-modal relationship is listed. By multi-modal relationship is intended the relationship between the co-occurring verbal and non verbal feedback expressions, which is annotated in terms of COMPLEMENT or CONTRADICT. Most of the verbal and non-verbal feedback expressions co-occur to complement each other.

Table 6.5 Specific function of co-occurring verbal and non-verbal feedback expressions.

FEEDBACK GIVE Functions	**Non-verbal expression**	**Multi-modal relationship**
CONTINUATION	SINGLE NOD, SINGLE JERK	COMPLEMENT
ACCEPTANCE	SINGLE NOD, REPEATED NODS, SINGLE JERK, SMILE	COMPLEMENT
NON-ACCEPTANCE	SINGLE NOD, SHAKE, EYEBROWS RAISING	COMPLEMENT
EXPRESSIVE	SHAKES, SMILE, EYEBROWS RAISING	COMPLEMENT, CONTRADICT

It would have been interesting to analyse what happens when verbal and non-verbal expressions co-occur to express feedback in terms of interaction between modalities; for instance by analysing whether the co-occurrence of non-verbal behaviour might influence the production of verbal-feedback in terms of articulatory and prosodic characteristics. Previous studies have shown that head movements often co-occur with intonation cues [Bertrand et al. 1995] which might mean that feedback words produced in co-occurrence with head movement might carry some focus, and as consequence show longer duration. Previous studies have in fact shown that prosodically marked items have a longer duration compared to the same items without prominent prosodic characteristics [Caspers 2003].

Unfortunately, the four dialogues analysed in this chapter do not allow for the accurate measurement of the actual duration of the identified head nods, nor for the accurate acoustic and tonal analysis of the verbal feedback expressions. For this reason it was not possible to test the hypothesis that the co-occurrence of non-verbal behaviour might influence the production of verbal-feedback in terms of articulatory and prosodic characteristics.

However it was possible to observe that short expressions with the function CONTINUATION YOU GO ON when produced in a minimally intrusive way tend to co-occur with minimal HEAD NODS and JERKS. This is because the dialogue participant who produces these short feedbacks has no intention to interrupt the other dialogue participant who is actually speaking.

S/he is rather willing to show an active listening attitude by producing minimal intrusive verbal and non-verbal feedback that serve the main function of showing continuation of attention and no intention to get the floor.

For longer feedback expressions, like repetitions, reformulation, with GIVE ACCEPTANCE or NON ACCEPTANCE function, and for feedback expressions with the function GIVE EXPRESSIVE, the co-occurring non-verbal expression tend to be more extensive, like REPEATED NODS or sequences of REPEATED NODS, SMILE and other FACIAL DISPLAYS (see table 6.3). Moreover the function GIVE EXPRESSIVE is often marked by some phonological and prosodic phenomena, like lengthening and variation of pitch contour.

6.5 Conclusions and Discussion

The aim of the study reported in this chapter was to show evidence that it is possible to categorise the semantic-pragmatic function of non-verbal feedback expressions by using the specific categories provided in the coding scheme and at the same time explore the realization of non-verbal expressions produced to signal feedback in real spontaneous dialogues.

The results seem to point out that the semantic-pragmatic function of feedback expressions can be classified by using the specific categories provided in the coding scheme. These categories are in fact independent of the modality in which feedback is expressed.

In the four analysed dialogues, a total of 250 FEEDBACK were identified: 65% of these consisted of verbal expressions only, 33% of co-occurring verbal and non-verbal expressions and 2% of non-verbal expressions only.

The low production of non-verbal feedback only in the four analysed dialogues can probably be explained by assuming that in dyadic conversations dialogue participants feel that they can produce verbal feedback without the risk of interrupting the other interlocutor and without disturbing the smooth unfolding of conversation. This explanation is consistence also with the result of the analysis of the distribution of feedback per semantic-pragmatic category, which shows that most of the identified feedback serves the function CONTINUATION YOU GO ON and is realised by means of short verbal expressions occurring either on their own or accompanied by short HEAD MOVEMENTS.

The most common HEAD MOVEMENTS produced to signal FEEDBACK are SINGLE NOD and JERK. These usually co-occur with feedback words such as m-like words and *ja* produced in a non-intrusive way.

For longer FEEDBACK, such as repetitions, reformulations, with GIVE ACCEPTANCE or NON ACCEPTANCE function, and for FEEDBACK with the function GIVE EXPRESSIVE, the co-occurring non-verbal expressions tend to

be more extensive, like REPEATED NODS or sequences of REPEATED NODS, SMILE and other FACIAL DISPLAYS.

In the four dialogues analysed in this chapter, it is possible to notice a "mimicking" phenomenon of the interlocutors that tends to produce head movements in response to the production of a head movements. However it was also noticed that small head movements, such as SINGLE NODS and JERKS, are produced even when there is no eyecontact between the dialogue participants. This might be habit, but it might also be that head movements and gestures are used even when interlocutors cannot see and interpret them because they help the speaker in the process of communicating ideas [Gullberg 1998; Goldin-Meadow 2003; Morsella & Krauss 2004].

Starting from these observations, it is possible to suppose that some co-occurring head-movements are produced unconsciously, as a way of maximizing the effectiveness of the only available channel (the auditory one) while other head movements are produced intentionally to modify the meaning of the co-occurring verbal expression when the visual channel is available for transmission. This assumption is mainly based on the observation that in order to convey a specific intentional meaning a movement has to be seen/received by the interlocutor, otherwise its communicative effect fails. If a speaker wishes to add some more information to his/her verbal production by means of an emphasizing gesture, for instance, he/she does it in a conscious way and supposing that the listener is looking at him/her and able to interpret his/her signal.

The materials analysed in this chapter are quite limited in their amount and in the fact that they represent just one cultural community. Moreover, the fact that the dialogues were recorded in a real environment constrained the quality of the recordings, which as a consequence, did not allow for an accurate analysis of how phenomena co-occurring in different modalities can influence each other.

Notwithstanding the several limitations shown by the materials analysed in this chapter, the advantages of using them for the investigation of non-verbal feedback phenomena outweigh the disadvantages they show. This is because they represent a good source for the study of feedback phenomena produced in spontaneous dialogues recorded in real communicative situations.

One more advantage of having these data available is that, being their activity type "information seeking", they could be compared to the available human-machine interactions belonging to the same activity type.

In the next chapter, the production of feedback phenomena observed in the spontaneous human-human dialogues analysed in this chapter is compared to the production of feedback phenomena in human-machine interaction with a Swedish experimental multi-modal dialogue system.

7 Feedback Phenomena in Human-Machine Interactions

7.1 Introduction

With the recent development in speech technologies, conversational speech interfaces are becoming more advanced and a larger number of users expect to be able to interact with their computer systems in the way they do with other people. This means that users of speech-based interfaces tend to integrate a larger number of human discourse features when interacting with computer systems and expect the interface to be able to produce and understand human-like behaviour.

Examples of human discourse features used during human-machine interactions are feedback expressions and turn-management signals. In human-human conversation, dialogue participants continuously give or elicit feedback to inform each other on the state of communication, and display turn-management signals to regulate the interaction and make it proceed smoothly. This is done by means of different kinds of verbal and non-verbal expressions that can have different semantic-pragmatic functions depending on the context in which they occur. In human-machine interactions, the production of feedback signals on the users' side tends to be less pervasive than in human-human communication [Okato et al. 1998]. This might depend on the fact that some users might find it uncomfortable to use feedback signals in machine interactions [Ward & Heeman 2000], probably because of the way in which interactions are designed. Most dialogue systems in fact still do not provide opportunities for the user to produce feedback signals and do not display feedback signals in the same way that humans do when they communicate with each other.

This chapter presents two studies that speculate on the possibilities of considering feedback expressions and turn-management signals as human discourse features that might be used to enhance smoothness in human-machine interactions. Study 1 is a comparison between verbal feedback phenomena in human-human and human-machine interactions, study 2 is a preliminary investigation of the production of non-verbal behaviour related to feedback, turn management and the visual expression of emotional attitudes by users of a multi-modal dialogue system.

7.2 Study 1 Verbal Feedback in Human-Human and Human-Machine Communication

The aim of this study is twofold: to verify that the categories used to label feedback expressions in human-human interactions are feasible also for feedback produced in human-machine interactions, and to get more insight in the production of feedback phenomena across different communicative situations. Having available dialogues belonging to the same activity type (that is "factual information seeking"), but to different communicative situations (human-human versus human-machine interactions) allows for the comparative investigation of the realization of feedback with the aim of finding which behaviour generalises across situations.

Moreover an auditory analysis and some acoustic measurements of the identified feedback words were carried out with the aim of verifying whether the acoustic characteristics of feedback expressions, such as duration and pitch contour, reflect their semantic-pragmatic function.

The investigation of the prosodic marking of short feedback expressions in relation to their specific communicative function is of fundamental interest when it comes to technological applications. Much current research towards human-like behaviour in spoken dialogue systems is oriented to testing the benefit of real-time prosodic analysis for interaction control and appropriate timing of feedback [Hirshberg 2002]. Prosodic cues have already been proven helpful when it comes to deciding the appropriate timing for speaking or remaining silent [Edlund & Heldner 2005; Edlund, Heldner & Gustafson 2005]. Similarly prosodic cues might result as helpful when it comes to the on-line interpretation of the specific communicative function of the short feedback expressions produced by users.

7.2.1 Materials

The comparison of the production of verbal feedback phenomena in human-human and human-machine dialogues was carried out using four dialogues selected from the GSLC corpus and four dialogues selected from the AdApt database.

The human-human dialogues are the same used to analyse feedback phenomena in Swedish in the study reported in chapter 6. These consist of spontaneous interactions between four different customers and the same travel agent, video recorded in a travel agency in Gothenburg (GSLC dialogues, for more details see section 3.3.1). These dialogues can be described as "factual information seeking" exchanges, where the customer asks the travel agent for information about timetables, visas, hotels, and the travel agent provides the information required.

The four human-machine interactions were selected from the first collection of the AdApt database. These interactions are referred to as AdApt Corpus I (for more details about the interactions see section 3.2.2).

AdApt was a Swedish experimental conversational multi-modal dialogue system, able to provide information about real estate in Stockholm. The dialogues in the AdApt database can also be described as "factual information seeking" since the users were instructed to interact with the system in order to find apartments in Stockholm that fulfilled certain criteria.

The results of a previous analysis of positive and negative users' feedback in the AdApt database [Bell & Gustafson 2000] showed that 94% of the users used feedback at least once in their interaction with the system, even if large individual variations were noted (i.e. some users gave much more feedback than others). For this study, four users (three male and one female) were randomly selected from those interactions in which the number of utterances containing feedback was at least 10% according to Bell and Gustafson's calculation [2000]. The recording set-up of the AdApt system was designed in such a way that it did not record what the users said while the embodied agent was speaking; as a consequence it is not possible to know whether short feedback expressions were produced but not recorded while the agent was talking, or if they were not produced at all.

7.2.2 Method

In this first study presented in this chapter, only verbal feedback phenomena are taken into account, this because the human-machine interactions selected from the first collection of the AdApt database were not video recorded, in contrast to the GSLC dialogues.

Verbal feedback expressions were identified as a reaction to the previous speech act and coded with the categories provided in the coding scheme illustrated in chapter 4. The same categories were used to code feedback both in human-human dialogues and human-machine interactions.

The acoustic analyses were performed using the tool WaveSurfer [Sjölander & Beskow 2000].

7.2.3 Results

Several quantitative measures, such as the percentage of contribution in each dialogue containing feedback and the type and function of feedback expressions, were retrieved. The specific semantic-pragmatic functions of the identified feedback expressions were compared across the two communicative situations. This information, together with the results of the auditory and acoustic analysis, provides an overall picture of the production of feedback phenomena in the two communicative situations [Cerrato 2002a].

7.2.3.1 Feedback Distribution

The distribution of verbal feedback expressions can be calculated in different ways. In this study, to allow for comparison across the two different communicative situations (human-human versus human-machine interactions) the distribution of feedback was calculated by counting how many contributions in each dialogue and for each dialogue participant contain at least one feedback expression and this number was related to the total number of contributions per dialogue and dialogue participant.

The percentage of contributions containing verbal feedback was calculated in both human-human and human-machine dialogues. For the four human-human dialogues the percentage is displayed per interlocutor (Agent and Customer) in figure 7.1. For the users in human machine interactions it is shown in figure 7.2 (only for the users, since the virtual agent did not give any explicit feedback).

The number of contributions containing at least one feedback expression does not necessarily correspond to the total number of feedback expressions produced, since a single contribution can include more than one feedback expression. For this reason another way to calculate the distribution of verbal feedback expressions is to count them in each dialogue and for each dialogue participant. This result for the four human-human dialogues is shown in chapter 6 in figure 6.2.

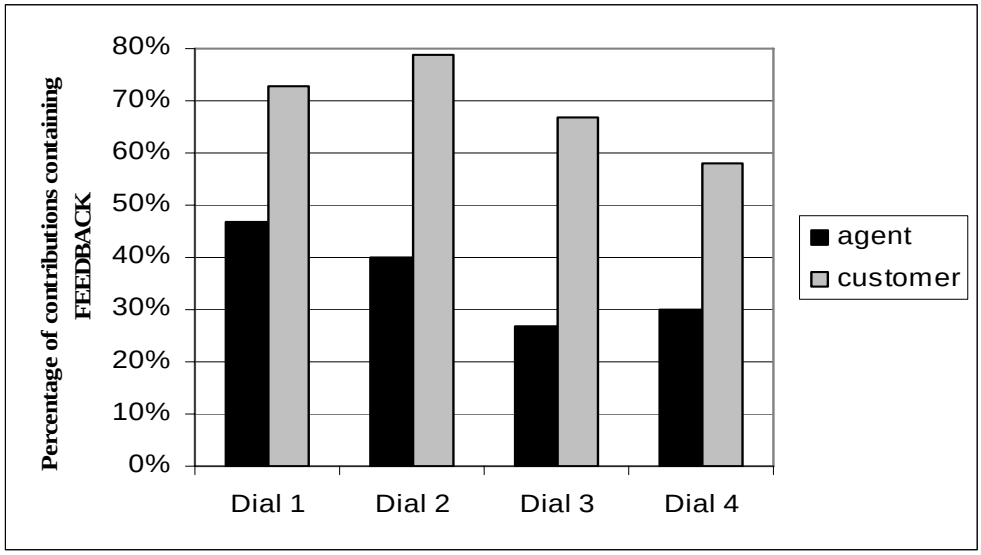

Figure 7.1 Percentage of contributions containing verbal feedback per dialogue participant (Agent vs. Customers) in the four human-human dialogues.

In these four human-human dialogues selected from the GSLC corpus, the two interlocutors have different roles: one is the travel agent, always the same female speaker, and her interlocutors are four different customers. The

percentage of contributions containing feedback is quite high for the customers across the four dialogues; it is higher than that of the agent in three of the four dialogues. Given the particular case of communicative situation, it is mostly the customer who has the role of the listener, and the travel agent who has the role of the speaker. The travel agent has in fact the right to maintain the turn until she supplies the information appropriate to the needs of the customer. As a consequence, the customers/listeners use a considerable amount of feedback to GIVE CONTINUATION YOU GO ON. This is because they wish to show their "active" participation in the interaction.

The travel agent elicits feedback more often than the customers: 12% of the feedback expressions produced by the agent have the direction type ELICIT, while only 2% of the customers' feedback expression has the direction type ELICIT.

The most common strategy adopted by the travel agent to ELICIT FEEDBACK REQUIRE ACCEPTANCE is by asking an explicit verification question, like for instance: *did you say you wanted to travel on Monday?*[19] or repeating or reformulating what the customer has said in her/his last contribution. So for instance if the agent asks: *how many of you are travelling?* and the customer answers by saying: *we are two*, then the agent might produce a reformulation with the function FEEDBACK GIVE ACCEPTANCE by saying: *you are two.*

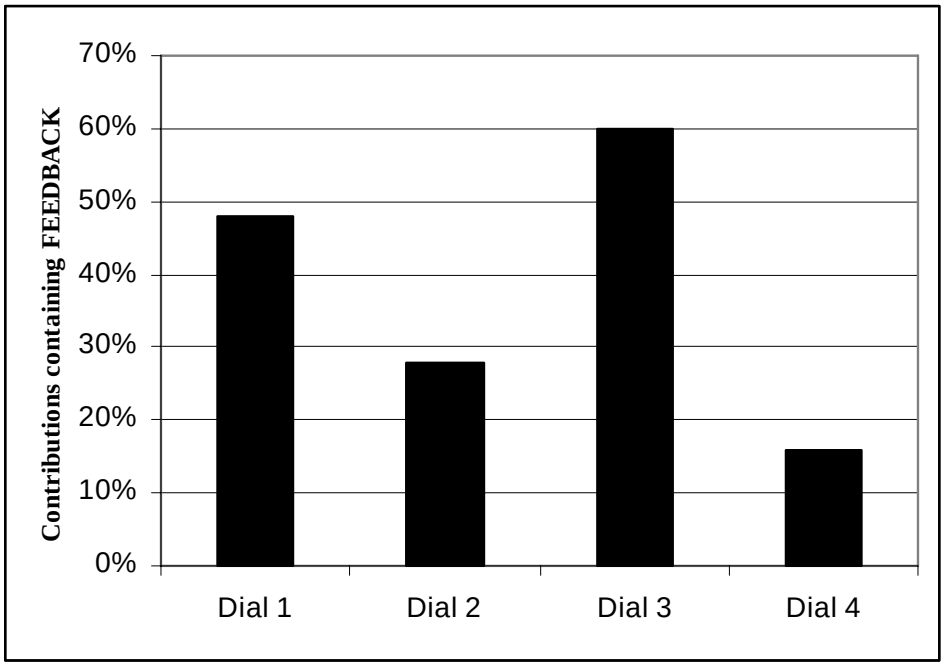

Figure 7.2 Percentage of contribution containing verbal feedback per user in human-machine interactions.

[19] The examples reported here are translated from the original transcription in Swedish.

Figure 7.2 shows the percentage of contributions containing verbal feedback per user in human-machine interactions. In these interactions the embodied agent does not produce any explicit feedback. The amount of feedback produced by the users ranges from 16% to 60%, however these figures are especially interesting if we consider that the virtual agent never elicited nor gave explicit feedback.

7.2.3.2 Feedback Functions

In the interactions analysed in this first study in this chapter, FEEDBACK shows mainly a GIVE direction type. In human-human dialogues 7% of feedback expressions have the direction ELICIT. Most of the identified FEEDBACK with ELICIT direction have the function REQUIRE ACCEPTANCE (FBElRA) (see chapter 6, figure 6.6).

The semantic-pragmatic functions of feedback expression with GIVE direction were coded using the following categories:

- CONTINUATION YOU GO ON (FBGiCY),
- CONTINUATION I GO ON (FBGiCI),
- ACCEPTANCE (FBGiA),
- NON-ACCEPTANCE (FBGiR),
- EXPRESSIVE (FBGiEx);

These categories are the ones proposed in the coding scheme in chapter 4 and used to code the semantic-pragmatic function of FEEDBACK GIVE expression in all the studies presented in this thesis. These categories provide an interpretation of feedback expressions in terms of type of reaction to the previous communicative act. This categorisation is applied both to the feedback expressions produced in the human-human dialogues and to those produced in human-machine interactions.

In figures 7.3a and 7.3b the feedback expressions produced in the four GSLC dialogues respectively by the customers and the travel agent are grouped according to the most frequent semantic-pragmatic categories: CONTINUATION YOU GO ON(FBGiCY), ACCEPTANCE (FBGiA), EXPRESSIVE (FBGiEx) and REQUIRE ACCEPTANCE (FBElRA). Only 2% of the total identified feedback belongs to categories other than these.

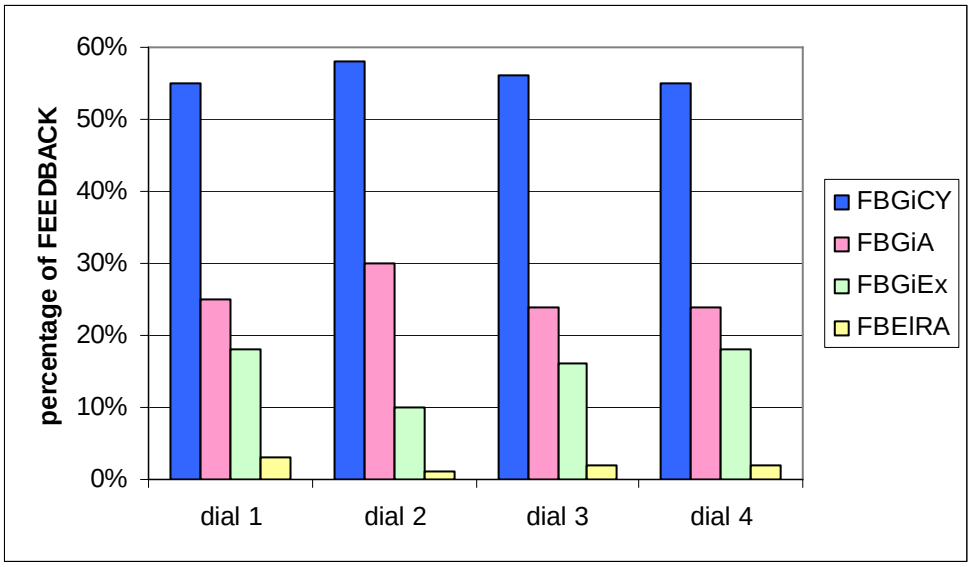

Figure 7.3a Percentage of customers' feedback expressions per category.

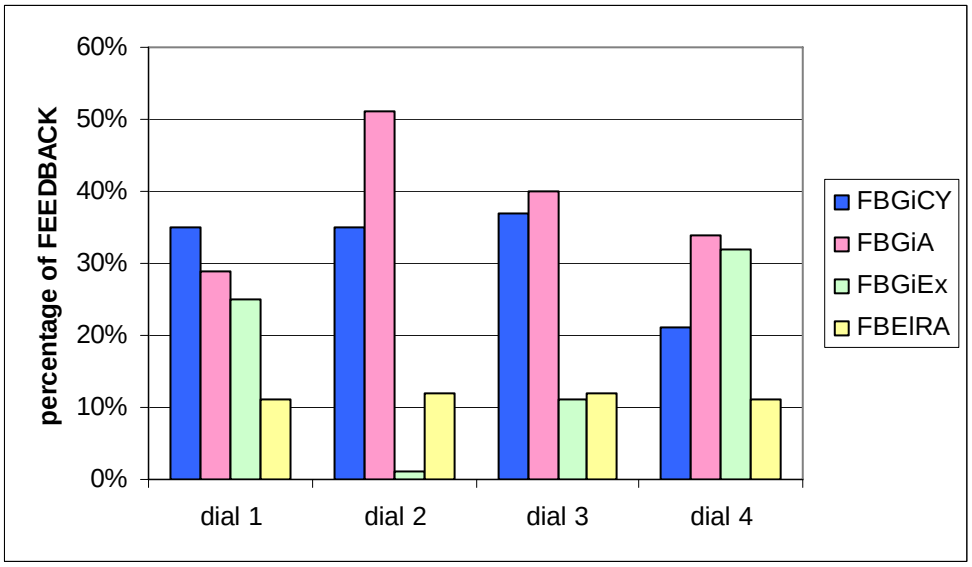

Figure 7.3b Percentage of agent's feedback expressions per category.

The most common function served by the feedback expressions produced by the customer is GIVE CONTINUATION YOU GO ON (FBGiCY), while for the agent is GIVE ACCEPTANCE (FBGiA).

The distribution of feedback per function and dialogue participants across dialogues shown in fig 7.3a and 7.3b can be interpreted in terms of cooperation strategies: in the dialogues the customer asks for some information and the agent provides the information required. For this reason the customer has mostly the role of active listener and tries to show his/her attention by giving feedback with CONTINUATION YOU GO ON (FBGiCY)

function, which shows attention and the willingness to go on without interrupting and taking the floor.

The agent has mostly the role of speaker and as a consequence she mainly produces feedback with the function GIVE ACCEPTANCE (FBGiA) to confirm that she has understood the request of the customers and she also elicits more feedback to REQUIRE ACCEPTANCE (FBElRA) than the customers to be sure that they understand what she is saying.

The feedback expressions belonging to the most common categories, GIVE CONTINUATION YOU GO ON and GIVE ACCEPTANCE consist mainly of *ja*, in its several phonetic realizations, m-like words and the short expression *nä*, often produced in a minimal non-intrusive way.

For the correct interpretation of these responses, it is important to consider how the "polarity" of the preceding communicative acts affects their function [Allwood, Nivre & Ahlsén 1992]. In Example 1 below, extracted from GSLC-Dial 3, the function of the expression *nä* is that of ACCEPTANCE of the information, since the preceding statement has a negative polarity. The agent is explaining the conditions for the booking of the hotel rooms she has just made and in contribution $G70 she says: if you want to have them you do not need to contact them. To this the customer reacts by producing the short response: *nä nä* which has been coded as GIVE FEEDBACK CONTINUATION YOU GO ON. The agents continues her explanation in contribution $G72 and says: if you do not want to have them or change anything /then/ you have to cancel the booking. To this completion of the explanation the customer reacts with another *nä,* reinforced by the adverb *precis* (exactly). This reaction has been coded as FEEDBACK GIVE ACCEPTANCE.

$G70:	*vill du ha dom så behöver du inte höra av dig/*
$C71:	***nä nä <FB;W;Gi;CY>***
$G72:	*vill ni inte ha dom eller ändra något /så / måste ni boka av*
$C73:	***nä precis* // <FB;W;Gi;A>**

Example 1 from GSLC-Dial 3: nä with different FEEDBACK functions.

For the customers these short expressions occur 90% of the times in a contribution of their own, for the travel agent in 80% of the times.

Figure 7.4 shows the percentage of feedback expressions produced respectively by the users of the dialogue system, the customers of the travel agency and the travel agent, grouped per semantic-pragmatic category; the embodied conversational agent in the AdApt system did not give any explicit feedback.

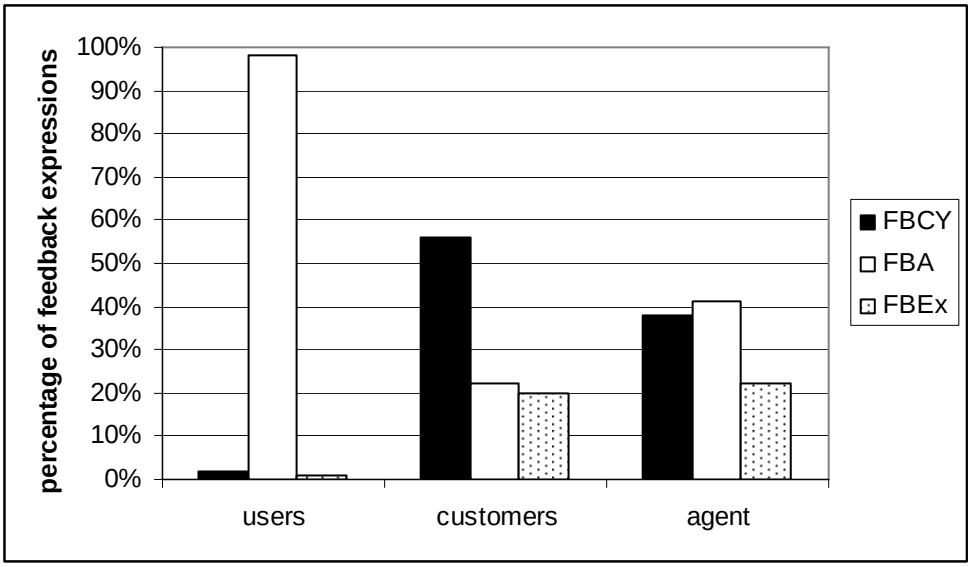

Figure 7.4 percentage of feedback expressions per category produced by users of the dialogue systems and customers and agent in the travel agency dialogues.

It is not possible to determine if in the four human-machine interactions short expressions were actually produced as a contribution of their own and not recorded, or did not occur at all. This is because of the way in which the dialogue system was designed: the system did not record what the users said while the embodied agent was speaking.

However, it was possible to observe that in the human-machine interactions, feedback words occur at the initial position of a longer utterance and their function is almost never CONTINUATION YOU GO ON, and never CONTINUATION I GO ON, but ACCEPTANCE of the information. This can of course depend on the constraint of the recording set-up, but it could also be that in the human-machine interactions FEEDBACK with CONTINUATION YOU GO ON function does not occur because of the way in which the interactions are structured. In these interactions the system produces standardised prompts consisting either of short questions such as: *var ska lägenheten ligga? hur många rum vill du ha*? *hur mycket får lägenheten kosta*? (where does the flat have to be placed? How many rooms do you wish? How much can the flat cost?) or of short descriptions of the available flats in Stockholm, such as: *det finns sju sådana lägenheter och de visas nu på kartan (*there are seven such flats and they are now shown on the map), *den gröna lägenheten har badkar* (the green flat has a bathtub). The longest prompt counts ca. 20 words and can consist of a more accurate description of a specific flat, as for instance: *det finns en trea på 75 kvadratmeter som ligger på Scheelegatan 28, lägenheten har kakelugn parkettgolv, takrosett och kabeltv* (There is a three-room flat of 75 square meters in *Scheelegatan* 28. This flat has a stove, parquet flooring, ceiling rose and cable television.).

The embodied agent in the system produces contributions which in general are shorter compared to those of the human travel agent and he does not show the rich set of "own communication management" (hesitations, re-starts and so on) and non-verbal behaviour employed by the human travel agent during communication. Hesitations and non-verbal behaviour can easily trigger the production of feedback, as shown in example 2.

A$42:	*och sen vet ja{g} inte hur de{t} kommer att vara nästa år* /
C$43:	S-NOD **<FB; HEAD MOVEMENT, S-NOD; Gi; CY>**
A$44:	*i och me{d} att de{t} är de{t} är årsskiftet och de{t} så kanske de{t}är väldi{g}t tryck*
C$45:	***ja*** **<FB;W;Gi;A>**
A$46:	*eller så är de{t} väldi{g}t lågt / de{t} är ingen som som vågar flyga*
C$47:	*ja nä* **<FB;W;Gi;A>**

Example 2 from GSLC Dial 1: a long contribution of the agent.

In contribution A$42 the travel agent, who is talking about the prices of flights and hotels to Brazil, says: then I do not know how it is going to be next year, and while uttering this contribution she looks at the customer and produces head shakes, shoulder shrugs and waves her hands. Even if these gestures have not been labelled as feedback elicit behaviour, they might have triggered the reaction of the customer who produces in fact a SINGLE HEAD NOD with the function FEEDBACK GIVE CONTINUATION YOU GO ON.

During her contribution A$44 the agent looks at the customer and produces several hand movements, which have not been labelled as feedback elicit gestures, but which however might have the effect of triggering more feedback from the customer. The customer in fact reacts with a short feedback word, with the function GIVE ACCEPTANCE.

In contribution A$46 the agent uses head movements to elicit feedback: she requires acceptance by producing waggles. This triggers the feedback acceptance of the customer in contribution C$77.

In the dialogue system, the agent did not produce any communicative non-verbal behaviour and this might be one of the reasons for the possible low production of FEEDBACK CONTINUATION YOU GO ON by the side of the users.

7.2.3.3 Acoustic Characteristics of Feedback

It is quite uncontroversial that acoustic cues can be used for the purpose of marking information structure at the discourse level. Acoustic correlates for different function of short expressions have been found for a variety of languages: English [Hirshberg & Nakatani 1996], Japanese [Ward &

Tsukahara 2000], Swedish, Italian [Cerrato 2002b; 2002c] and Dutch [Caspers 2003].

Unfortunately the audio quality of the recordings does not allow for accurate acoustic analysis of the speech materials, and as a consequence it was not possible to obtain precise measures of the F0 of the feedback expressions in the four GSLC dialogues. However it was possible to notice the following characteristics:

- feedback words such as *ja, mh,, nä*, produced in a non-intrusive way, have a rising pitch contour and are produced to signal an active listening attitude and show that listeners wish to respond without interrupting to take the turn;
- feedback words with slightly rising or rising-falling pitch contour show either the listeners' intention to take the turn or the acceptance of the information received;
- feedback words with varying pitch contour, such as rising-falling generally indicate an expressive feedback.

The characteristic rising pitch contour shown by feedback words produced in a non-intrusive way is a typical continuation contour. A raised F0 is considered a marker of non assertiveness [Ohala 1983] and in fact the minimal non-intrusive responses have been categorised as feedback CONTINUATION YOU GO ON (CY) which is a category that indicates the willingness to show attention and the intention to go on in the interaction, but not explicitly acceptance of the received information.

The non-intrusive minimal responses seem to be systematically produced in appropriate points of the production of the main speaker, namely in correspondence to a short pause and mostly at the end of a grammatical clause. If overlap occurs, it is only partial, because the feedback expression always starts before the start of the main speaker's new clause.

When speakers wish to convey a particular point of view with their feedback expressions, they tend either to produce reduplicated expressions, such as: *jaja, jaha, nähe, mhm* with varying pitch contour, or add a reinforcing expression to *ja*, such as *visst, precis, just det*.

In example 3, taken from GSLC Dial 1, the customer expresses her surprise in hearing about the information of getting a paid stop-over in Paris during her flight to Brazil.

A$96:	*de{t} är byte eller så är det övernattning i Paris/ man får eh/ man får den betald då*
C$97:	***jaha*!<FB;W;Gi;FBEx>**

Example 3 from GSLC Dial 1: jaha as FEEDBACK GIVE EXPRESSIVE.

In the GSLC human-human dialogues, the very few existing realizations of short expressions produced at the beginning of a longer utterance showed a slightly rising pitch contour.

In the human-machine interactions from the AdApt Corpus I, the short expressions were always produced at the beginning of a longer contribution and they showed a rising pitch contour.

7.2.4 Conclusions and Discussion

Notwithstanding the differences between human-human dialogues and human-machine interactions, it was possible to find some common ways of expressing feedback; in particular the expressions used to accept the information received appear to be used in both kinds of communicative situations. In human-human dialogues they are produced by both the interlocutors, in the analysed human-machine interactions the embodied agent did not produce any explicit feedback.

Short non-intrusive feedback expression were produced in a large number of contributions in human-human dialogues, but it was not possible to know whether they were produced as own contributions in the human-machine interactions or not.

As concerning the acoustic characteristics of feedback expressions, in particular pitch contour seems to reflect the function conveyed by feedback. These results provide important cues to the interpretation of the specific semantic-pragmatic function of users' feedback. These cues could be exploited in the development of more advanced speech-based interfaces, able to interpret and produce human-like behaviour related to feedback.

Given that feedback expressions have such an important role in human-human communicative interaction and since both participants in a dialogue produce them in great numbers, it is essential that speech-based interfaces, and in particular multi-modal dialogue systems displaying embodied conversational agents, should be able to recognize and interpret them and also produce them in an appropriate way. The appropriate production of feedback signals by the system has been proven to enhance not only the interaction between human users and dialogue systems [Takeuchi & Nagao 1993; Rajan et al. 2001], but also human satisfaction [Okato et al. 1998].

7.3 Study 2 Non-Verbal Behaviour in Human-Machine interactions

The second study presented in this chapter focuses on the analysis of the production of non-verbal behaviour of users in interactions with a Swedish multi-modal dialogue system with an embodied conversational agent. Besides non-verbal feedback phenomena, also non-verbal behaviour

signalling turn management and emotional attitudes are analysed in this study. This investigation is based on the assumption that users of multi-modal dialogue systems with an embodied agent tend to employ human discourse features when they interact with the embodied agent. On the basis of this assumption, it is hypothesised that the production of non-verbal behaviour that signals feedback and turn management, and the physical signals of emotional attitude of the users towards the system during the interactions, could be interpreted as an index of the fluency and naturalness of the interaction, and therefore used as additional metrics for user satisfaction.

This study is part of a wider investigation carried out with the aim of proposing alternative evaluation metrics for the evaluation of user satisfaction in interactions with multi-modal dialogue systems (see also [Cerrato & Ekeklint 2004]). These new metrics focus on the users, rather than on the system, the assumption being that the intentional use of prosodic variation and the production of communicative non-verbal signals by users can give an indication of their attitude towards the system and might also be used to evaluate the users' overall experience of the interaction with the system.

7.3.1 Materials

Video recordings of six users' interactions with the experimental multi-modal dialogue system AdApt were analysed for this study, which were selected from the second collection carried out with the AdApt system and are referred to as AdApt Corpus II (for more information see section 3.3.2). The users in the AdApt Corpus II were video recorded, as well as audio-recorded, both when listening to the instructions given by the test leader and when interacting with the system.

The six users (three female and three male) selected for this study belong to the sub-group of recordings of the AdApt system set-up with presence of the agent turn-taking gestures. In this set-up the agent produced gestures such as changing of gaze direction, eyebrow rising, head tilting to show when he was busy thinking and to signal turn-taking.

7.3.2 Method

The video recordings of the interaction between the six users and the dialogue system AdApt, as well as the video recordings of the instruction phase, were analysed in order to identify non-verbal feedback and turn-management phenomena.

The non-verbal communicative behaviour investigated in this study include facial displays, hand movements and body postures that users produce during the interaction with the system, with the specific

communicative function of giving or eliciting feedback, signalling turn management or showing an emotional attitude.

The analysis of the visual correlates of the emotional attitude carried out in this study is not based on the identification of the classic set of emotions proposed by Ekman [1993], but rather on the identification of visible expressions that might give an indication of the users' attitude towards the system. The user's attitude towards the system is judged considering two levels: one related to the involvement in the interaction and one related to the bodily behaviour during the interaction. At the involvement level, the user attitude towards the interaction is judged in terms of engagement, amusement and irritation. At the bodily level, the user behaviour is interpreted in terms of tenseness, tiredness and frustration.

Non-verbal feedback phenomena were identified and annotated using the coding scheme presented in chapter 4. For the annotation of turn-management phenomena the MUMIN coding scheme was followed [Allwood et al. 2005]. The categories for turn management are shown in table 7.1.

Table 7.1 Turn-management annotation features.

TURN MANAGEMENT	Turn gain	Turn take Turn accept
	Turn end	Turn yield Turn offer Turn complete
	Turn hold	Turn hold

The annotation of the emotional attitude was done looking at the body postures of the users during the interactions, and by using an open group of binary features, among which engagement-disengagement, boredom-amusement, annoyance-pleasure, frustration-satisfaction, irritation-calmness. It was assumed that each user has a neutral emotion by "default" when starting the interaction with the system. The course of the interaction with the system may trigger other emotional attitudes. As a consequence the user might show some facial expressions and body postures that communicate boredom, tiredness, amusement, irritation, frustration and so on. However an in-depth analysis of the visual correlates of the users' emotional state is beyond the scope of this investigation[20].

[20] More detailed investigations aiming at analysing and reproducing objective visual correlates of emotional states in talking heads and aiming at providing technological baselines and methodologies for comparative evaluations of visual correlates of emotional speech in talking heads have been carried out in the framework of the European project PF-Star.

The analysis and annotation of the non-verbal communicative behaviour was performed with the support of WaveSurfer with a video-plug in, which facilitates the analysis of visually accessible information in temporal alignment with speech.

7.3.3 Results

A comparison of the distribution of feedback production in the instruction phase and in the interactions with the multi-modal system was carried out. A subjective judgement of conversational fluency was obtained by looking at the answer given by the selected six users to a question about their experience of the smoothness of the interaction with the system. Finally, the non-verbal expressions that signal feedback and turn management and the physical signals of emotional attitude of the users towards the system were related to an overall measure of user satisfaction, which had been previously calculated on the same users' interactions.

7.3.3.1 Non-Verbal Feedback and Turn management

The amount of communicative non-verbal behaviour that a speaker produces might depend on the personal keenness to produce non-verbal behaviour. The only way to verify this is to have some kind of baseline measurement of the user's non-verbal behaviour production in human-human communication.

Besides the recording of the interactions with the dialogue system, a five-minute recording for each user listening to the test leader's instructions, before starting the actual interaction with the system, was available. While the test leader reads the instructions, all the users showed an active listening attitude, which means that they gave verbal and non-verbal feedback to the person instructing them. Their communicative non-verbal behaviour consisted mainly of SINGLE NOD and REPEATED NODS, SINGLE JERKS and SMILES with the function of giving FEEDBACK CONTINUATION YOU GO ON or FEEDBACK ACCEPTANCE. However, few turn-management signals were also produced, mainly head movements.

Figure 7.5 shows the total number of communicative non-verbal behaviour produced during the instruction phase, compared to the instances of non-verbal communicative behaviour produced during the interaction with the dialogue system. All users produce more non-verbal communicative behaviour while listening to the test leader, than when interacting with the system, even if the interaction with the test leader lasted between 4 and 5 minutes and the interaction with the system lasted between 25 and 30 minutes.

It might be argued that the 5-minute recordings of each user listening to the test leader represent a different communicative situation compared to the interaction between the users and the dialogue system, since while in the

interaction with the test leader the users mainly show a listening attitude, in the actual interaction with the system they have a more active role which put them in the position of asking the system for different information regarding the available apartments in Stockholm city. However the results of this comparison show that the production of non-verbal communicative behaviour in human-machine interaction is quite low compared to the production of non-verbal behaviour by the same users in interactions with a human being.

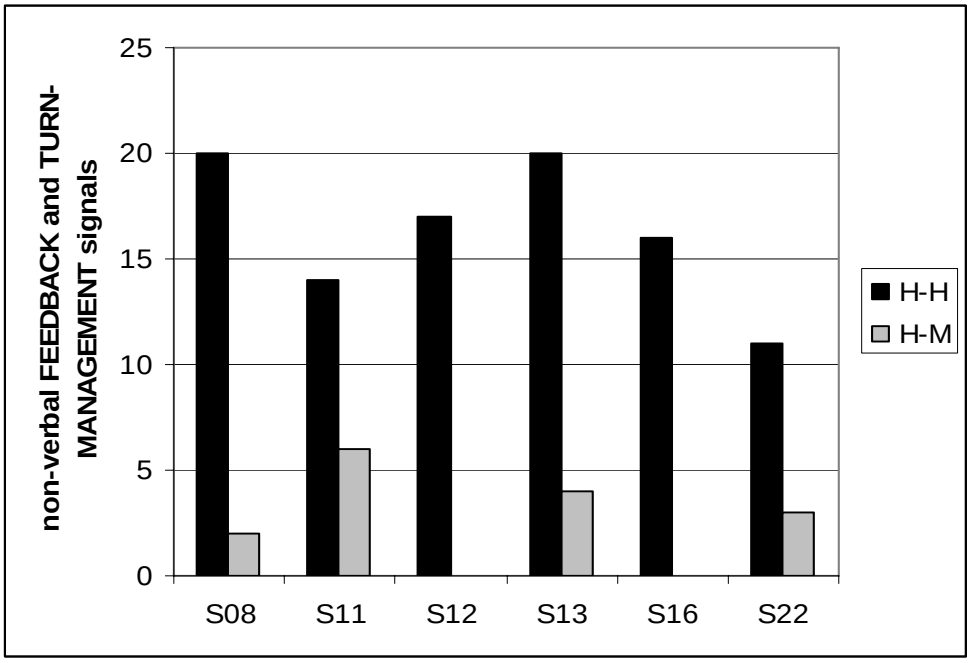

Figure 7.5 Number of instances of non-verbal feedback and turn management produced by each user in the instruction phase (human-human: H-H) and in the interaction with the system (human-machine H-M).

7.3.3.2 Conversational Fluency

Conversational fluency can be defined as the smooth unfolding of conversation, which can depend on several objective factors, such as:

- time spent repairing communication (breakdowns): if need for clarification is low then fluency is high;
- exchange of information and ideas: if information is easily and successfully shared, then fluency is high;
- speaking time shared: if the interlocutors use an equal time talking and if few silences and few interruptions occur, then fluency is high.

Since a scarce feedback production can cause communication breakdowns, it is possible to suppose that there is a connection between feedback production and fluency.

However, a subjective judgement of conversational fluency can also be obtained by directly asking the users of a given system about how they experienced the interaction with the system in terms of smoothness or fluency of the interaction. Under the second acquisition of the AdApt database, at the end of each interaction with the system, each user was asked to fill in a questionnaire containing several questions that could catch the subjective judgement of the user. The answer to the following question: "Did you think that the interaction with the system worked smoothly?" was taken into account in order to get an indication of how the users experienced the interaction with the system. The answer to this question can relate to several factors which might be correlated to the quantitative measures used for evaluation (number of communication breakdowns, task success, number of user contribution, word error rate and so on), but can also relate to whether or not the users feel at ease while interacting with the system. This condition of "feeling at ease" can also be influenced by many factors and it is not necessarily correlated to the quantitative measures used for evaluation; it might be, for instance, influenced by the visual realism of the embodied agent, by the emotional condition of the user, by the particular interest or by how familiar the user is with machine interaction, by the topic of the interaction and so on.

7.3.3.3 Emotional Attitude

It is here assumed that the physical signals of emotional attitude of the users towards the system during the interactions can give an indication of whether or not they were feeling at ease while interacting with the system. Therefore it is hypothesized that if the users do not feel at ease when interacting with the system, they tend not to produce non-verbal communicative behaviour related to feedback and turn management, which in turn can lead to an unsmooth unfolding of the interaction.

Important cues to emotions can be given by the posture that the user holds during the interaction. This assumption is supported by some results obtained in the field of psychology and AI [Laban 1976; Damasio 1994; Höök 2002] according to which emotions reside both in the mind and in the body of human beings. This means that emotion felt as a state can be displayed by means of facial displays, hand gestures and body postures; as a consequence, it is possible to suppose that particular gestures and body postures might encourage or constrain the expression of emotions. This assumption seems to be supported by the bodily behaviour of the six users in the AdApt Corpus II, reported in the last row on table 7.2.

During the interaction with the system, five of them sat mostly with their arms crossed in a tense and still position, or were constrained by holding the

microphone in one hand and holding the other hand under their chin. This posture may have limited them in their production of communicative non-verbal behaviour during the interaction with the system. For instance, user S12, who does not produce any communicative non-verbal behaviour during the interaction, appears very tense. She keeps still during most of the interaction, she hyper-articulates speech and gives the impression of being quite frustrated when the system does not understand her requests.

Users S22 and S13, who both have rather high user satisfaction scores, appear very engaged during the first half of the interaction with the system, and it is in this first half that they make their communicative non-verbal behaviour. User S22 in the second half of the interaction shows evident signs of tiredness, such as placing one of her hands under her chin or sinking down on the desk hanging on her elbows. User S16 does not appear very engaged in the interaction; he yawns several times, and fidgets in the chair. He gets very bored during the interaction and this is signalled both through his voice, which undergoes some variations in pitch (lowering), and through his facial expression and mainly his body postures: corners of the mouth downward, gaze directed in space, hanging body posture with his head bent on one side, or supported by a hand positioned under his chin or holding one side of his face.

7.3.3.4 Non-Verbal Communicative Behaviour and User Satisfaction

The number of non-verbal feedback and turn-management behaviour, the score given to the question about the smoothness of the interaction and the description of the emotional attitude of the users towards the system were related to an overall measure of user satisfaction, calculated in an investigation previously performed on the same materials [Hjalmarsson 2002]. This earlier investigation consisted of an assessment of the PARADISE paradigm [Walker et al.1997] as a feasible evaluation tool for the AdApt system. The PARADISE paradigm includes several metrics for the evaluation of user satisfaction, consisting of a combination of dialogue quality measures, dialogue efficiency measure and task success. Some changes were made to the metrics proposed in the PARADISE paradigm in order to adjust them to the specific features of the AdApt system. The data for the users' satisfaction measure were collected by means of a user survey based on the questions proposed in the PARADISE paradigm, which concern both modular and subjective evaluation of the system: evaluation of TTS and ASR Performance, evaluation of Task Ease, User Expertise, System Response, Expected Behaviour and Future Use.

Some questions regarding the benefit of the specific graphical interface were added to the questionnaire with the intention to capture the multi-modal features of the AdApt System, for a total of ten questions. All question responses ranged from the value "no, almost never" to "yes, almost

always"[21].Each response was mapped to an integer (from 1 to 5) so that the final measure of user satisfaction would be given by a number ranging between 10 and 50: the higher the number, the more satisfied the user.

While in PARADISE a questionnaire was completed after each task, in the AdApt survey the users filled in just one questionnaire at the end of their interaction. For this reason the user satisfaction measure given in Hjalmarsson's survey can be considered as an overall evaluation of the dialogue and cannot be associated with a specific task.

The production of user's non-verbal communicative behaviour was related to the measure of overall user satisfaction obtained by Hjalmarrson. Table 7.2 shows the number of communicative non-verbal behaviour of the six users in interaction with the system --Feedback (FB) and turn management (TMn)-- mapped against the overall score of user satisfaction, and to the answer given to the question about smoothness of interaction.

None of the users seems to have enjoyed the interaction; in particular S08 did not seem to have liked it at all. This male user is in fact the one who gave the lowest score to the question about the smoothness of the interaction and obtained one of the lowest user satisfaction scores. Under the interaction with the system, this user appears quite irritated because the system does not reply to his requests.

The user satisfaction score shows cross-gender differences: female users all gave higher scores than male users.

[21] The whole scale was: no; almost never; seldom; sometimes; yes; often; yes almost always.

Table 7.2 Number of communicative non-verbal behaviour of the six users in interaction with the system, mapped against the measure of user satisfaction, the answer given to the question about smoothness of interaction, the judged involvement in the interaction and the bodily behaviour of the users.

	S08 M	**S16 M**	**S11 M**	**S12 F**	**S22 F**	**S13 F**
Non-Verbal FB &TMn in H-M	2	0	6	0	3	4
User satisfaction score (10-50)	20	23	26	27	30	31
Smoothness (1-5)	1	2	2	3	3	3
Involvement	Engaged at the start	Disenga-aged	Engaged	Engaged	Engaged at the beginn-ing	Engaged at the start
Bodily behaviour	Tense	Tense, still posi-tion	Tense, still posi-tion	Tense, still posi-tion	Comfortable at the start, Uncomfort-able at the end	Relaxed at the start

7.3.4 Conclusions and Discussion

With the exception of user S11, who produces six turn-taking gestures, consisting of eyebrow raising and a little head nod at the end of the uttered request to the system, the non-verbal communicative behaviour produced during the interactions by the other users is very little, if at all. This result is inconsistent with the assumption that users of multi-modal dialogue systems with an embodied agent tend to employ human discourse features when they interact with the embodied agent. One possible interpretation of this lack of communicative non-verbal behaviour is that the users, aware of the fact that the agent cannot see and interpret their non-verbal communicative behaviour, intentionally tend not to produce them. In fact user S11, who shows the highest number of non-verbal communicative behaviour, produces only visual cues related to turn management, which are probably triggered by the visual turn-management cues given by the agent.

Since the agent does not produce any visual cues related to feedback, the users tend not to do that either. This explanation seems to be consistent with the results reported in chapter 6 that showed that in human-human spontaneous dialogues, most of the non-verbal behaviour related to feedback are produced when there is eyecontact between the interlocutors. This kind of eyecontact was not possible to establish with the embodied conversational agent in the AdApt dialogue system, and this might have limited the visual triggering of feedback noticed in the human-human

dialogues. Moreover, in communicative exchanges, gazing is considered as a signal of attentiveness [Argyle 1988], while the lack of gaze is experienced as a sign of passiveness, disinterest. If this is the case, it is worth considering that an appropriate production of non-verbal human-like features from the side of the agent (in particular those related to FEEDBACK and TURN MANAGEMENT, which are strictly connected to gazing [Kendon 1967, Novick et al. 1996]) might not only trigger the production of non-verbal communicative behaviour from the side of the users, but also make the users feel that the agent is interested in the conversational interaction.

Several attempts to provide computer systems with the ability to produce non-verbal behaviour related to FEEDBACK and TURN MANAGEMENT have been already carried out [Rajan et al. 2001, Thorrisón 2002]. More recently, efficient and robust algorithms to recognize non-verbal behaviour related to feedback have also been proposed [Morency & Darrel 2006].

Enabling embodied conversational agents to interact with humans in an effective way requires both the understanding of how communicative non-verbal behaviour is naturally performed by humans and the possibility to capture the exact dynamics of the non-verbal behaviour produced by humans. Having available high precision data makes it possible to control communicative non-verbal behaviour in embodied conversational agents. In the next chapter the focus is on the collection and analysis of high precision data, able to capture the dynamics of facial displays produced by humans in natural conversations.

8 3D-Multi-modal Corpora

8.1 Introduction

In this chapter the collections of 3D-data acquired by means of the opto-electronic motion capture system Qualisys Mac Reflex[22] are presented. The system is able to capture the dynamics of facial displays with high precision.

The possibility of analysing data collected by means of the opto-electronic system has opened new frontiers in the investigation of facial displays that serve important communicative functions, such as: signalling the position of focus [Hällgren & Lyberg 1998; Beskow, Granström & House 2006], signalling feedback and turn management [Cerrato 2004], and showing different expressions of emotions [Nordstrand et al. 2004; Magno Caldognetto et al. 2004],

The first and second study presented in this chapter deal with detailed analyses of head movements, in particular head nods, related to FEEDBACK. In the third study an automatic method for the detection of head nods is illustrated.

Previous studies on head nods have mainly focussed on the analysis of the distribution and semantic function of head nods in conversational speech [Mc Clave 2000, Allwood & Cerrato 2003] and even of the physical properties of head nods. An attempt to quantify the extent of head movements was made by Birdwhistell [Birdwhistell 1970]. He assumed that all movements of the body, including head nods, are directly linked to linguistic structure and proposed a hierarchical system of units of movement in which lower-level units (kines) combined to form higher-level units (kinemes). A kine is an isolable feature of movement, while a kineme is defined as a group of movements having the same meaning in the American culture.

Birdwhistell based his category system on a model taken from the categories of verbal communication (allophone, phone, phoneme and morpheme), this way “a kineme is similar to a phoneme because it consists of a group of movements which are not identical, but which may be used interchangeably without affecting social meaning" [Knapp 1972, p. 94-95].

[22] Qualisys MacReflex Motion Tracking System: http://www.qualisys.se (June 2006)

Birdwhistell considered head nods as distinct kinesic units and estimated that a similar population of movers will make a full 15-degree nod which can extend from about 0.5 sec to around 1.5 sec. The velocity, not the duration, is significant here.

Birdwhistell defined also kinic variants with a velocity range from about 8 degrees per frame (1/24 sec) to around 3 degrees per frame. Head movements outside these specifications belong to different units. This attempt to quantify head nods is one of the first documented in the literature.

With the data recorded with motion capture techniques it is possible to obtain more precise and reliable data for the quantification of head nods. The recordings of spontaneous speech in real situations are, in fact, inadequate for detailed measurements of head movements and present several other limitations since they do not allow for total control of the acquired data and often present a series of distortions that can depend on the presence of background noise or the wrong placement of the recording apparatus, etc.

However the recordings of spontaneous speech by means of motion capture techniques is not constraint-free either. They need to be performed in a lab environment and require the use of elicitation techniques, which might limit the spontaneity of expressions.

Even if the lab environment does not allow for a complete spontaneity and naturalness of the interactions, it results in high precision and high quality data. Moreover the use of a specific elicitation technique, though limiting the spontaneity of the interaction, can lead to the acquisition of more controlled and structured data. Still a complete control over the production of non-verbal behaviour is unattainable, due to the impossibility to predict when exactly non-verbal behaviour is going to occur during speech production.

Data acquired with motion capture systems can be used to control facial displays in synthetic talking heads. The reproduction of facial displays can be performed by two different methods: one consists in re-using the registered dynamic sequences of natural recorded behaviour (this process is referred to as re-synthesis) the other consists in generalising them (this is referred to as data-driven process).

Re-synthesis [Beskow, Engvall & Granström 2003] is a simple process that allows for a very realistic reproduction of the registered facial displays. However with this method it is not possible to automatically reproduce non-verbal behaviour in co-occurrence with arbitrary speech text. This is possible if a data-driven strategy is employed for controlling communicative movements.

By means of a data-driven process it is possible to generalise from dynamic sequences of natural recorded behaviour and thereby better capture and model the variability that is present in human expression.

A data-driven technique has so far successfully been applied to control the articulatory movements of expressive speech in a Swedish talking head [Beskow & Nordenberg 2005; Beskow & Cerrato 2007].

By analogy with this data-driven strategy applied to obtain expressive speech, it seems likely to foresee that models for each head movement that is considered to have an important communicative function could be trained. This requires a large database of annotated movements for training and testing the models. Collecting a structured database of head movements is not an easy task, first of all because it is not possible to completely control the production of non-verbal behaviour, and secondly because the manual annotation of non-verbal behaviour is a time-consuming and quite subjective task.

One possibility to facilitate the annotation process is to carry out an automatic detection of non-verbal behaviour. The aim of the third study presented in this chapter is therefore to propose a method for the automatic detection of communicative head nods. This method can be used to obtain an adequate number of items to train and test a model for implementation in talking heads.

8.2 Study 1-Head Movements Signalling Feedback

The aim of this study is twofold: to explore the feasibility of using the Qualisys Mac Reflex motion tracking system to acquire dialogic speech as well as to investigate the potential of analysing how specific head movements are used to signal feedback and show evidence that it is possible to measure and quantify the entity of these movements in the acquired data[23].

So far recordings with optoelectronic systems had mainly focussed on the acquisition of short prompted utterances for the purpose of studying articulation [Hällgren & Lyberg 1998; Magno Caldognetto & Zmarich 1999, Granström, House & Beskow 2002], but also for the purpose of estimating face motion from the speech acoustic [Yehia, Kuratate & Vakiotakis-Bateson 2002]. The novelty of the method here proposed consist in the acquisition of semi-spontaneous dialogic speech using an opto-electronic motion tracking system for the purpose of studying feedback phenomena. The data collected in the first study presented in this chapter are analysed with focus on those head movements that co-occur with the production of short verbal feedback expressions. By looking at the curves representing each movement, it is possible to see whether:

- there is a one-to-one relationship between a specific feedback verbal expression and the co-occurring head movement;

[23] The collection of the data used in this study and part of the analysis were conducted together with Mustafa Skhiri [Cerrato & Skhiri 2003].

- each category of movement shows a general pattern;
- there is a one-to-one relationship between a specific head movement and a specific feedback semantic-pragmatic function.

8.2.1 Materials

The study of multi-modal communication requires much effort in collecting, processing and analysing data. Given the high complexity of multi-modal communicative behaviour, it is not easy to create recording set-ups and use elicitation techniques that are feasible for the acquisition of spontaneous controlled speech; as a consequence there are no standardized techniques or procedure that could be applied. For this reason the first data acquisition with the motion capture system was mainly an attempt to explore the potential of the recording set-up and of the elicitation technique for the purpose of acquiring dialogic speech.

The set-up used for the data acquisition presented in this chapter allows the recording of audio and visual data: audio data is recorded on a DAT-tape and visual data is recorded both by means of a digital video camera and with the optical motion tracking system Qualisys Mac Reflex.

Attaching infrared reflecting markers to the subject's face enables the system to register the 3D-coordinates for each marker at a frame-rate of 60Hz, i.e. every 17ms. The markers are ca. 5 millimetres wide and reflect infra-red light, this way they are visible in the dark and easily traceable by the four infra-red sensitive cameras.

3D recordings are obtained by combining 2D information calculating the 3D coordinates from the four cameras with different viewing angles. Before any measurements can be taken, the system needs to be calibrated to determine the geometrical relation between the image planes of the cameras and the coordinate system of the volume to be measured.

A Sony digital video camera DCR-PC-115 E, focusing on the subject with the marker on his/her face was placed 2 metres away from him in the recording studio. The video-recording signal was digitalized before being used for the detailed analysis.

A microphone SHURE Model 16A was placed at an appropriate fixed distance from the subjects in order to assure good quality of audio recording.

The movements of the markers in three dimensions were stored together with the recorded acoustical and video signals.

The recording system cannot record chunks longer than 60 seconds, so the data are stored in several data files.

Thirteen hemispherical markers were used for these recording. Six markers were attached on the subject's face, two at the base of the neck to be able to register torso movements and five on a specially prepared spectacle frame that helped to recover the rigid 3D motion of the head (see figure 8.1 for a reproduction of the markers-set-up).

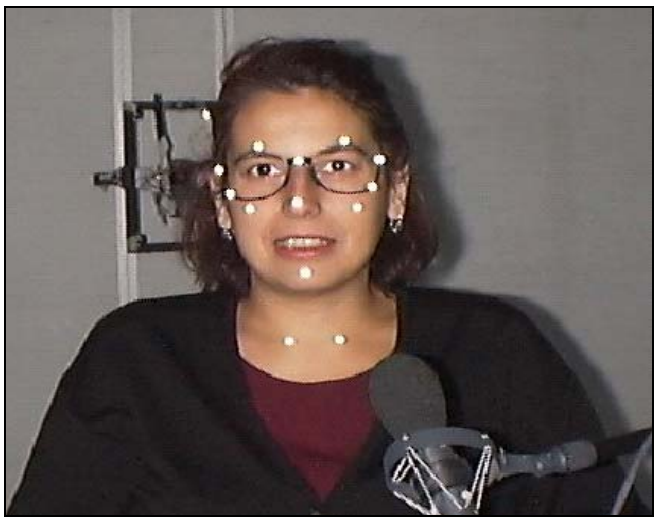

Figure 8.1 Position of the reflecting markers during the recordings[24]

With this first data acquisition 3 dialogues were recorded, unfortunately one of them could not be used for the analysis since during most of the dialogue the subject leaned to one side hanging her head on her shoulder. This way it was not possible to measure the head movements. For this reason only 2 dialogues could be used (3D-Dial 1 and 3D-Dial 2).

Two male subjects (from now onward subject-1 and subject-2) were recorded with the markers on their face. They were Swedish students at Linköping university of the age between 25 and 30. They were instructed to interact with a female interlocutor in a spontaneous way. Since it was foreseen that the results could be implemented in animated conversational agents, a communicative scenario similar to one that might arise between a user and an embodied conversational agent in a dialogue system was reproduced. In this scenario, that can be defined as “factual information seeking”, there are two dialogue participants: the "information seeker" and the "information giver". The information exchanged is relative to movies, plots, schedules and so on.

In both dialogues the "information giver" was the subject with the markers on his face, and the information seeker was his interlocutor, a female speaker, who asked several questions about the plots of movies, the actors and so on. Each recording session lasted about 14 minutes. During the first couple of minutes of each session the participants got acquainted with each other and with the recording environment, in order to feel at ease when starting the actual task. None of the subject thought that wearing the markers and the glasses during the recording was uncomfortable. 3D-Dial 1 lasts 10.30 minutes and subject-1 produces 97 contributions in it. 3D-Dial 2 lasts 7 minutes and subject-2 produces 80 contributions in it.

[24] The person in figure 8.1 was not one of the subjects recorded in the dialogues, she only poses to show the position of the reflecting markers.

8.2.2 Method

The video recordings acquired with the digital camera and those obtained with the tracking system were not automatically synchronised. In order to be able to synchronise them it was necessary to make a tagging of the phenomena to be analysed on the digital video. This was done by inserting tagging points by means of Multitool [Allwood et al. 2003].

The two dialogues were orthographically transcribed with the support of Multitool. Once an expression was identified as FEEDBACK and a head movement related to it was tagged, the categories reported in table 8.1 were used for the annotation.

To code the possible function of the identified head movements related to feedback the labels reported in table 8.2a and 8.2 b were used. These are the same categories illustrated in chapter 4.

Table 8.1 Categories and labels for head movements.

Type of non-verbal expression	
Category	**Labels**
SINGLE NOD (DOWN)	S-NOD
REPEATED NODS (DOWN)	R-NOD
SINGLE JERK (BACKWARDS UP)	S-JERK
REPEATED JERKS (BACKWARDS UP)	R-JERK
SINGLE SLOW BACKWARDS UP	BACKUP
MOVE FORWARD	FORWARD
MOVE BACKWARD	BACK
SINGLE TILT (SIDEWAYS)	S-TILT
REPEATED TILTS (SIDEWAYS)	R-TILT
SIDE-TURN	SIDE-TURN
SHAKE (REPEATED)	SHAKE
WAGGLE	WAGGLE
OTHER	OTHER

Table 8.2a Labels used to code the explicit semantic-pragmatic function of expressions that give feedback.

FEEDBACK GIVE	
Category	**Labels**
CONTINUATION **I** GO ON	**FB**Gi**CI**
CONTINUATION **Y**OU GO ON	**FB**Gi**CY**
ACCEPTANCE	**FB**Gi**A**
NON-**A**CCEPTANCE (**R**EFUSAL)	**FB**Gi**R**
EXPRESSIVE	**FB**Gi**Ex**

Table 8.2.b Labels used to code the communicative function of expressions that elicit feedback.

FEEDBACK ELICIT	
Category	**Labels**
CHECK ATTENTION	**FB**E|**ChA**
REQUIRE ACCEPTANCE	**FB**E|**RA**
MORE INFORMATION	**FB**E|**M**

The coding was displayed, in alignment with the dialogue transcription, on the multi-tier partiture of Multitool.

Each movement which co-occurred with a verbal feedback expression was tagged with Multitool, and thanks to this tagging it was possible to isolate the coded movement in the measurement data recorded with Qualisys. Unfortunately since the recording system cannot record chunks longer than 60 seconds, and the dialogues were much longer than 1 minute, it has occurred that c.a. 16% of the non-verbal expressions tagged on the digital video recordings were not found in the tracking data, since they occurred exactly in the break points.

The measurement data are rich in information related to the different movements the subjects made. The thirteen markers can, in fact, gauge both head movements and other facial displays (like eyebrow movements, cheek displacement and so on). However for this study the analysis is limited to the data related to the head movements interpreted as FEEDBACK. To analyse these movements the coordinates of the marker placed on the middle of the glasses were considered.

For each identified head movement a 2D curve was plotted. The curve displays the amplitude of the head movement in millimetres on the Y axis and the duration of the gesture in milliseconds on the X axis.

The curves were plotted by means of WaveSurfer.

8.2.3 Results

Table 8.3 shows a list of the occurrence of HEAD MOVEMENTS with FEEDBACK function that were observed and coded in the audio-visual recordings.

Table 8.3 Occurrences of the head movements related to feedback

Head movement	**Related expression**	**Feedback Function**	**Subject-1**	**Subject-2**
SINGLE NOD	*ja, mh, ja visst*	Gi**CY,** Gi**A**	4	2
REPEATED NODS	*ja, mh, ja visst, just det*	Gi**CY,** Gi**A,** El**RA**	7	3
SINGLE JERK	*ja, jaha*	Gi**CY,** Gi**A**	4	2
SHAKE	*nej*	Gi**R**	6	2
WAGGLE	*jag vet inte, jag tror inte det*	Gi**Ex,** El**RA**	2	1
SIDE TURN		Gi**CY,** Gi**Ex**	3	1

A total of 32 head movements related to feedback were tagged on the digital video recordings for subject-1, of these 26 could be analysed looking at the 3D data.

A total of 13 head movements related to FEEDBACK were tagged for subject-2, of these 11 could be analysed looking at the 3D data.

The total number of head movements related to FEEDBACK tagged for each subject represented 50% of their entire production of head movements in the dialogues. Beside head movements related to FEEDBACK, the two subjects produced head movements to signal FOCUS, TURN-MANAGEMENT functions and in co-occurrence with yes-no answers. FOCUS was signalled by a head nod and turn management was mostly signalled by sequences of quick head movements, such as side turn, co-occurring with gaze direction and eyebrow movements. A forward or backward movement of the whole trunk could also be used to signal turn management. POSITIVE ANSWERS co-occurred mostly with repeated head nods and negative answers with shakes.

Feedback words such as *ja* and *mh* produced with the semantic-pragmatic function CONTINUATION YOU GO ON (CY) often co-occur with NOD and JERK. These movements have been coded as "single" or "repeated". In the case of a head nod for instance, this means that it is possible to observe the head go up and down once (SINGLE NOD) or more than once (REPEATED NODS).

The difference between SINGLE NOD and REPEATED NOD is evident to detect on the measurement data, since every nod is represented by a single arc/peak as shown in figure 8.2, for a SINGLE NOD and figure 8.3 for a REPEATED NOD.

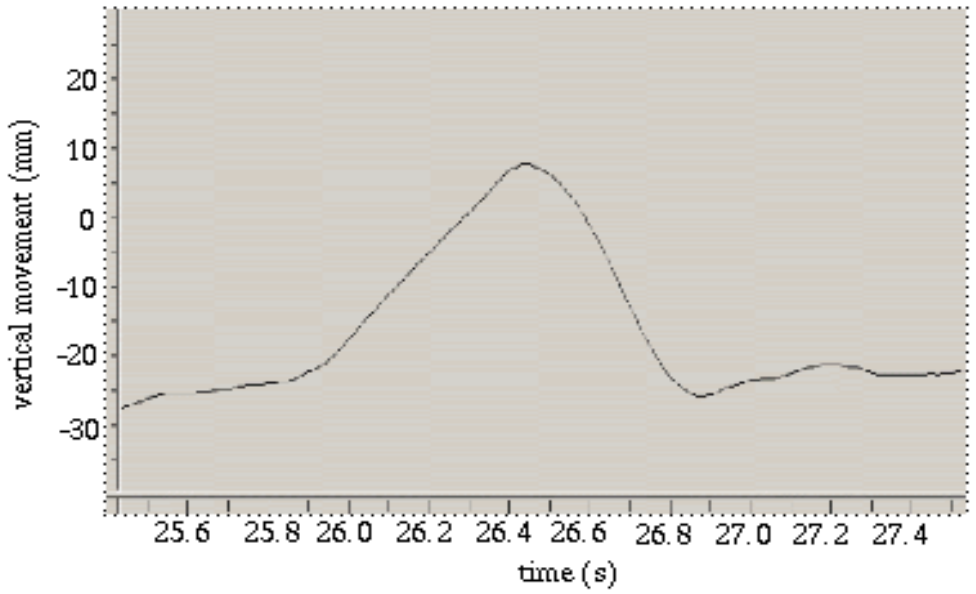

Figure 8.2 Curve of a SINGLE NOD *produced together with the feedback expression "mh" with* GIVE CONTINUATION YOU GO ON *function, by subject-1.*

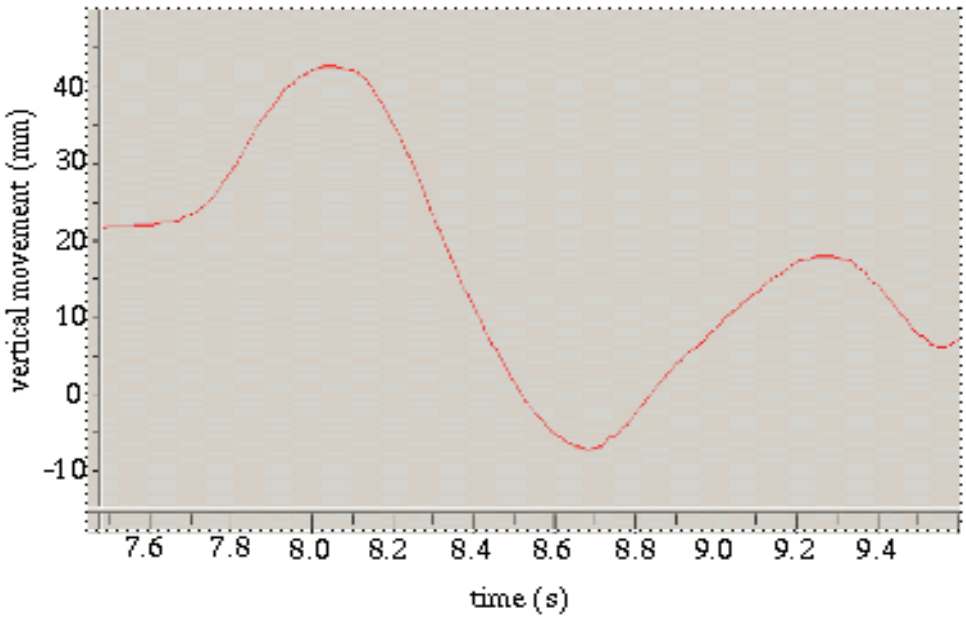

Figure 8.3 Curve of REPEATED NOD *produced with the feedback expression "ja visst" with* GIVE FEEDBACK ACCEPTANCE *function, by subject-1.*

The curves shown in figure 8.3 and figure 8.4 represent the movement coded as REPEATED NOD produced by subject-1 and subject-2 respectively. This REPEATED NOD co-occurred with two different verbal feedback expressions: *ja visst* and *just det* respectively. These two expressions were used to convey the same communicative function, which is GIVE FEEDBACK

ACCEPTANCE. This exemplifies that it is not possible to establish a one-to-one relationship between a specific verbal feedback expression and a specific head movement. However even if the curve in figure 8.4 shows three peaks and the curve in figure 8.3 shows two peaks (each peak corresponding to a nod) the two curves show a similar shape.

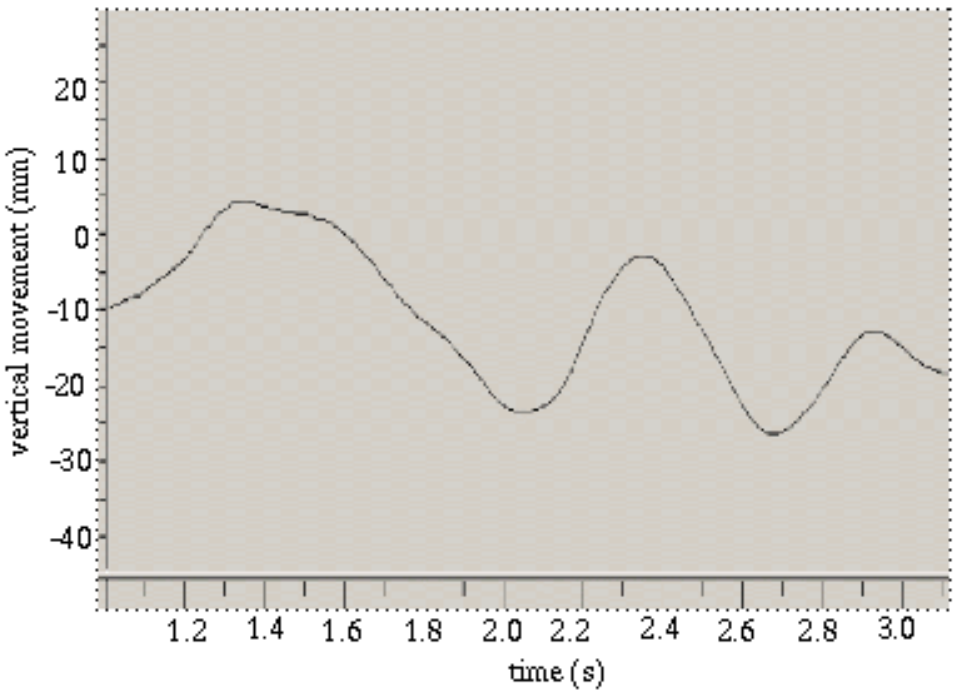

Figure 8.4 Curve of REPEATED NOD produced with the feedback expression "just det" with GIVE FEEDBACK ACCEPTANCE function, by subject-2

Figure 8.5 shows the curve for another head movement produced by the same speaker: a JERK produced when saying *ja* with CONTINUATION YOU GO ON function.

The curves representing the SINGLE NOD in figure 8.2 and JERK in figure 8.5 show different characteristics. For the SINGLE NOD the displacement of the marker placed on the middle of the glasses is of 36 mm, while for the JERK the displacement is 55 mm. The average displacement in mm for the four instances of SINGLE NOD produced by subject-1 is 30 mm, while for the four instances of JERK is 52 mm. The average duration of the SINGLE NOD is 0.25 seconds, of the JERK is 0.36 seconds. These results seem to support the idea that it is possible to identify a general pattern for each specific movement.

The curves in figure 8.2 and 8.5 are different, but both the head movements (SINGLE NOD and JERK) that they represent were produced by subject-1 with the same function: CONTINUATION YOU GO ON. This result exemplifies that it is not possible to establish a one-to-one correspondence

between a movement and a function. In fact different movements can be produced with the same function.

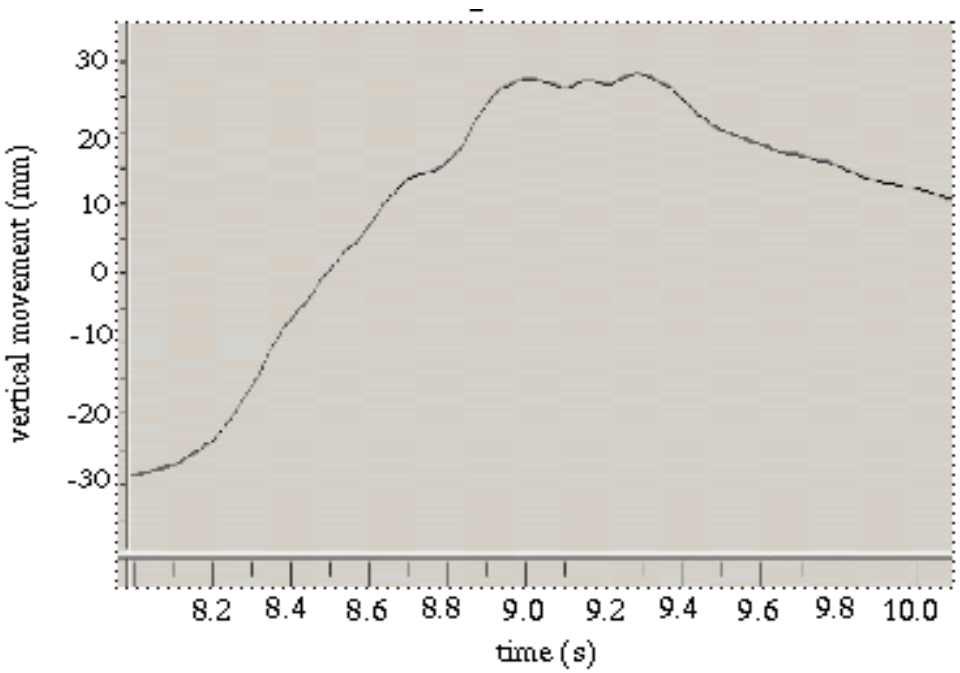

Figure 8.5 Curve of a JERK, which is a backward movement of the head, produced together with the feedback word "ja" with CONTINUATION YOU GO ON function, by subject-1.

8.2.4 Conclusions and Discussion

This first study was mostly explorative and aimed at evaluating the feasibility of the acquisition method and the elicitation technique for the collection of interactive dialogues. The usefulness of the acquired data for the aim of analysing head movements and providing data to control facial display in synthetic conversational agents was also looked at.

The recording set-up used in this data collection led to the acquisition of data related to the extent and duration of the different head movements; the elicitation technique appeared to induce the subject to interact in a quite spontaneous way, even if sometimes the interlocutors make long pauses to think about what to say.

The number of examples of head movements analysed in this study is quite limited; this limitation is partly due to the constraints of the recording set-up and partly to the fact that elicitation of spontaneous non-verbal communicative behaviour is not an easy matter. While different eliciting techniques have been developed to induce subjects to produce more or less spontaneous speech in somewhat controlled situations [Gybbon, Mertins & Moore 2000; Campbell 2001], it is still very hard to define a good method of eliciting spontaneous facial displays and other communicative non-verbal

behaviour in a controlled and time-limited recording session. This is primarily due to the fact that communicative non-verbal behaviour has not been studied and coded as much as speech, so it is quite difficult to predict exactly when they might occur in prompted semi-spontaneous and spontaneous speech.

Despite the limited number of identified head movements, it was possible to observe that a one-to-one relationship between a specific verbal feedback expression and a specific head movement cannot be established since different feedback words can co-occur with the same head movement (for instance *ja* and m-like words can co-occur with SINGLE NOD) nor is it possible to establish a one-to-one relationship between a specific head movement and a specific semantic-pragmatic function. In fact different head movements can be produced to convey the same functions, as for instance SINGLE NOD and REPEATED NOD to GIVE FEEDBACK ACCEPTANCE or SINGLE NOD and JERKS to GIVE FEEDBACK CONTINUATION YOU GO ON. However it was possible to observe that short expressions with the function CONTINUATION YOU GO ON when produced in a minimally intrusive way tend to co-occur with minimal HEAD NODS and JERKS. This observation is consistent with the results shown in chapter 6 and points out that HEAD NOD and JERK are typical movements involved during the production of non-intrusive FEEDBACK.

HEAD NOD and JERK produce specific patterns and can be easily measured and quantified and eventually implemented in animated talking heads.

Even if the outcome of this first attempt was quite positive, it is important to point out also the limitations of this experimental set-up.

Some of the limitations were technical and could be improved in successive collections of data, for instance the fact that only the subject with the markers on his face was video recorded cut out the possibility to observe how dialogue participants mimicked each other's behaviour and how they exchanged gaze and so on. Mutual gaze and gaze direction are in fact important visual turn-management signals and play an important role for the production of feedback. In the successive data acquisition the video recordings were performed by using two video cameras in order to record both interlocutors.

Other constraints in this first data acquisition were due to the recording time buffer. The dialogues recorded were not controlled in terms of time, while the recording buffer was only 60 seconds long, which means that some of the actual dialogues got lost between successive recording-chunks. For this reason in the following acquisition-sessions the subjects were instructed and guided to interact in dialogues of the length of a maximum of 60 seconds.

Moreover in the successive data acquisition the video recordings made with the digital video camera were automatically synchronised with the tracking system, by means of a synchronisation signal produced by the

Qualisys system and recorded on one channel of the DAT-tape as well as on one audio channel of the video recordings.

Given the high precision of the measurements obtained from the data acquired with the opto-electronic system, it was foreseen that the acquired data could be useful to gain valuable knowledge into how to control facial displays in synthetic talking heads.

8.3 Study 2-Linguistic Functions of Head Nods

The second study presented in this chapter is a more thorough investigation of head nods, with the aim of finding which specific communicative functions, beside FEEDBACK, they can carry out in spoken Swedish and provide a precise description of head nods in terms of their shape and duration, which might be exploited for the implementation of more natural head nods in the design of embodied conversational agents [Cerrato 2005b].

Beside the investigation of the functions that head nods can convey, this study aims at testing the hypothesis that minimal feedback expressions having the feedback function of giving CONTINUATION tend to co-occur with minimal non-verbal expressions, while more composite verbal feedback expression, that carry out feedback function other than CONTINUATION, tend to co-occur with more extensive non-verbal expressions.

This hypothesis is based on previous results obtained analysing verbal feedback in English, Swedish, and Italian [Jurafsky et al. 1998; Cerrato 2002b; Cerrato 2003]. These results suggest that short verbal feedback expressions, such as *ja, sì*, and m-like words having the semantic-pragmatic function GIVE CONTINUATION YOU GO ON show shorter duration and lower energy than other more complex verbal feedback expressions having more complex feedback functions such as ACCEPTANCE, EXPRESSIVE. These short verbal expressions are intended to be non-intrusive and have the function of showing that the interlocutor is following the interaction and is not yet willing to express ACCEPTANCE or to take the floor, for this reason it is likely to hypothesize that the non-verbal expressions co-occurring with minimal unobtrusive feedback verbal expression might also be minimal.

8.3.1 Materials

In order to carry out a more detailed investigation of the realization of head nods it was necessary to perform a further data acquisition with the opto-electronic system Qualisys[25]. This data acquisition was carried out under the framework of the European project PF-Star[26], in which one of the

[25] This data collection was performed in collaboration with Jonas Beskow, Magnus Nordstrand and Gunilla Svanfeldt.

[26] PF-Star www.pfstar.itc.it (December 2006).

initial main activities was the collection of audio-visual speech corpora and the definition of annotation formats.

For the study reported here, ten short dialogues were collected. These short dialogues are part of the PF-Star Corpus 3 (see section 3.3.4). For these recordings one semi-professional Swedish actor served as the subject. It was decided to choose an actor since part of the data collected under the framework of the PF-Star project was meant to serve as a source for the study of visual correlates of expressive speech, so the actor was asked to utter several sentences with different expressions of emotions. For the acquisition of the dialogues, the actor was instructed to interact with one of the experimenters (with whom he was well acquainted) in the most possible spontaneous way. From now onward the actor will be referred to as subject-S and the experimenter as subject-M. The dialogues can be defined as semi-spontaneous since the two subjects had a short script describing the scenario and the task to perform, and they had to improvise the dialogue. The scenario given to the two dialogue participants was that of a "travel agency", the different tasks to perform consisted of asking information about travels, booking train and flight tickets and so on. (A list of the ten dialogues and a short description of each scenario is shown in section 3.3.3.4.)

The focus of the 3D recording was on the subject with the markers on his face. However the other subject was also recorded, by using two SONY DV digital video cameras. The two participants alternate in their roles, this way in five dialogues subject-S (the one with the markers on his face) plays the role of travel agent, while his interlocutor plays the role of the customer, and in the other five dialogues the roles are switched. In order to avoid the same buffer problem as in the first acquisition (see section 8.2.4) the interlocutors were guided to hold the dialogues of the length of maximum 1 minute with the help of an experimenter who signalled when there were 5 seconds left to the end of the recording buffer.

Each dialogue counts between 10 and 16 contributions per interlocutor. The total number of labelled head nods is 93 for subject-S and 101 for subject-M. Subject-S had 29 IR-sensitive markers attached on his face, of which four markers were used as reference markers, instead of the glasses. The marker setup in this data acquisition corresponds to MPEG-4 Feature point (FP) configuration (see section 3.3.4). This configuration is a compact and standardised scheme for describing human facial displays (i.e. movements of the face and of the head) and it is therefore more suitable for the reproduction of human facial displays in talking heads [Beskow et al. 2004b].

8.3.2 Method

Annotation, segmentation and measurement of the duration of head nods in the audio-visual material were carried out with the help of WaveSurfer

provided with a video plug-in, which allows seeing the video recordings in .mpeg format synchronised with the speech analysis panels. Moreover it is possible to display the chosen location dimension of the 3D data on another panel. The marker on the nose tip was used as reference for the detection of head nods.

Figure 8.6 is a screenshot of WaveSurfer showing subject S uttering one of the expressive utterances in PF Star corpus 2[27] with a *questioning* expression. The displacement of the marker on the nose tip is shown on the panel over the spectrogram.

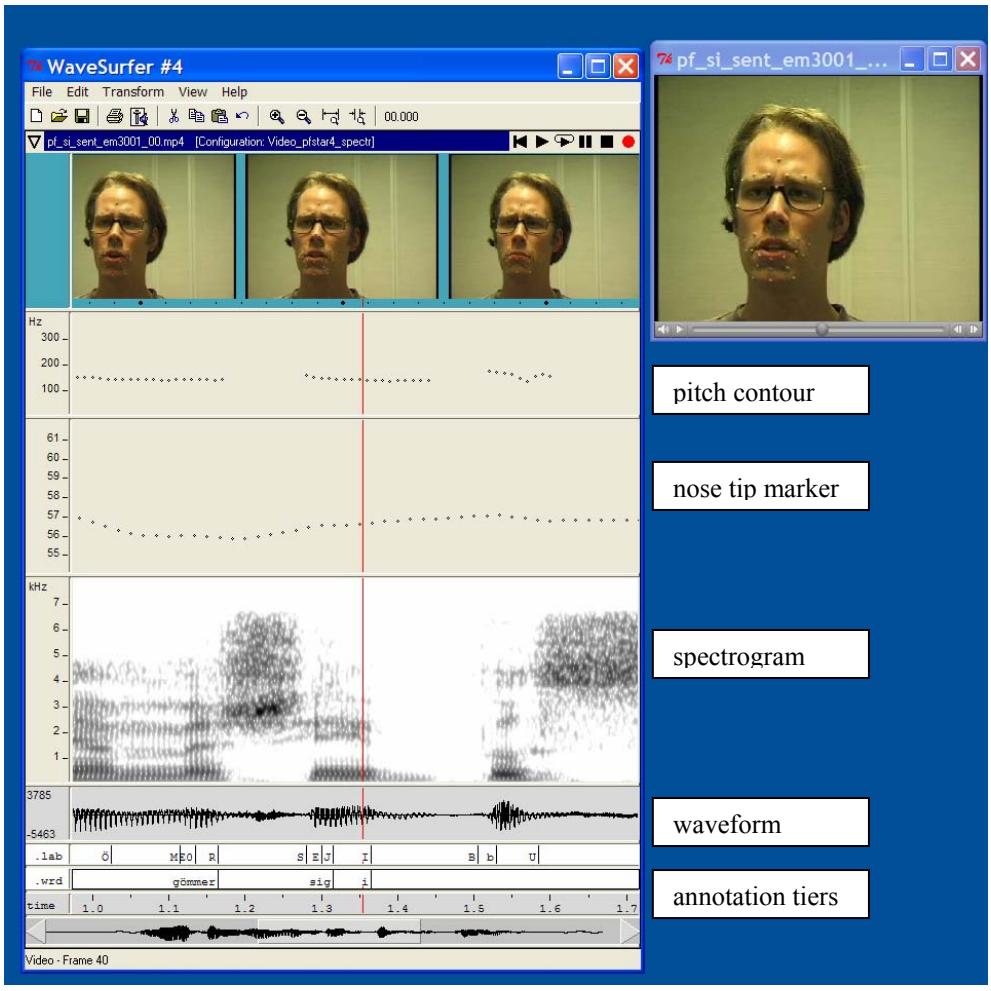

Figure 8.6 A screenshot of WaveSurfer with the video display in synchronization with the speech analysis panels and the transcription tiers.

The annotation tiers shown in figure 8.6 are only two: one for the phonetic transcription and one for the word transcription. For the annotation

[27] For the collection of PF Star corpus 2 the subject was wearing spectacles that had five markers attached to them, these markers served as reference to factor out head movements.

of the type and function of the head nods analysed in the second study reported in this chapter a multi-layer annotation was performed. This allowed annotating head nods in synchrony with the annotation of the verbal expression produced at the same time.

The multi-tier coding was carried out for each subject using the following five tiers:

Transcription: shows the orthographic transcription approximately segmented per contribution.
HM: is the tier for the coding of head movements.
V: is the tier for the transcription of the verbal expressions co-occurring with the production of head nods.
Semantic-Pragmatic function: is the tier for the coding of the semantic-pragmatic functions of the identified head nods.
Multi-modal relationship: is the tier for the multi-modal relationship annotated in terms of dependency or independency.

The head nod type could be SINGLE NOD (S-Nod) or a REPEATED NOD (R-Nods). The following function categories were defined *a priori*, based on previous observation of head nods and on literature references [Allwood & Cerrato 2003; Mc Clave 2000; Knapp & Hall 2002; Graf et al. 2002]:

- FEEDBACK,
- TURN MANAGEMENT,
- POLITENESS,
- POSITIVE ANSWERS,
- FOCUS,
- EMPHASIS,
- HESITATION.

Table 8.4 is a scheme of the category functions of head-nods and their labels used in the annotation.

Table 8.4 Coding scheme for the functions of head nods and their labels.

FUNCTION CATEGORIES	LABELS
FEEDBACK	**FB**
GIVE CONTINUATION (I GO ON)	GiCI
GIVE CONTINUATION (YOU GO ON)	GiCY
GIVE ACCEPTANCE	GiA
GIVE NON-ACCEPTANCE (REFUSAL)	GiR
GIVE EXPRESSIVE	ElEx
REQUIRE ACCEPTANCE	ElRA
CHECK ATTENTION	ElChA
DESIRE FOR MORE INFORMATION	El Mo
TURN MANAGEMENT	**TMn**
TURN GAIN	TG
TURN END	TE
TURN HOLD	TH
POLITENESS	**Pol**
POSITIVE ANSWERS (REPLY POSITIVE)	**RP**
FOCUS	**Focus**
EMPHASIS	**Emph**
HESITATION	**He**

The categories used to code the function of feedback expressions are those used in the first study and reported in tables 8.2a and 8.2b. The categories used to code TURN-MANAGEMENT phenomena are those proposed in the MUMIN annotation [Allwood et al. 2005] and used also in the annotation illustrated in section 7.3.

As concerns POSITIVE ANSWERS, labelled as RP, the difference between a positive feedback and a POSITIVE ANSWER is quite subtle. However the criteria followed to assign the label of affirmative response was that of looking for a positive answer to a polar question.

The label Pol is used to code the production of a single slow head nod, which has the function of showing politeness, courtesy. This slow nod is usually produced at the end of an interaction when the interlocutors greet and thank each other by saying some courtesy words, such as thanks, thank you.

The category HESITATION is intended for the annotation of head nods that might co-occur with hesitations and self-corrections.

Head nods are produced also to signal focus on words or constituents or to signal emphasis; in this case the categories have been defined as FOCUS and EMPHASIS, labelled respectively as Focus and Emph.

In order to code head nods and their functions it is necessary to identify them in the first place. To do so it is crucial to carefully analyse the digital video recordings and take contextual information into account, which means interpreting and categorising head nods in terms of explicit reactions to the

previous communicative act. For subject-S, the one with the markers on his face, it was possible to access the 3D data and have a more complete picture of the type of produced head movement and of their starting and ending point. For subject-M the identification relied only on the visual information given by the digital video recordings. Each identified head nod was assigned a function label.

Head nods produced with the function of feedback are easy to identify in the dialogues, since they often co-occur with short verbal feedback expressions such as: *ja* and m-like words.

For the head nods having other functions than feedback sometimes identification could be problematic, since they co-occur with different words or utterances and often simultaneously with other kinds of movements (as for instance with a forward or backward movement of the whole trunk, with eyebrow movements, and other facial expressions) or as a continuum sequence with other movements, as for instance before or after a jerk (i.e. a fast backward movement of the head) or a tilt (i.e. a single movement of the head leaning on one side) and so on.

8.3.3 Results

The distribution of the identified head nods is shown as the number of occurrences per function for each subject. Tables 8.5a and 8.5b show the distribution of SINGLE and REPEATED NOD respectively, and their semantic-pragmatic function for subject-S (the one with the markers on his face) in the role of agent and customer in all ten dialogues.

Tables 8.6a and 8.6b show the distribution of SINGLE and REPEATED NOD respectively and their semantic-pragmatic function for subject-M in the role of agent and customer in all ten dialogues.

In the ten dialogues each subject had both the role of agent and customer, so in order to see whether the production of head nods varies depending on the role the speaker has in the interaction, the distribution of head nods was calculated in all dialogues per subject and per role.

Table 8.5a Distribution of single head nods produced by subject-S in the role of agent and customer.

Function	**Subject-S Agent**	**Subject-S Customer**
	S-Nod	
FBGiA	5	3
FBGiCY	7	8
FBElRA	3	1
Pol	6	2
Emph	4	2
Focus	3	3
FBGiR		2
FBGiEx		1
FBGiCI	1	1
TMn		
Total	52	

Table 8.5b Distribution of repeated head nods produced by subject-S in the role of agent and customer.

Function	**Subject-S Agent**	**Subject-S Customer**
	R-Nods	
FBGiA	10	7
FBGiCY	2	
FBElRA	3	9
Pol		
Emph	2	
Focus	2	
FBGiR	3	
FBGiEx		
FBGiCI	1	2
TMn		
Total	41	

Table 8.6a Distribution of single head nods produced by subject-M in the role of agent and customer.

Function	**Subject-M agent**	**Subject-M customer**
	S-Nod	
FBGiA	5	6
FBGiCY	4	4
FBElRA	2	3
Pol	4	4
Emph	4	4
Focus	5	7
FBGiR	1	
FBGiCI		2
FBGiEx		
TMn: Turn Yield	3	
Total	58	

Table 8.6b Distribution of repeated head nods produced by subject-M in the role of agent and customer.

Function	**Subject-M agent**	**Subject-M customer**
	R-Nod	
FBGiA	9	8
FBGiCY	3	1
FBElRA	8	8
Pol		
Emph		
Focus	2	
FBGiR		
FBGiCI		1
FBGiEx		3
TMn		
Total	43	

The only relevant difference in the production of head nods that seems dependent on the role the speaker has in the interaction is the higher number of head nods produced to REQUEST ACCEPTANCE by subject-S in the role of customer, compared to the role of agent. Given the little difference noticed, the number of occurrences of head nods has been collapsed in figures 8.7 and 8.8, which show the distribution of SINGLE and REPEATED NOD respectively for each semantic-pragmatic function.

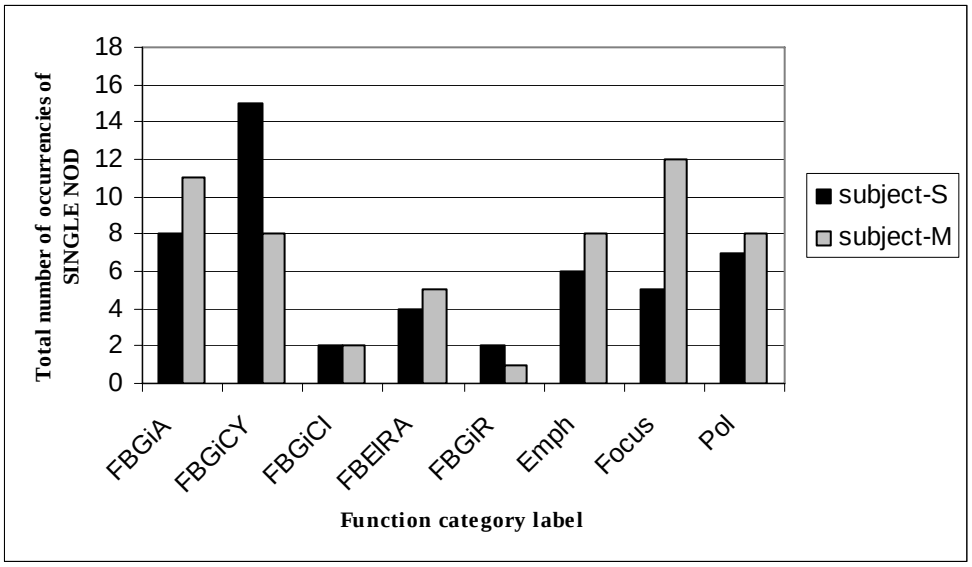

Figure 8.7 Total number of occurrence of SINGLE NOD for each semantic-pragmatic function for both subjects.

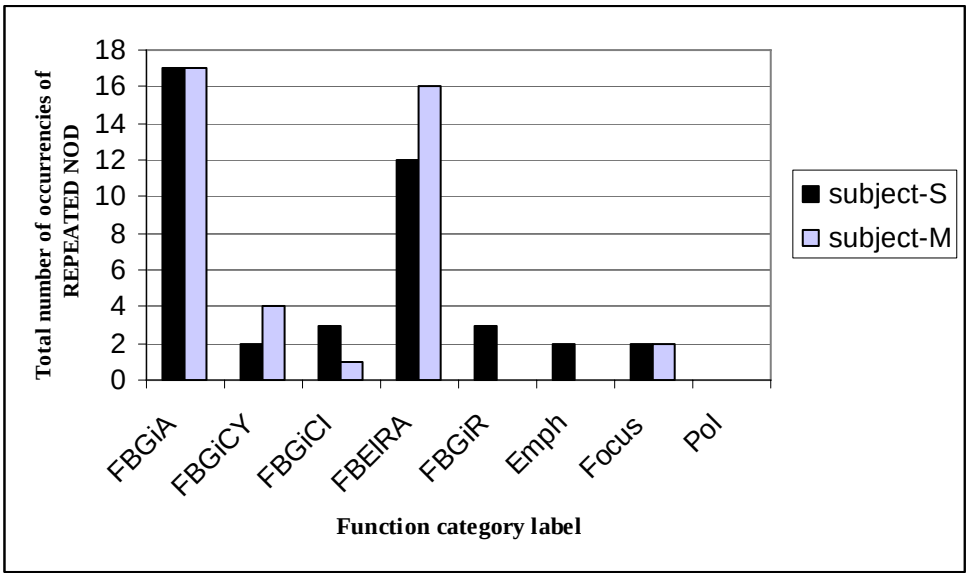

Figure 8.8 Total number of occurrence of REPEATED NOD for each semantic-pragmatic function for both subjects.

The total number of SINGLE NOD and REPEATED NOD is similar for both subjects: 93 for subject-S and 101 for subject-M, however the distribution of nods per function is slightly different across subjects.

Subject-S produces more SINGLE NOD with the function FEEDBACK GIVE CONTINUATION YOU GO ON (FBGiCY), while subject-M produces more SINGLE NOD with the function FEEDBACK GIVE ACCEPTANCE (FBGiA). Moreover subject-M produces more SINGLE NOD to signal FOCUS.

Both speakers produce REPEATED NOD to signal FEEDBACK, mostly to REQUIRE ACCEPTANCE (FBElRA) and GIVE ACCEPTANCE (FBGiA). Subject-S produces SINGLE and REPEATED NOD with the function NON-ACCEPTANCE (FBGiR), which means that the information received is not accepted, either because of misperception, misunderstanding or disagreement. Usually negative feedback is signalled by head shakes rather than by head nods.

There is a relevant difference in the distribution of SINGLE and REPEATED NOD in that SINGLE NOD is mostly produced as FEEDBACK GIVE CONTINUATION YOU GO ON (FBGiCY), this mostly for subject-S; very few times is REPEATED NOD produced with this function, it is instead frequently produced, by both speakers, as FEEDBACK GIVE ACCEPTANCE (FBGiA) and REQUIRE ACCEPTANCE (FBElRA).

SINGLE NODS are used by both subjects to show POLITENESS, which is never done with REPEATED NOD, however the SINGLE NOD produced with the function of showing POLITENESS is slower compared to the others.

Tables 8.7a and 8.7b show the average duration and standard deviation for SINGLE NOD according to different functions. Temporal values were measured for subject-S by looking at the displacement of the marker on the nose tip, and for subject-M by looking at the digital video recordings.

These results show that the SINGLE NODS produced with the function FEEDBACK GIVE CONTINUATION YOU GO ON are shorter compared to the SINGLE NODS produced with other functions.

The longest single nods are those co-occurring with courtesy words (Pol). Several examples of a SINGLE NOD produced to accompany a thank or a greeting word at the end of an interaction have been labelled in the 10 dialogues, in average these head nods that accompany courtesy words have a duration of 0.71 sec. Given its characteristics this head nod can be interpreted as a "reduced form" of a courtesy bow[28] [Eibl-Eibesfeldt 1970].

[28] Bowing is the act of lowering the head, or sometimes the entire upper body from the waist, as a social gesture. Bowing serves several functions: to greet, to defer, to show courtesy, and to pray. This is common around the world, but is especially prominent in Oriental cultures.

Table 8.7a Duration in seconds of SINGLE NODS per semantic-pragmatic function for Subject-S.

Function	Duration in seconds	Stand.dev.
Pol	0.65	0.19
FBGiA	0.64	0.22
Focus	0.58	0.20
FBGiCY	0.37	0.16

Table 8.7b Duration in seconds of SINGLE NODS per semantic-pragmatic function for Subject-M.

Function	Duration in seconds	Stand.dev.
Pol	0.78	0.14
Focus	0.57	0.16
FBGiA	0.50	0.17
FBGiCY	0.43	0.10

Because of the low number of instances of head nods it was not useful to run statistic analysis on the data. However, despite the low number of instances of head nods found in the analysed data, it is possible to observe interesting tendencies, for instance the results of the duration analysis give an indication that the duration of head nods may be related to the function they carry out in the communicative situation: when the function is FEEDBACK GIVE CONTINUATION YOU GO ON (FBGiCY) the head nods tend to be shorter than when the function is FEEDBACK GIVE ACCEPTANCE (FBGiA), FOCUS (Focus) and POLITENESS (Pol)**.**

8.3.3.1 Semantic analysis

In this session some examples extracted from the dialogues are reported and discussed in order to show when the head nod is produced and what its function/meaning in the given context is.

The first three examples are excerpted from PF-Star-Dial 4, in which subject-S plays the role of the customer and subject-M that of the travel agent. In this dialogue, which counts a total of 30 contributions, subject-S wishes to buy a packet-travel to Italy. He has previously been in contact with the travel agent and in the situation they meet in the agency to look at some possibilities and take a decision.

In example 1 subject-M says to the travel agent: *du skulle titta upp på några paket for mig och min fru* (you should look for some packets for me and my wife). The word *paket* (packets) carries the focus and co-occurs with a single head nod which has been labelled as Focus. At this request the travel agent (M) replies by saying *ja precis* (yes exactly), with the function FEEDBACK GIVE CONTINUATION I GO ON. This expression shows the intention of keeping the floor. Subject-M does not produce any head nod in this contribution, but he raises his eyebrows to show that he wishes to keep

the floor, he in fact continues by illustrating the possible alternatives in the following contribution: *det var lite olika alternativ där med all-inclusive och ehm* (there are some different alternatives with all-inclusive and ehm). Before subject-M continues to speak in contribution 11, subject-S produces a short m-like wordin his contribution 10. This short m-like word has the function FEEDBACK GIVE CONTINUATION YOU GO ON and does not signal the intention to take the floor. This expression co-occurs with a short single head nod of the duration of 0.22 sec. These head nods are quite minimal: they have an average duration of 0.40 sec and they usually co-occur with short verbal feedback expressions such as *yes, mh, ok*. These short verbal expressions with a CONTINUATION YOU GO ON function show shorter duration and lower energy than other more complex verbal feedback expressions having even more complex feedback functions (ACCEPTANCE, EXPRESSIVE). This is because they are intended to be unobtrusive and simply show that the interlocutor is following the interaction and is not yet willing to express a judgment or to take the floor.

In contribution 12 subject-S says *{j}a precis* (yes exactly) as a reaction to subject-M contribution 11 in which he illustrates the possible alternatives for the travel, in this case the verbal feedback expression co-occurs with repeated head nods (2 for precision) with the function FEEDBACK GIVE CONTINUATION YOU GO ON.

$S8:	*du skulle titta upp på några* ***paket*** **<W;S-Nod;Focus>** *för mig och min fru*
$M9:	***ja precis*** **<FB;Ph;EBRa;CI>**
$S10	***mm*** **<FB;W;S-Nod;Gi;CY>**
$M11	*det var lite olika alternativ där med all-inclusive och ehm*
$S12	***{j}a precis*** **<FB;W;S-Nod;Gi;CY>**
$M13	***{j}a*** **<FB;W;S-Nod;Gi;CI>**// *och det beror lite grann på vilken// vad ni kommer att välja då om ni vill åka till* ***Torino*****<W;S-Nod;Focus>** *eller om ni vill åka ner kanske till södra delen mot Rom*

Example 1 from PF-Star-Dial 4.

The second excerpt from PF-Star-Dial 4 offers instances of other categories of FEEDBACK. In contribution 25 subject-S, the customer, asks: *ja, men vad har du för alternativ då?* (yes, but what alternatives do you have, then?, referring to the package travels, and the travel agent answers: *ja* (yes) realized with a lengthening of the final vowel and accompanied by a short single head nod. The lengthening signals that the travel agent wishes to keep the floor and go on speaking to fulfil the customer request. In fact he continues his contribution by saying: *det finns den här tiodagarsresan då* (there is this 10-day package). In this utterance subject-S puts the focus on

tiodagarsresa by producing a short single head nod which has also the function FEEDBACK ELICIT REQUEST ACCEPTANCE (FBElRA).

In fact subject-M, as a reply, produces a short m-like word accompanied by a short single head nod with the function FEEDBACK GIVE CONTINUATION YOU GO ON; he produces this short feedback to show CONTINUATION YOU GO ON, but not yet ACCEPTANCE, since the information given by subject-M is not complete yet. The rest of the information comes in the following contribution when Subject-S, the travel agent, says that the package costs 8000 Swedish crowns per person.

$S25:	*ja men vad har du för alternativ då*
$M26:	*ja*[+] **<FB;W;S-Nod;Gi;CI>** / */det finns den här* ***tiodagarsresan*** **<FB;R-Nod;El;RA>** då/
$S27	***mm*** **<FB;W;S-Nod;Gi;CY>**
$M28:	*som / den* ***kostar*****<W;S-Nod;Focus>** *åttatusen per person*
$S29:	***{j}a just de{t}*****<FB;W;Gi;A>**

Example 2 from PF-Star-Dial 4.

Example 3 from PF-Star-Dial 4 is an instance of a single head nod with the function FEEDBACK GIVE ACCEPTANCE (FBGiA). Subject-M, the customer, asks: *vilken flygbolag är det?* (which flight company is it?) and the travel agent answers: *{j}a det är SAS* (yes, it is SAS), emphasizing the word SAS which is simultaneously produced with a short single head nod. SAS is the Scandinavian Airlines, which in this context represents a kind of warranty for a good and safe flight. To this the customer says: *ok*, accompanied by a short single nod, which has the function FEEDBACK GIVE ACCEPTANCE (FBGiA).

S$ 36:	*vilken flygbolag är det?*
M$37:	*{j}a det är* ***SAS*** **<W;S-Nod;Emph>**
S$38:	***ok*** **<FB;W;S-Nod;Gi;A>**

Example 3 from PF-Star-Dial 4.

Example 4, is excerpted from PF-Star-Dial 10, and shows an instance of the feedback category GIVE EXPRESSIVE, which means a feedback that explicitly shows an expressive/emotional reaction.

In this dialogue Subject-M is the customer, he is complaining about the fact that during a holiday in a holiday resort managed by the travel company he and his wife got stomach flu, as a consequence he is asking for a reimbursement. The travel agent tries to clarify that they cannot pay any refunds since they cannot be sure that the stomach flu was actually caused by the food and drinks consumed at the resort. As a reaction to this refusal by the travel agent the customer, quite annoyed, almost blackmails the travel agent by saying: *ok då får jag väl vända mig till konsumentombudsmannen*

då (ok then I will contact the Consumer Protection Organization) with a annoyed tone of voice and producing repeated head nods.

S$23:	***ok då får jag väl vända mig till konsumentombudsmannen då* <FB;R-Nods;Gi;Ex; /El;RA>**
M$24:	*ehm* ***ja och det är det är väl kanske det du får väl göra då* <FB;S;R-Nods;Gi;A>**

Example 4 from PF-Star-Dial 10.

This is an example of FEEDBACK GIVE EXPRESSIVE which at the same time is also a REQUEST FOR ACCEPTANCE, since the customer is not only showing his annoyed attitude, but he is actually trying to get a reaction from the travel agent. To this the travel agent replies: *ehm ja och det är det är väl kanske det du får väl göra då* (ehm ok maybe this is what you can do) and also produces repeated head nods. This is also a good example of how the production of head nods by one interlocutor triggers the production of head nods of the other interlocutor.

In example 5, from PF-Star-Dial 3, the two interlocutors have concluded their interaction and the travel agent says: *då bokar jag en sån* (so I will book it). To this the customer reacts by saying: *ok tack så mycket* (ok thank you so much) and produces a head nod while uttering the word *tack* (thanks), to which the travel agent replies with another *tack* accompanied by a single slow head nod. These two head nods have been labelled as POLITENESS (Pol), since they have been produced to thank the interlocutor and show courtesy. These head nods are in average respectively 0.76 and 0.67 seconds long, which is longer than the head nods produced with FEEDBACK functions.

S$38:	*då bokar jag en så{da}n*
M$39:	***ok*<FB;W;Gi;A> *tack så mycket* <Ph;S-Nod;Pol>**
S$40:	***tack* <W;S-Nod;Pol>**

Example 5 from PF-Star-Dial 3.

8.3.4 Conclusions and Discussion

The results of this study aiming at investigating the communicative function of head nods in Swedish dialogic speech shows that in 70% of cases the function of head nods is related to FEEDBACK. Besides FEEDBACK head nods are produced to signal FOCUS and EMPHASIS, POLITENESS, TURN YIELD and to give POSITIVE ANSWERS.

Very few instances of head nods produced as TURN-MANAGEMENT signals have been identified in the analysed materials. This might depend on the fact that TURN MANAGEMENT is signalled by means of other FACIAL

DISPLAYS and gestures, as for instance gazing [Kendon 1967] and hand gestures [Knapp & Hall 2002].

The hypothesis that short, minimal head nods might be related to short verbal feedback expression carrying out the function FEEDBACK GIVE CONTINUATION YOU GO ON, seems to be supported mainly in the case of subject-S. In fact, if for subject-S it is true that minimal head nods are mainly produced to accompany short verbal expressions having the function of CONTINUATION YOU GO ON, it is also true that for subject-M, minimal single head nods are produced also when the function of feedback is GIVE ACCEPTANCE**.** This result is consistent with the results shown in chapter 6 and with the trend observed in the realization of head nods in the materials analysed in study 1 in this chapter. These results clearly point out that NOD and JERK are typical movements involved during the production of non-intrusive FEEDBACK.

One of the most clear and interesting results of this investigation is the different use of SINGLE and REPEATED NOD: SINGLE NOD tends to be produced to show CONTINUATION, very seldom is REPEATED NOD produced with this function, rather REPEATED NOD is more frequently produced to GIVE ACCEPTANCE and to REQUIRE ACCEPTANCE.

The results of the duration analysis show that SINGLE NOD can have different durations depending on the different communicative functions they serve. If these differences in the duration will be proven to be significant on a bigger amount of data, they could represent a distinctive cue for the different communicative functions that head nods can carry out. This cue could be then exploited in the implementation of communicative head nods in talking heads used in human machine interfaces.

The implementation of communicative head nods in talking heads requires a large database of annotated head nods for the training and testing of the models in the data-driven process.

Since the manual annotation of a large database is a time consuming and even subjective task, one possibility to facilitate the annotation is to perform an automatic detection of head nods on the acquired data with the optoelectronic system. The third study presented in this chapter proposes therefore, a method for the automatic detection of head nods.

8.4 Study 3-Automatic Detection of Head Nods

The aim of this study[29] is to propose a method for automatic detection of head movements, in particular head nods.

The automatic detection of head nods is a reliable means to retrieve an adequate number of items to train and test head movement models for implementation in talking heads.

[29] This study was carried out with Gunilla Svanfeldt [Cerrato & Svanfeldt 2005].

The material used for this study consists of short sentences as well as of dialogic speech produced by a Swedish actor who was recorded by means of an optical motion capture system.

The method for automatic head nods detection is based on criteria for slope, amplitude and a minimum number of consecutive frames. The criteria are tuned on head nods that have been manually annotated. These parameters can be varied to detect different kinds of head movements and can also be combined with further parameters in order to detect facial gestures, such as eyebrow displacements.

For this study the focus was on the detection of head nods, since in earlier studies they have been found to be important visual cues in particular for signalling feedback.

8.4.1 Materials

The material used for this study has also been selected from the Swedish Multi-modal PF-Star Corpus 3 (see section 3.3.4). Two sets of data have been used for this study:

- 39 short sentences[30] uttered with a confirming expression and a controlled variable position of the focus. These sentences were used to train the head nods detector;
- 10 short dialogues of the length of 1 minute each, set in a travel agency scenario. These dialogues were used to evaluate the head nods detector. (These are the same ten dialogues used to carry out study 2).

During the annotation of the PF-Star Corpus 2 [Beskow et al. 2004b], which included short sentences such as: *grannen knackade på dörren*, *damen vattnade blommorna, båten seglade forbi* (the neighbour knocked on the door, the lady watered the flowers, the boat sailed by), providing good phonetic coverage and uttered with different expressions of emotions, consisting of: *confident, confirming, questioning, insecure, happy,* and *angry*, plus *neutral*, it was noticed that the actor produced evident head movements when he uttered the sentences with a *confirming* expression.

In PF-Star Corpus 3 (see section 3.3.4), which included the same short sentences as in PF-Star Corpus 2, uttered with the same expressions of emotions and in addition with a varying position of the focus, it was noticed that for the sentences uttered with a *confirming* expression, the production of head movements was connected with the prosodic structure of the text. Figure 8.9 illustrates this phenomenon: the displacement of the marker on

[30] In PF-Star corpus 3, 15 sentences with a variable position of the focus and different expressions of emotion were recorded. The sentences uttered with a confirming expression were originally 15 and they were uttered with a variable position of the focus (3 different positions per each sentence) thus yielding 45 sentences. Of these 45 sentences 6 could not be used, for this reason the number of sentences used for the training of the method is 39.

the nose-tip (y-direction) is here plotted during the production of the sentence *damen vattnade blommorna* (The lady watered the plants). The sentence was uttered thrice, with a different position of the focus:

- the upper plot shows the marker displacement for the sentence with focus on DAMEN,
- the middle plot shows the marker displacement for the sentence with focus on VATTNADE,
- the lower plot shows the marker displacement for the sentence with focus on BLOMMÅRNA.

Because the sentences uttered with a variable position of the focus and with a *confirming* expression were produced with head nods on the words that received the focus, they were selected from PF-Star corpus 3 for the tuning of the criteria for the parameters in the head-nod detection method.

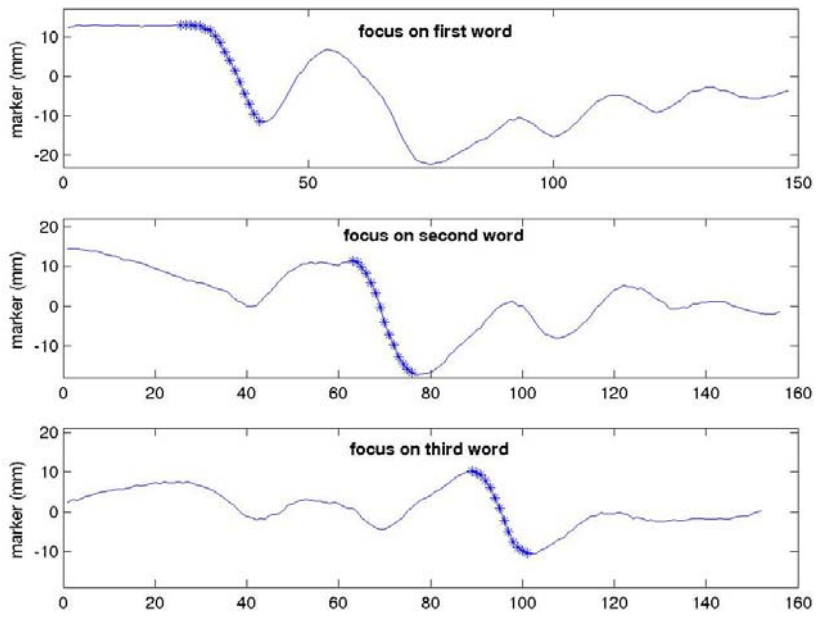

Figure 8.9 Displacement of the marker on the nose-tip (y-axis) per time (x-axis, in frames) during the production of the sentence "damen vattnade blommorna" with varying position of the focus.

8.4.2 Method

The 39 sentences were manually annotated by two annotators who independently from each other identified the head nods and marked, in each sentence, the most prominent of the head nods. The annotations showed a good level of inter-agreement. However, since there exists no gold standard (the total number of presumptive nods is not known); the conventional calculation of inter-agreement was not applicable. Therefore it can be

simply reported that both annotators found 139 nods each, of which 131 were identified by both annotators, which means 94% of agreement.

In the ten dialogues, the annotation of the head movements was performed by only one annotator who identified a total of 93 head nods in the production of the subject with the marker on his face --subject-S--. The annotator assigned a communicative function to each head nod. Several functions categories were defined *a priori*, based on previous observation of head nods and on literature references. Functions include: FEEDBACK, TURN MANAGEMENT, POLITENESS, POSITIVE ANSWERS FOCUS and EMPHASIS. However 70% of the identified head nods were assigned a FEEDBACK category.

The annotation of head nods in the audio-visual material was carried out with the help of WaveSurfer provided with a video plug-in, that allows to see the video files in .mpeg format together with the 3D data.

A very simple approach was chosen for head nods detection. The marker on the nose was used as reference for the detection of nods. It was thus assumed that the movements of this marker are representative for the movements of the head. Of course, other movements as well, such as changes in body posture, can cause the same displacement of the marker as a nod might do. This potential source of error could easily be avoided by looking at the rotation angle rather than point displacements.

The chosen parameters were: slope, length (in frames) and (vertical) amplitude. The slope finds the candidates at first and then the other criteria are applied. Only the negative slopes were considered, since this is the most prominent feature of a nod. Therefore, the temporal length of a nod corresponds to approximately half the total length of the total cycle.

In order to find suitable criteria for the parameters, only one parameter at a time was varied, and a record was kept over the "under/over" hit rate of the automatic process as compared to the manual marking of head nods done by human annotators.

Then the criteria that gave the best results were chosen. Since it was earlier noticed that in the sentences uttered with the confirming expression subject-S produced lots of head nods, these sentences were chosen as training material. The minimum criteria values for the automatic detector that were reached were:

- minimum number of frames: 7 (=116ms)
- minimum amplitude: 4 mm
- minimum slope: -0.3 mm/frames (=-5mm/ms)

In 39 sentences with the confirming expression in total 139 nods were identified by the two annotators, although some disagreement existed. Compared to one of the annotators, the automatic process missed 4 and found 3 additional nods. Compared to the other annotator, 3 nods were missed out, and 2 additional were found by the automatic process.

The vertical displacement of the marker on the nose tip during the sentence *Båten seglade förbi* (The boat sailed by) is showed in figure 8.10. Head nods are displayed as vertical location of the nose tip as a function of time. The vertical movements are shown on the y-axis, and time is displayed on the x-axis. The upper (blue) line is the actual location, where the bold line denotes the automatically found nods. The two lower thick lines indicate where the annotators have found nods. The lines beneath are the same trajectory, only displaced 5 (resp 10) mm for clarity reasons. The (red) bold lines on those correspond to the annotators marking of nods. This way the result can easily be compared.

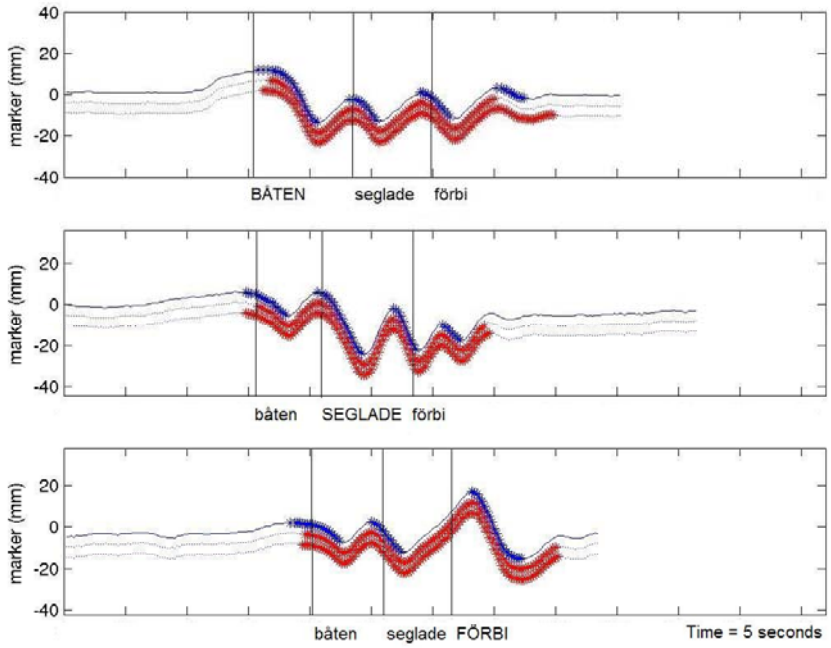

Figure 8.10 Displacement of the marker on the nose tip during the sentence "Båten seglade förbi" (The boat sailed by).

8.4.3 Evaluation

In the ten dialogues 93 head nods had been manually identified and their specific feedback function labelled in the given context. The automatic detection was carried out on the ten dialogues and then the mismatches in the two annotations (manual vs. automatic) were analysed.

The evaluation of the method on the dialogues showed the complexity of spontaneous speech in comparison to the controlled sentences, on which the criteria for detection were based. In the dialogues the head movements were not as smooth and cyclic as in the sentences that were used for training. However, 95 % of all manually annotated nods were found by the automatic process.

Figure 8.11 illustrates the detection of head nods in dialogic speech. The vertical displacement of the marker on the nose tip is shown on the y-axis, and the time is displayed in frames on the x-axis (each frame corresponds to 17 msec, so 300 frames correspond to ca. 5 seconds). The three panels show a sequence of the duration of 15 seconds of one of the recorded dialogues. In each panel the upper line is the actual location of the marker for each frame, and the thicker (blue) line denotes where the automatic detector has found a nod.

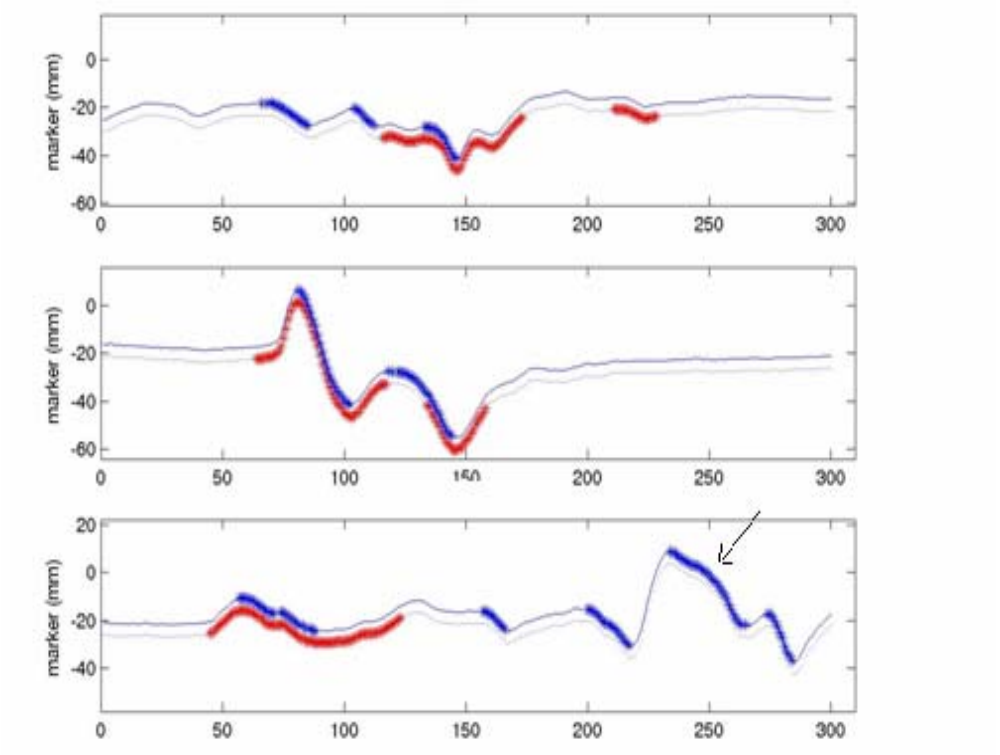

Figure 8.11 Displacement of the marker on the nose-tip in mm (y- axis) per time (x-axis, in frames, 300 frames = 5 seconds) during the production of dialogic speech with complex head movements.

The lower (red) bold lines correspond to the annotator's marking of nods. While in the example of annotation shown in figure 8.10 there were two (red) bold lines (corresponding to manual identification of nods made by the two annotators) in figure 8.11 the red line is only one since only one annotator manually identified nods in the dialogues.

Several intervals were identified as nods by the automatic detection, although they were not marked as such by the manual annotation, as for instance the last displacement marked by the arrow in figure 8.11.

One reason for the "over hit rate" of the detector might be that the criteria were too lax due to the differences between the production of the prompted utterances and the spontaneous dialogues.

In the prompted utterances used for the training of the detector the head movements seem to be more regular, while in the dialogic speech they show more complex patterns. What the automatic detector identifies as a nod, given the fact that only the vertical movement of the head is considered, might not have been manually annotated as such in dialogic speech since additional sideway movements might have been present. In other words what to the human eye does not look like a nod, might be detected as such by the automatic recognition process.

In order to check whether there might be any truth is this explanation, another criterion was added to the detector. The criterion concerned maximum sideway displacement, which means that in the second run of the automatic detector the sideways movements were restricted.

This sideway displacement restriction makes the number of "over hits" decrease by almost 48%. However, it also causes the loss of some of the previous correctly annotated nods (22%).

Figure 8.12 shows the same three segments of one of the dialogues as shown in figure 8.11, but this time with the additional criterion for the restriction of sideways movements. With the additional criterion for the restriction of sideways movement the last displacement marked by the arrow in 8.12 is not identified as a head nod.

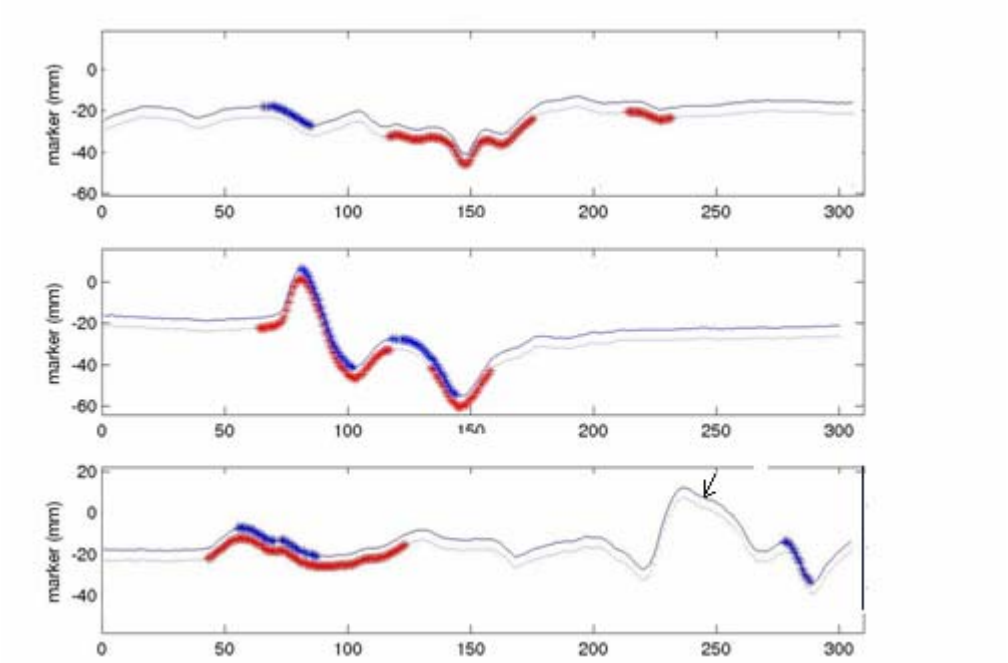

Figure 8.12 Displacement of the marker on the nose-tip in mm (y-direction) per time (x-direction, in frames, 300 frames = 5 seconds) during the production of dialogic speech: same three segments as in figure 8.11 with an additional criterion in the x-direction, for the restriction of sideways movements.

8.4.4 Conclusions and Discussion

The results of the evaluation of the automatic detector of head nods showed the complexity of spontaneous speech in comparison to the controlled sentences, on which the criteria for detection were based.

The automatic detector is able to identify head nods in general, but it is not able to recognize different types of head nods that serve different communicative functions. This might depend on the fact that there is not a one-to-one relationship between a movement and its function, but it might as well be due to the fact that only criteria for minimum values were used, and maybe in order to separate different nod functions, it is likely that also maximum criteria should be used. Also, the parameters were treated one by one, and perhaps more precise results would be achieved if the combinations were to be considered and the criteria set according to that.

One feature that would need to be added is repetitiveness, since both annotators seems to have fewer constraints on, for example, amplitude when the nod is in a series of nods.

There is also the question of how much sideways movement that is to be allowed for a head movement to be considered a head nod. When sideway movements are restricted, some of the movements that had been identified

as head nods by the annotators are not "let through" the automatic process. This means that vertical and sideways displacement on their own are not enough for the accurate automatic detection of head nods. Probably some other criteria need to be taken into account, since the distinguishing factor is not in the movement itself, but probably in the linguistic-conversational context. Since many head movements are quite subtle, it must be the use of contextual information that helps people to recognize them in communicative situations. For instance at the end of a sentence, or of a contribution, the speaker might expect the listener to produce a head nod to signal feedback.

Incorporating contextual information has been proven successful in visual feedback recognition [Morency & Darrel 2006], so it might be possible that even our automatic system for the detection of head movements might benefit from additional contextual information.

Even so this method seems to be a valuable help to detect head nods, as well as other facial displays, in large multi-modal corpora collected by means of motion capture, where the manual annotation would be too time-consuming and costly.

The automatic detection of head nods, as well as of other communicative non-verbal behaviour is a good method to obtain a large amount of data for the training and testing of a data-driven model of non-verbal behaviour in talking heads.

8.5 General Conclusion

In this chapter the acquisition of 3D data recorded by means of the opto-electronic system Mac Reflex Qualisys has been described and some examples of the analysis of the trajectories of head movements related to FEEDBACK have been presented. Moreover a method for automatic detection of head movements, in particular head nods signalling FEEDBACK, has been developed.

The automatic detection of head nods is suggested as a possible means to retrieve an adequate number of items to train and test head movement models for implementation in talking heads.

In the following chapter the possibility to use the results of the analysis of 3D data to train and test models for the implementation of head movements and facial displays in talking heads is discussed.

9 Implementation and Evaluation

9.1 Introduction

One of the final aims of the studies reported in this thesis is to provide data that could be used to control facial displays related to non-verbal communicative behaviour --in particular feedback-- in synthetic conversational agents. The assumption was that the implementation of non-verbal communicative behaviour in talking heads could improve their effectiveness in interactions with human users and could make users experience talking heads as more natural and pleasant.

The *modus operandi* followed in this thesis starts with the investigation of these phenomena in spontaneous human-human interaction, in order to find regularities in human communicative behaviour that could be reproduced in talking heads, and ends up with more detailed analyses of data collected in a lab environment, with an acquisition set-up that allows capturing of the dynamics of facial displays for the sake of reproduction in talking heads.

The big challenge lies therefore in trying to reproduce non-verbal behaviour in talking heads and in evaluating their appropriateness and effectiveness in a real usage context. Since human users are very sensitive and critical concerning non-verbal behaviour the agents must act naturally in order to be believable.

In order to implement natural, effective and believable, head movements that signal feedback functions in talking heads, a greater amount of instances of head movements than what was available after the analysis described in chapter 8 is necessary. For this reason, the goal of implementing head nods related to feedback was not pursuable.

However, the collection of data performed by means of the opto-electronic system under the framework of the European project PF-Star (illustrated in chapter 3, section 3.3.4) also included a controlled set of utterances acted with the six basic emotions. This controlled set of data related to the visual expression of emotions allowed for the reproduction of expressive visual articulation in a virtual talking head.

The expressive articulation was assessed by means of an evaluation test in which external observers had to judge the emotional expression shown by the talking head, which was inserted in an interactive scenario.

By inserting the talking head in a simple scenario where it carried out effective interactions with a user in a given realistic usage context, it was possible to employ the expressive articulation as a non-verbal means of showing expressive feedback.

Before illustrating the evaluation test design and results, this chapter describes and discusses other evaluation paradigms which have been employed to assess the reproduction of non-verbal behaviour in animated talking agents.

9.2 Implementation, why and how

Technological and scientific development has made it possible to faithfully synthesize and animate human faces and heads, and to synchronise their articulation to speech. Examples are available in many languages such as: English [Massaro 1998] Swedish [Beskow 2003] and Italian [Cosi, Fusaro & Tisato 2003].

Visual speech information has been proven to increase speech intelligibility, especially under acoustically degraded conditions [Sumby & Pollack 1954; Summerfield 1979; Massaro 1998; Beskow, Granström & Spens 2002]. In the same way it is assumed that when speech is presented together with communicative non-verbal behaviour, it may result in a more robust, more natural and more efficient communication.

Existing implementations of communicative (non-verbal) signals in talking heads or embodied conversational agents are often based on prototypical descriptions of human-human communication found in psychology literature or on observations conducted in a non-systematic way, which means that reproduction of the behaviour in talking heads is based either on formalized findings or on intuition rather than on empirical data.

These implementations are unable to display the degree of variability and dynamics exhibited in human facial expressions in real communicative situations. Rather, they risk being characterized as stereotypical and predictable. To compensate for this, often a certain degree of manual tuning is involved for the application. This makes it possible to achieve effective displays for limited scenarios, even if augmenting and improving the agent's communicative repertoire requires substantial manual labour.

In order to emulate the degree of variability found in human-human facial expressions it is therefore suggested to acquire dynamic data with motion capture systems, which allows controlling facial displays in synthetic talking heads. The reproduction of facial displays can be performed by two different methods: one consists in re-using the registered dynamic sequences of natural recorded behaviour --this process is referred to as re-synthesis-- the other consists in generalizing them, referred to as data-driven process.

Re-synthesis is the simplest process; it exploits the 3D point trajectories recorded by the motion capture system, which are previously converted into MPEG-4 FAPs. The FAPs, facial animation parameters, specify the movements of a number of feature points in the face, and are normalized with respect to face dimensions, to be independent of the specific face model. Thus it is possible to drive the face from points measured on a face that differs in geometry with respect to the model.

The outcome of re-synthesis is a very realistic reproduction of the registered facial displays, however this method does not allow for the possibility of reproducing non-verbal expressions in co-occurrence with arbitrary speech text. This is possible if a data-driven strategy is employed for controlling communicative movements.

A data-driven process can generalise from dynamic sequences of natural recorded behaviour and thereby better capture and model the variability that is present in human expression [Beskow and Nordenberg 2005]. A data-driven technique has so far successfully been applied to control the articulatory movements of expressive speech in a Swedish talking head [Beskow & Nordenberg 2005; Beskow & Cerrato 2007].

An articulatory control model based on Cohen & Massaro [1993] has been trained on the articulatory parameter trajectories recorded from a Swedish actor, so as to learn to predict the patterns and produce articulatory movements for novel (arbitrary) Swedish speech. In order to account for emotional expression, seven separate articulatory control models were trained: one for each of the six basic emotions [according to Ekman 1982] and one for neutral speech. To train these models, short sentences uttered with different expression of emotions were recorded using the opto-electronic system Qualisys. The 3D point trajectories recorded by the Qualisys system were first converted into MPEG-4 FAPs. Then a Principal Component Analysis (PCA) was carried out on the FAP-data for each emotion. The top 10 principal components were able to explain 99% of the variation in the original FAP data streams. Each of the top 10 PCs were modelled individually using the Cohen-Massaro model of coarticulation [Cohen & Massaro 1993].

By analogy with this data-driven strategy applied to obtain expressive speech, it seems likely to foresee that models for each head movement that is considered to have an important communicative function could be trained.

9.3 Evaluation, why and how

Collecting data and implementing human communicative behaviour is however not the whole story. If it is not possible to prove that a certain feature can improve some aspect of the communicative experience, it becomes difficult to motivate its implementation. For this reason it is recommendable to perform evaluation tests.

Talking heads and/or embodied conversational agents (ECAs) can be evaluated at the micro and macro level. Evaluation at the micro-level consists in testing whether the designer's model, as implemented in a talking head, is understood by subjects in the intended way.

This kind of evaluation is usually carried out by means of judgment studies, in which human subjects are asked to judge a model that has been implemented in a talking head or in an embodied conversational agent (ECA), which is not placed in any given realistic usage context. (See for instance Massaro et al. [2001], Beskow [2003] for micro-evaluation of an articulatory model; Krahmer et al. [2003] for evaluation of a model of ECA's personality; Fabri, Moore & Hobbs [2002] for evaluation of emotional expressions).

The recommended evaluation paradigm for evaluation at the micro-level is to compare ECA with NO ECA (for a baseline) and with a HUMAN (for the golden standard). By macro-level evaluation is meant the evaluation of a functional talking head or embodied agent capable of interaction in a given realistic usage context, as for instance in a dialogue system. Macro-evaluation is also referred to as system-level evaluation. This kind of evaluation considers different aspects of the interaction and of the ECA itself. Evaluation paradigms at the macro-level can be designed to assess how fluent and successful the interaction with an embodied agent is, the user's experience of the interaction, the benefit of the presence of the embodied agent for the sake of the interaction, and the most appropriate characteristics of the embodied agent for a given context.

Because interactive dialogue systems with ECAs are still very much at the prototypical stage, little has been done to evaluate the different aspects of ECAs at the macro-level. It is probably still too early to conduct comprehensive, definitive empirical studies that might cover all the aspects of ECAs evaluation since the research field is still very young and no standard evaluation methodology have been developed yet, even if attempts have been made to provide some general directions [Sanders & Scholtz 2000; Noor 2004; Ruttkay, Dormann & Noot 2004].

9.3.1 Micro-Level Evaluation

An example of evaluation at the micro-level is offered by an experiment run to evaluate the expressiveness of a talking head within the EU-project PF-Star[31].

The design principles for the experiment are inspired by Ahlberg, Pandzic and You's [2002] evaluation procedure for MPEG-4 facial

[31] A whole work package of the project was dedicated to synthesis of facial expressions of emotions, and much effort was spent to design a methodology for evaluation of the emotional facial displays performed by 3D animated talking heads.

animation players. They propose to measure the expressiveness of a synthetic face through the accuracy rate of human observers who recognize the facial expression, and to compare the expressions of the synthetic face with those of the "original" human face, upon which they are based.

The original aim of the experiment, which consisted of a cross-cultural evaluation of expressiveness in synthetic faces [Beskow et al. 2004a], was to assess the adequacy of the emotional facial displays performed by Italian and Swedish talking heads. In this discussion only the data relative to Swedish will be considered, with the aim of discussing issues related to the evaluation of expressiveness of a synthetic face compared to the expressiveness of a natural face.

The main assumption on which the experiment was based is that an accurate analysis and understanding of the way in which humans use facial expressions can provide valuable insight into how to control the synthetic ECA's facial expressions, hopefully leading to the implementation of natural-looking expressions.

Preparation of data involved: recording an actor uttering a series of stimuli acted with three emotions using the opto-electronic system Qualisys[32], production of the related MPEG-4 FAPs (Facial Animation Parameters) files, and animation of the FAPs sequences using the synthetic face. The materials used for the test consist of ten non-sense words uttered three times each with two emotions: angry and happy, plus neutral.

The synthetic 3D face model used in the study is made up of approximately 1,500 polygons, and adheres to the MPEG-4 Facial Animation (FA) standard. The FAPs are normalized according to the MPEG-4 FA standard, so that they are speaker-independent. The point trajectories obtained from the motion tracking systems described above were converted into FAP streams with custom made software. The FAP streams were then used to animate the synthetic faces.

One group of Swedish university students (30 volunteers from the Dept. of Linguistics of the University of Stockholm and the Dept. of Speech, Music and Hearing at KTH) were confronted with 24 video-files: 12 showing the Swedish actor, and 12 showing the talking head uttering the stimuli ABBA and ADDA, produced with two emotional expressions (*happy* and *angry*) plus *neutral.* The stimuli were played in random order, without the audio.

After each presented video-file, the participants were asked to choose, on the answer sheet, among the three available labels for the emotional expressions. At the end of the experimental session, they were also asked to fill in a short questionnaire about their impressions concerning the faces.

[32] The data used to prepare the stimuli was selected from the PF-Star Corpus 2, presented in chapter 3.

The average percentages of correct recognition in table 9.1 show that even if the human face got higher rates than the synthetic face, the recognition rates for the synthetic face are quite good.

Table 9.1 Average percentages of correct recognition for each emotion.

Emotion	**SW actor**	**SW synthetic face**
Angry	81%	66%
Happy	88%	77%
Neutral	91%	79%
All	87%	74%

This result might be interpreted in a positive way as an indication that the models for expressive speech are quite effective, and the process of re-synthesising the human behaviour from the observed reality gave a faithful reproduction.

However this evaluation paradigm presents a series of limitations. The main limitation consists in the fact that both the recorded expressions of emotions and their reproduction in the talking heads were de-contextualised. The material consisted of a restricted number of short non-sense words uttered in isolation by an actor with different emotional expressions. The actor was sitting in a silent booth with 35 markers applied on this face and with five video cameras recording him. It could be argued that this kind of speech is not very representative of human communicative behaviour, because nobody in everyday life usually sits in a silent booth with markers glued on the face and utters non-sense words in isolation, with different expressions of emotions!

The production of such short non-sense words with different emotions out of the context cannot be an optimal way of mediating emotions. Context plays in fact a crucial role both in emotion expressions and recognition. Effective accurate mediation of emotion is closely linked with the situation and other related communicative signals, therefore a reliable interpretation of facial expressions cannot work independently of the context in which they are displayed.

To overcome this limitation it is therefore advisable to record materials consisting of more spontaneous interactions between two participants, which might be a better source for instances of natural expressions of emotions and communicative non-verbal behaviour. This is what has been done under the framework of the PF-Star project with the collection of the PF-Star Corpus 3 (see section 3.3.4). In the ten semi-spontaneous dialogues recorded with the motion capture system, several non-verbal-behaviour related to the expression of feedback, turn management, and emotional attitude were identified. These data represent a quite valuable source for implementation of non-verbal behaviour in talking heads which could be evaluated in more realistic scenarios.

However even in these data, the limitation of the markers glued on the face and the video cameras will still persist. Unfortunately there is a problem here and this is an old unsolved problem, which Labov defined as the “observer paradox”: how to record the way people speak when they are not being observed [Labov 1972, p. 181]. Overcoming the observer paradox is impracticable.

9.3.2 Macro-Level Evaluation

Macro-level is a synonym of “system-level”, which means that the effectiveness, appropriateness and usability of talking heads are evaluated in the context of a functional interactive system. Many factors intervene at the macro-level, and it becomes very difficult to control all the variables and to interpret the results of the evaluation studies in the correct way. However, because interactive dialogue systems with talking heads are still very much at the prototypical stage, it is still too early to conduct comprehensive, definitive empirical studies that might cover all the aspects of evaluation.

Usually evaluation at the macro-level is carried out by asking users to answer questions about how they experienced the interaction with the system and whether they benefit from the presence of the talking head. The assumption here is that by evaluating users’ satisfaction in the interaction with the system, it is possible to get subjective measures that can be used for the evaluation of the overall performance of the system [Walker, Kamm & Litman 2000].

Using subjective measures for the overall evaluation of a system can be misleading, since the judgments of the users can be influenced by many factors and moreover there are particular questions, which are difficult to formulate, such as whether the interface has influenced the users' feelings and expectations during the interaction.

Furthermore there are the issues of costs, time and users' integrity. Even if the individual point of view of the users is very important for the aim of evaluation, it is often difficult to collect individual judgments and combine them with other component level-based metrics in order to formulate generalizations in the final evaluation.

A way of avoiding the complexity of overall system evaluation but keeping at higher level of evaluation is to simulate an effective interaction between a talking head and a user in a realistic scenario. This can be done by using the Wizard of Oz technique.

In WOZ experiments a user interacts with what appears to be a computer system, but is in fact a simulation provided by either a human (referred to as the wizard) or the combination of a human and a computer. This technique is simple and flexible. For the evaluation of talking heads it has been used to find out how users are likely to interact with a system endowed with a talking head, or to evaluate different versions of the talking head showing different implemented characteristics [Edlund & Nordstrand 2002].

With the WOZ technique it could be possible to assess the feasibility and appropriateness of different characteristics of the talking head in a given context. For instance it could be assessed whether the interactions could benefit from the presence of a given non-verbal behaviour implemented in the talking head.

Evaluation of interactions with dialogue systems endowed with talking heads is quite hard for many reasons, mainly because interactions are difficult to record under normalised and easily reproducible conditions, and because they are highly dependent on user behaviour.

One way to evaluate talking heads characteristics in a realistic scenario and in an interactive situation with a user, but at the same time avoiding the dependency from user's behaviour, is to simulate an effective interaction between a talking head and a user who always behaves in the same way and then present these simulated interactions to external observers who are asked to judge whether the talking head characteristics were appropriate/effective in that given scenario.

An example of this method is given by a couple of experiments carried out at TMH-KTH. The first experiment aimed at gaining more insight into the relative importance of specific prosodic and visual parameters for giving feedback implemented in a talking head [Granström, House & Swerts 2002]. The experiment was conducted by using a synthetic talking head whose prosodic and visual features were orthogonally varied in order to create stimuli that were presented to subjects who judged them as affirmative or negative feedback signals. The talking head was inserted in an interactive situation in a simple scenario, in which it had the role of a travel agent interacting with a human voice, which was intended to represent a client/user. The subjects could only hear the voice of the user/client, which was a natural recording of a male voice, exactly the same in every stimulus, and see and hear the talking face. The talking face contribution was manipulated by orthogonally varying 6 parameters (smile, head movements, eyebrows, eye closure, F0 contour, delay). Two possible settings were used for each parameter, one which was hypothesized to lead to affirmative feedback responses and one which was hypothesized to lead to negative responses.

The perceptual test was run in the following way: subjects were asked to look and listen to the stimuli exchange and judge whether the talking head was signalling understanding, acceptance of the utterance produced by the human client (i.e. positive feedback) or uncertainty about it (negative feedback).

The results showed that the parameters which had the most influence on subjects' judgements were, in rank order: smile, pitch contour, eyebrows and head movements. The conclusion of this study was that subjects are sensitive to both acoustic and visual parameters when they have to judge utterances as affirmative or negative.

The set-up of this experiment is quite interesting, since it offers the possibility to directly control a number of parameters and to evaluate their role in the perception of the talking head in an interactive context (interaction between a human being and an "agent"); however the method shows some limitations. The first limitation concerns the "observation of the reality". As the authors state "parameter settings were largely created by intuition observing human productions" and this is more an interpretation of the reality than a reproduction of it. Another limitation could be seen in the number of parameters which were considered in the experiment: only six, probably the most "evident" from the researchers' point of view, but we are not sure that they are the most important/effective. However in the experiment it is possible to manipulate only a limited number of parameters at a time and this rules out the observation of other parameters. But on the other hand, augmenting the parameters to check can lead to a potential enormity of combination of parameters, which might result in unfeasibility.

Avoiding this limitation, though incurring other kinds of limitation, could be done by carrying out a direct observation of human communicative behaviour (i.e. based on empirical data). This is what was done in another evaluation session, designed to evaluate the implemented models of visual expression of emotions based on the data recorded with the motion capture system Qualisys.

9.4 Evaluation of the Expressiveness of a Swedish Talking Head

An evaluation test was designed with the aim of assessing the expressive visual articulation of a newly developed talking head. The synthesis, which is an initial attempt to synthesize expressive visual articulation using an MPEG-4 based virtual talking head [Beskow & Cerrato 2004], is data-driven[33], trained on a corpus consisting of 75 short sentences uttered with the six Ekman basic emotions (*happiness, sadness, surprise, disgust, fear, anger* plus *neutral*). The corpus was collected with the Qualisys system. These sentences are part of the PF-Star Corpus 3, illustrated in section 3.3.4

Each emotion was modelled separately using principal component analysis and a parametric coarticulation model (For more details on the synthesis see [Beskow & Nordenberg 2005; Beskow & Cerrato 2007]).

In order to evaluate the expressiveness of the data-driven synthesis, an experiment was designed and performed. Our talking head was inserted in an interactive situation in a simple scenario where it carried out effective interactions with a user in a given realistic usage context.

[33] The data-driven synthesis was implemented by Jonas Beskow, with whom this evaluation study was planned and conducted [Beskow & Cerrato 2007].

The interactions were then presented to external observers who were asked to judge the emotion expressed by the talking head.

9.4.1 Test Design

To quantify the perceived expressiveness of the talking head, a perceptual experiment consisting of two sub-tasks was designed and conducted.
Task 1 aimed at:

- evaluating the expressiveness of the talking head on the basis of the accuracy rate of human observers recognizing the emotional facial expressions in a given context,
- assessing the strength of the visual modality on the recognition of emotion,
- determining the appropriateness of a given expressive feedback in a given context and checking whether the context would bias the responses of the participants.

Task 2 aimed at:

- Determining whether the observers preferred the expressive talking head or the non-expressive one.

The talking head was inserted in an interactive situation in a simple scenario, where it had the role of a language tutor, interacting with a human user, who was intended to represent a student of Swedish trying to learn to pronounce the word *Linköping* (the name of a Swedish city). For task 1, three contexts were designed. Each context was thought to lead to an expected emotional reaction on the side of the tutor that could be judged by external observers.

This learning environment seemed to be the most appropriate for our purposes, since it allows the possibility to evaluate different expressions of emotions that can be interpreted as feedback expressions in the given context. In a learning environment, in fact, the ability to show emotion through facial expressions is central to ensure the quality of the tutor-learner interaction [Cooper, Brna & Martins 2000]. The emotions expressed by the talking head/tutor in the experiment were: *happiness, sadness* and *anger* plus *neutral*.

The voice of the tutor was a pre-recorded male voice, with acted emotional expressions. The voice of the user was a pre-recorded female voice, with acted pronunciation mistakes.

The tutor's voice was phonetically labelled using forced alignment [Sjölander & Heldner 2004]. Talking head animations were then synthesized using the different expressing speech models. The pre-recorded voices were automatically combined into video files forming the stimuli for the experiment.

The interactions were presented to external observers who were asked to judge the expressiveness of the talking head/tutor. The participants in the

experiment could see and hear the talking head while the user could only be heard.

The talking head/tutor is shown in figure 9.1.

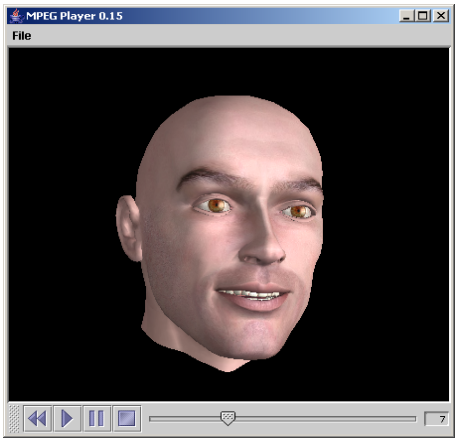

Figure 9.1 The talking head/tutor.

All the interactions started in the same way: with the learner saying a sentence containing the mispronounced word *Linköping*. The reaction of the talking head/tutor to the pronunciation mistake consisted in the production of the word *Linköping* with the correct pronunciation, to which the learner reacts by repeating the word either with the correct pronunciation or with the same mispronunciation as before or with a new pronunciation mistake.

Each of these three reactions is thought to lead to an expected emotional feedback on the side of the talking head/tutor, as schematized in table 9.2.

Table 9.2 The three contexts used in task 1.

Context	**Learner's pronunciation**	**Expected tutor's feedback**
1	Correct	Happy
2	Same mispronunciation	Angry
3	New mispronunciation	Sad

Besides the stimuli with the "appropriate" emotional feedback (*happy*, *angry* and *sad*), stimuli with all the possible combinations of auditory and visual stimuli for the three contexts were created; this way obtaining a total of 48 stimuli (4 emotions, 2 modalities, 3 contexts, 2 conditions = 48 stimuli).

In other words the three emotions plus neutral were presented with consistent auditory and visual stimuli, for instance a sad visual expression

with a sad voice and also with a mismatch between the auditory and visual stimuli, for instance a sad visual expression with a happy voice.

The presentation of all these conditions provides an informative picture of how the two speech modalities are processed, and this way it is possible to assess which modality (auditory or visual) plays a more relevant role in the recognition of emotion.

Ten Swedish native speakers (five male and five female around 25-30 years) volunteered to participate in the experiment. They were instructed to watch and listen to the talking head/tutor interacting with the human/learner and judge which emotion the talking head was expressing in the last turn of each interaction. They were instructed to mark their answer on a separate answering sheet, choosing among the four possible answers: *angry, neutral, happy,* and *sad*. The test-session lasted about 15 minutes.

9.4.2 Test Results

The total number of responses for each emotion, for the three different situations (happy, angry and sad reply) is reported in figure 9.2. These results show that the situation where the tutor had an angry reaction is the only one that biases the responses of the participants in the test, however angry is generally overrepresented.

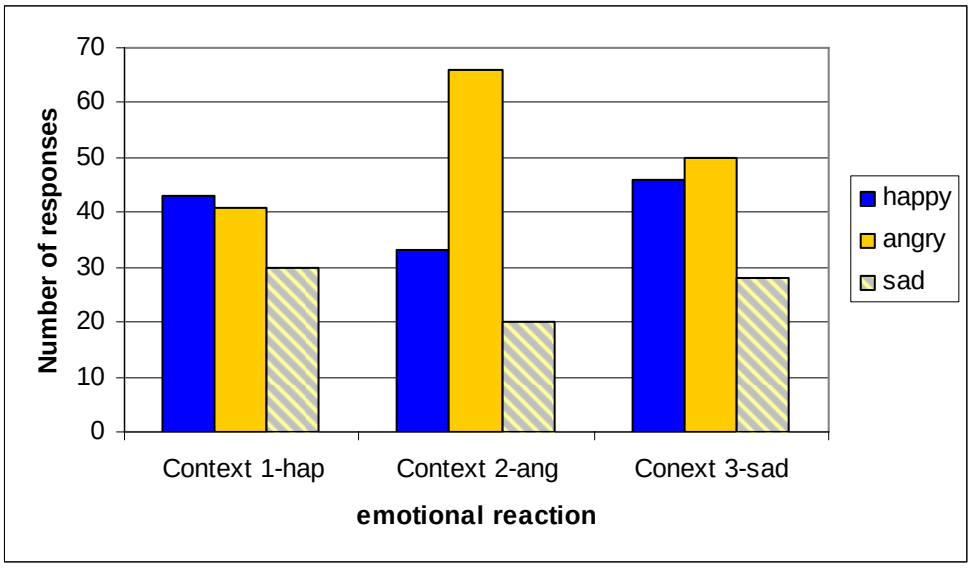

Figure 9.2 Total number of responses for each emotional reaction for the three different contexts.

The results, collapsed for the three situations, are reported in the confusion matrix in table 9.3. The matrix shows the distribution of the responses (neu = neutral, hap = happy, sad = sad, ang = angry) for all the stimuli combinations (auditory + visual). The stimuli combinations include

both consistent and inconsistent stimuli. “v” before the emotion stands for visual, “a” for auditory. The results show that for consistent stimuli, emotion recognition is quite good: 87% for neutral and happy, 70% for sad and 93% for angry (the figures in red).

For inconsistent stimuli the visual stimuli have a stronger influence than the auditory ones on the recognition of the expressive reaction. The only exception is for the stimuli a-ang/v-sad, and a-ang/v-neu, where the auditory stimuli (i.e. a-ang) seem to have a stronger role on the recognition.

For those stimuli consisting of the combination of the auditory neutral expression with the visual emotional expressions the responses of the subjects are clearly influenced by the visual stimuli.

Table 9.3 Confusion matrix for the recognition test.

Stimulus		**Responses**			
Auditory	**Visual**	**neu**	**hap**	**sad**	**ang**
aneu	**Vneu**	**87**			
	Vhap	13	**73**	7	7
	Vsad	23	0	**43**	33
	Vang	30	0	10	**60**
ahap	**Vneu**	**57**	40	0	3
	Vhap	3	**87**	0	10
	Vsad	23	13	**33**	30
	Vang	10	10	3	**83**
asad	**Vneu**	**83**	3	13	0
	Vhap	20	**67**	13	0
	Vsad	13	0	**70**	17
	Vang	20	7	20	**73**
aang	**Vneu**	**53**	0	10	40
	Vhap	0	**60**	0	40
	Vsad	13	0	**23**	63
	Vang	0	3	3	**93**

For those stimuli consisting of the combination of neutral visual expression and auditory expressions of emotion, the responses of the subjects seem to be influenced mostly by the auditory stimuli, in particular for angry.

In almost all cases, the participants in the experiment choose among the four possible responses one of the two emotions employed in the stimulus, with prevalent influence from the visual stimulus. In a few cases there is a scattered distribution of the responses, and this occurs when the two stimuli are contradictory, as for instance in the case of v-sad/a-hap and v-hap/a-sad. In these two cases one would expect that the two contradicting stimuli

neutralize each other and that the response of the participants would be neutral, but the results show instead a dispersion of the responses for the mis-matched stimuli, in particular for v-sad/a-hap. For the stimuli v-hap/a-sad, happy is the prevailing response with 67%, but in this case 20% of the confusion goes to neutral.

One of the implications of such results, which suggest that the visual modality has a stronger influence than the auditory one on the recognition of emotion, could be the possibility of eventually producing expressive audio-visual speech synthesis by simply adding visual correlates of emotions to a neutral speech synthesis.

9.4.3 Task 2: Preference Judgement

In order to determine whether the expressive talking head was preferred to the non-expressive one, a second task was carried out using six of the stimuli in task 1. The same participants as in task 1 were presented two alternative versions of the interactions: with and without visual expressiveness. They were asked to observe the six stimuli and judge which of the interactions they preferred.

At the end of the test they were also asked whether they would like to use such a system with a language tutor capable of expressing emotions, and they were given the possibility to express their own comments related to the experiment.

The results, reported in table 9.4, show that the participants in the experiment seem to prefer the stimuli with visual expression of emotion for *angry*, while for *happy*, 80% of the participants said that they preferred the stimulus without the visual expression. This result is likely to depend on the fact that in the *happy* visual expression, the tutor was smiling in such a way that his gums were shown, which did not appeal the participants in the test (many of them expressed negative comments on the happy expression).

For *sad*, 50% of the participants showed preference for the stimuli with the visual expression of emotion and 50% without. *Sad* was also the worst recognized emotion among the others.

All the participants said that they would like to use a dialogue system with a language tutor capable of showing expressive feedback, even if they would not like him to show an angry reaction as in the test. However they commented that the angry expression was exaggerated.

Table 9.4 Percentages of preferences for the different stimuli

Stimuli	Happy	Angry	Sad
With visual expression of emotion	20%	90%	50%
With neutral visual expression of emotion	80%	10%	50%

9.5 Conclusions and Discussion

The results of an experiment run to evaluate the expressiveness of the data-driven synthesis are encouraging and they clearly show that the visual expression plays a more relevant role in the recognition of expression of emotions than the auditory one. However, there are still some obvious artefacts in the output of the expressive synthesis used in the evaluation test. These artefacts can be traced back to inconsistencies in the training data, for instance the fact that the happy expression displayed too much of the gums above the upper teeth.

The method used for the assessment of the expressiveness of the talking head offers the possibility to evaluate talking head characteristics in a realistic usage scenario and in an interactive situation with a user, even without having a functional interactive dialogue system available.

The disadvantage of this method is that it prevents the possibility to evaluate user experience, since the judgments are not given by the direct users of the simulated system, but by external observers. This means that it is not really possible to assess whether the users would actually benefit from the presence of the talking head in the envisioned real usage application.

Moreover the evaluation paradigm presented here still suffers from the limitation that the recorded expressions of emotions used to train the models were recorded out of context, even if for the actual evaluation tests a scenario was created in which the talking head was inserted in a realistic usage context.

The results of the preference judgement seem to point out that it is not a simple question to find an adequate context for letting a talking head express the basic emotions. Not all the emotions seem to be appropriate in the context of the language tutor interaction with a student, and different users seem to have different expectations from the talking head/tutor.

When designing evaluation tests aiming at assessing the characteristics of talking heads at the macro-level, it is important to assess also whether the physical characteristics of the talking head are optimal for the envisioned application. If the learning environment chosen for the evaluation test seemed to be the most appropriate for the evaluation of different expressions of emotions that can be interpreted as feedback expressions in the given

context, at the same time the emotions expressed by the talking head/tutor in the experiment were probably not the most appropriate for the envisioned application.

10 Final Discussion

10.1 Summary

This thesis deals with human communicative behaviour related to feedback, which is analysed across languages (Italian and Swedish), modalities (auditory versus visual) and in different communicative situations (human-human versus human-machine dialogues). The intention is to give more insight into human feedback production and to provide a method to collect valuable data that could be used to control facial displays related to visual feedback in synthetic conversational agents.

The introductory part of this thesis (chapter 2) offers an historical perspective and reviews the state of the art of studies about feedback phenomena in human-human and human-machine interactions.

The procedure followed in this thesis implies initial analyses of communicative phenomena in spontaneous human-human dialogues (chapter 5 and 6) and human-machine interactions (chapter 7) selected from existing available databases and corpora (an overview of the materials used for the investigations is given in chapter 3).

The aim of these initial investigations is to learn about regularities in human communicative behaviour that could be transferred to talking heads. Then, for the sake of reproduction in talking heads, the thesis includes further detailed analyses of data which were collected in a lab environment with an acquisition set-up that allows capturing the dynamics of facial displays in dialogue situations (chapter 8).

Finally the possibilities of transferring human communicative behaviour to a talking head are discussed and some evaluation paradigms are illustrated (chapter 9).

The original contributions of this thesis are to be seen in the method developed and tested for the analysis of feedback phenomena (the method of analysis is thoroughly explained in chapter 4, in particular the coding scheme), and in the data acquisition technique suggested for the recording of 3D-data related to non-verbal feedback phenomena in semi-spontaneous dialogues (the data acquisition technique is illustrated in chapter 8).

Thanks to the method of analysis developed in this thesis it has been possible to obtain interesting results related to the acoustic and visual characteristics of short feedback expressions.

10.2 The Method

The method followed to analyse the data throughout the thesis consists in the identification of feedback in terms of reaction to the previous communicative act, and its annotations by means of a specific coding scheme.

The coding scheme is the key to all the investigations carried out in this thesis. It allows analysing and categorizing verbal and non-verbal feedback expressions according to their type, direction and the semantic-pragmatic function they convey in the given context.

The results of the analysis carried out on different kinds of materials have shown that the coding scheme seems to be feasible for annotation of feedback phenomena across languages (Italian and Swedish), modalities (auditory and visual) and communicative situations (human-human versus human-machine interactions).

The tags used for the annotation help the automatic retrieval of several quantitative measures, such as the number of occurrences of feedback expressions, their type and position. This information provides an overall picture of the distribution of feedback expressions. A more detailed picture of the specific functions that feedback expressions can carry out in the given context is provided by the coded explicit semantic-pragmatic function and by some acoustic and visual characteristics of the expressions under observation.

Producing a set of categories and relative labels to annotate a corpus presupposes that there should be one correct interpretation for each phenomenon under analysis. This might be true for the analysis at the syntactic, morphological, and phonological level, since specific annotation systems that have been prepared and tested for the past 100 years are available. However for the analysis and categorization of the semantic-pragmatic function of verbal and non-verbal feedback phenomena the methodology is still being shaped. As a consequence one of the risks in the investigation of the semantic-pragmatic function of verbal and non-verbal communicative behaviour is that of subjective interpretations.

At the syntactic level a noun is a noun, and even if there might be some lexical items that have an ambiguous syntactic status, the context will help in disclosing the role that the item receives in the given situation. At the semantic-pragmatic level a given communicative behaviour can be interpreted by external observers in different ways, because people are sensitive to different aspects of information, and also because at this level, apparently, there is not a one-to-one relationship between a communicative behaviour and a meaning. Even if the contextual information plays a very important role for the correct interpretation of the phenomena at the semantic-pragmatic level, there is not one correct interpretation of a phenomenon, and the only way to validate the annotations is to rely on statistical evaluation methods.

In this thesis the reliability of the categories designed to annotate the semantic-pragmatic functions of feedback expressions has been tested running the following three tests:

- Stability test, or inter-variance test, which checks whether the same coder varies his/her judgments over time.
- Reproducibility test, or inter-coders-variance, which checks the agreement in the codings of two coders.
- Accuracy test, which compares the codings produced by two coders to a standard, if a standard is available.

The results obtained can be considered as positive. The reliability and ease of use of the categories in the coding scheme and feasibility across languages is indicated by the scores of the *Kappa* coefficients, which range between 0.6 and 0.94, thus indicating a fair degree of agreement for the assignment of the pre-defined semantic-pragmatic categories for feedback functions (see sections 5.2.3 and 5.2.3).

Considering the fact that assigning pre-defined theoretical categories always implies a dose of subjectivity, the results obtained in the reproducibility test can indeed be considered as positive.

The risk of subjective interpretations arises also in the case of annotation of non-verbal communicative behaviour, at a level that can be called "gestural" or "non-verbal". The typology and communicative function of a given expression (a facial display, a hand or arm gesture, and a body posture) might be interpreted in different ways by different observers, because there is not always one single correct interpretation of a phenomenon.

A formal assessment of the degree of agreement in the identification of non-verbal feedback phenomena, and the assignment of the semantic-pragmatic categories for feedback function to non-verbal feedback phenomena has not been run on the materials analysed in this thesis. However the reliability and feasibility of the categories designed to code the semantic-pragmatic functions of non-verbal feedback has been tested during the MUMIN workshop on "Multi-modal Annotation" held in 2004 at KTH, Stockholm [Allwood et al. 2005; 2006].

The results of the reproducibility test run by two non-expert coders who independently coded the semantic-pragmatic functions of facial displays related to feedback in a one-minute clip extracted from a TV talk-show in Danish, showed *Kappa* scores ranging between 0.68 and 0.9. This result has been taken as a positive indication of the validity of the categories for the annotation of non-verbal behaviour, not only across languages, but also across modalities.

10.3 The Materials

The materials used for the investigations presented in this thesis span from spontaneous conversations video recorded in real communicative situations, to semi-spontaneous dialogues obtained with different eliciting techniques, such as map-task and information-seeking scenarios.

The spectrum of varying communicative situations, recording set-ups and speaking styles offered by the materials analysed in this thesis, represented a challenging testing ground for the method developed for the purpose of analysing feedback phenomena. At the same time using a variety of materials for the analysis of feedback phenomena offered the possibility to get a more varied picture of the production of feedback phenomena in different communicative situations.

The initial idea was to use available resources in Italian and Swedish. However this presented several limitations, which were mainly due to the fact that the original purposes of the data collections were different from the actual purpose of the analysis carried out in this thesis. Moreover it turned out to be unfeasible to get comparable materials recorded in similar circumstances in both Italian and Swedish. As a consequence a cross-linguistic study could be carried out only on the audio recordings of map-task dialogues.

The original idea of using only already existing available materials had to give way to the need of collecting a specific corpus of data that could better capture the dynamics of facial displays related to feedback phenomena.

10.3.1 The Data Acquisition

Two data acquisitions have been performed with a recording set-up that allows recording audio-visual tri-dimensional data: audio data is recorded on a DAT-tape and visual data is recorded both by means of one or two digital video camera/s and with the optical motion tracking system Qualisys. Attaching infrared reflecting markers to the subject's face enables the system to register the 3D-coordinates for each marker at a frame-rate of 60Hz, i.e. every 17ms (see section 3.3.3 and 3.3.4). This recording set-up was shown to be feasible for the recording of high precision and high quality data that are helpful both to get more insight into human non-verbal communicative behaviour and to provide data for the reproduction of facial displays in talking heads.

So far optical motion capture systems had been used to register articulatory movements as well as facial displays in prompted speech, and these data have been successfully applied to model visual articulation and expressiveness in talking heads; by analogy, it seems likely to foresee that models for each facial display and in particular for each head movement that is considered to have an important communicative function could be trained by using dialogic speech recorded by means of motion capture systems.

However the acquisition technique presents several limitations, due to the constraints of the recording set-up and the elicitation technique. The most evident effect of these limitations is the relative small dimension of the acquired corpora.

The use of a specific elicitation technique was necessary to acquire a set of controlled and structured data. However even if attempts were made to try to elicit spontaneous interactions in scenarios which were thought to lead to the production of several instances of non-verbal feedback behaviour, it was not achievable to obtain a complete control over the production of non-verbal behaviour, due to the impossibility to predict when exactly they might occur during speech production.

Notwithstanding the constraints, the audio-visual recordings of spontaneous dialogues and of elicited semi-spontaneous dialogues have been very useful sources for the systematic investigation of feedback phenomena on an empirical basis.

The use of a detailed coding scheme and the help of dedicated tools for audio-visual analysis have facilitated the process of annotation and retrieval of the data.

10.4 The Tools for Audio-Visual Analysis

The experience with the tools for audio-visual analysis has shown the usefulness of these tools, but also their fragility, which is mainly due to the fact that they are often developed by researchers for their own research purposes and therefore often suffer from several deficiencies when it comes to wider applications. Moreover they are susceptible to changing video formats and often incompatible with different available platforms.

However thanks to these tools and to the helpfulness of their developers, it has been possible to analyse audio-visual materials and obtain interesting results that show evidence that it is possible to identify a general pattern for each specific head movement, even if it is not possible to identify a one-to-one relationship between a specific verbal feedback and a specific head movement, nor between a specific head movement and its semantic-pragmatic function.

10.5 Acoustic Characteristics of Feedback Words

The analysis of the acoustic characteristics of feedback words showed that, notwithstanding the observable variability in the realization of feedback across speakers, languages and communicative situations, some regular behaviour could be observed.

The results of the acoustic analysis of feedback words such as *ja, sí* and m-like words in Italian and Swedish show evidence that acoustic characteristics, such as duration and F0 contour, reflect the

semantic-pragmatic function carried out by feedback words both in Swedish and in Italian, even if different F0 contours and durations are produced in the two languages to express the same function. In particular the results showed that the difference in the durations of Italian *sì* with different semantic-pragmatic function resulted to be significant, especially for the function GIVE CONTINUATION YOU GO ON (see section 5.4.1).

Moreover the results of the perceptual test run to verify whether the acoustic characteristics of feedback words extracted from their context can be considered as reliable cues for the identification of their semantic-pragmatic function showed that, in particular for the Italian stimuli, subjects obtained high recognition scores in the identification task (see section 5.4.7.2). This result might depend on the fact that different semantic-pragmatic functions are signalled by different acoustic characteristics.

It is therefore possible to conclude that the investigation of the prosodic marking of short expressions in relation to their specific communicative function is of fundamental interest when it comes to technological applications. A dialogue system might benefit from an on-line prosodic analysis of human-users' production to better interpret the intention of human utterances. This might hopefully contribute to enhance human-machine interactions.

10.6 Visual Characteristics of Feedback

The results of the study aiming at investigating the communicative function of head nods in Swedish dialogic speech (see section 8.3) show that in 70% of cases the function of head nods is related to feedback. Besides FEEDBACK head nods are produced to signal FOCUS and EMPHASIS, POLITENESS, TURN YIELD and to give POSITIVE ANSWERS.

The results of the analysis of head movements produced during dialogic speech in Swedish to signal feedback show some interesting trends as for instance that nods and jerks are the most common head movements used to signal feedback. In particular one of the most clear and interesting results of the investigation of head nods related to feedback is the different use of SINGLE and REPEATED NOD: SINGLE NOD tends to be produced to show CONTINUATION, REPEATED NOD is seldom produced with this function, rather REPEATED NOD is more frequently produced to GIVE ACCEPTANCE and to REQUIRE ACCEPTANCE.

Thanks to the high precision of the data it was possible to test the hypothesis based on the observations made in chapter 6 that short, minimal head nods might be related to short verbal feedback expression carrying out the function FEEDBACK GIVE CONTINUATION YOU GO ON. This resulted to be true especially for the subject who was recorded with the markers glued on his face (subject-S in section 8.3.3).

The results of the duration analysis show that SINGLE NOD can have different durations depending on the different communicative functions. If these differences in the duration will be proven to be significant on a bigger amount of data, they could represent a distinctive cue for the different communicative functions that head nods can carry out. This cue could be then exploited in the implementation of communicative head nods in talking heads used in human-machine interfaces.

Enabling embodied conversational agents to interact with humans in an effective way requires both the understanding of how communicative non-verbal behaviour is naturally performed by humans and the possibility to capture the exact dynamics of the non-verbal behaviour produced by humans.

Having available high precision data makes it possible to control communicative non-verbal behaviour in embodied conversational agents.

The promising results obtained with the automatic head nod detector (see section 8.4.3) also envisage the possibility of obtaining larger annotated databases for the training and testing of eventual data-driven models of head movements in talking head.

10.7 Concluding Remarks

The results provided in this thesis are not meant to be exhaustive; however they show how pervasive feedback phenomena are in human-human communication and survey the most common types and specific functions carried by feedback in different communicative situations. The results underline in particular the fact that the prosodic characteristics of feedback expressions, as well as the visual information carried out by some facial expressions and head movements that signal feedback in spoken communicative interactions, is without doubt extremely important. Nevertheless it has to be borne in mind that it is not enough to simply consider the phonetic form of verbal feedback expressions and the physical form of non-verbal feedback expressions, since the specific functions carried out by feedback expressions are highly dependent on the context in which they occur. In particular the timing and their precise placement within a sequence of speech are of fundamental importance for the correct interpretation of feedback.

The expectation is that a dialogue system might benefit from an on-line prosodic and contextual analysis of human-users' production to better interpret the intention of human utterances. Moreover it is expected that users interacting with a conversational embodied system which is able to provide and understand verbal and non-verbal feedback, might experience the interaction as more human-like. The appropriate production of feedback signals by the system has been proven to enhance not only the interaction

between human users and dialogue systems [Takeuchi & Nagao 1993; Rajan et al. 2001], but also human satisfaction [Okato et al. 1998].

Given the fact that in human-human communication the production of verbal and non-verbal feedback is a pervasive phenomenon, it is recommended that the production of feedback should be a natural behaviour also in human-machine communication. The suggestion is therefore to enable embodied conversational agents to produce appropriate feedback signals during interactions with humans. To do this in an effective way requires both the understanding of how communicative non-verbal behaviour is naturally performed by humans and the possibility to capture the exact dynamics of the non-verbal behaviour produced by humans.

The production of feedback by the system should not be based on intuition, but rather on the results of empirical studies which show how the production of feedback is distributed in human-human interactions in different communicative situations. However the reproduction of the rich inventory of feedback phenomena observed in human production might not be feasible, for this reason the categorization of the specific type and semantic-pragmatic functions of verbal and non-verbal feedback proposed in this thesis could be a good solution for implementation. A good starting point could be the exploitation of one of the most clear and interesting results of the investigation of head movements reported in chapter 8 of this thesis, that is the different use of SINGLE and REPEATED NOD to signal different feedback function.

10.8 Potential Future Work

Having available a reliable method for the analysis of feedback phenomena in different modalities, a feasible data acquisition set-up for the recording of 3D-data related to non-verbal feedback phenomena in dialogic speech and an automatic tool for the detection of head movements, the remaining challenge is the reproduction and evaluation of the non-verbal phenomena related to communicative feedback in talking heads.

However in order to build and train models it is necessary to have a greater amount of data available. For this reason the collection of more data that can provide further quantitative results which might be better exploited in the field of human-machine interfaces is advocated.

Once the data is available and the implementation of non-verbal phenomena related to communicative feedback in talking head is performed, an accurate evaluation needs to be designed and run in order to assess the effectiveness and appropriateness of the visual expressions related to communicative feedback… but this sounds like the interesting topic of another thesis, this one ends here!

References

Ahlberg, J., Pandzic, I. S., You, L. (2002). 'Evaluating MPEG-4 Facial Animation Players' In Pandzic, I. S. & Forchhimer, R. (Eds), *MPEG-4 Facial Animation: the standard, implementation and applications*, Wiley & Sons, Chichester: 287-291.

Allwood, J. (1988). Om det svenska systemet för språklig återkoppling. In Linell, P., Adelswärd, V., Nilsson, T. & Pettersson P.A. (Eds.) *Svenskans Beskrivning* 16, Vol.1. (SIC 21a) University of Linköping, Tema Kommunikation: 89-106.

Allwood, J. (1993). Feedback in Second Language Acquisition. In Perdue C. (Ed.) *Adult Language Acquisition. Cross-Linguistic Perspectives* Cambridge, Cambridge University Press, II: 196-235.

Allwood, J. (2001a). Cooperation and Flexibility in Multi-modal Communication. In Bunt, H. & Beun R. J. (Eds.) *Cooperative Multi-modal Communication. Lecture Notes in Computer Science* 2155, Springer Verlag, Berlin/Heidelbert: 113-123.

Allwood, J. (2001b). Dialog Coding - Function and Grammar: Gothenburg Coding Schemas. *Gothenburg Papers in Theoretical Linguistics 85,* Dept of Linguistics, University of Gothenburg: 1-67.

Allwood, J., Björnberg, M., Grönqvist, L., Ahlsén, E. & Ottesjö, C. (2000). The Spoken Language Corpus at the Linguistics Department. *Forum Qualitative Social Research,* 1(3): 22-32.

Allwood, J. & Cerrato, L. (2003). A study of gestural feedback expressions. In Paggio, P., Jokinen, K. & Jönsson, A. (Eds.) *Proc. of the First Nordic Symposium on Multi-modal Communication.* Copenhagen: 7-20.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. & Paggio, P. (2005). The MUMIN Annotation Scheme for Feedback, Turn management and Sequencing. In Allwood, J., Dorriots, B. & Nicholson, S. (Eds.) *Proc. of the 2nd Nordic Symposium on Multi-modal Communication.* Gothenburg, Sweden: 91-109.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C. & Paggio, P. (2006). A Coding Scheme for the Annotation of Feedback, Turn management and Sequencing Phenomena. In Martin et al (Eds.) *Proc. of the LREC Workshop on Multi-modal Corpora. From Multi-modal Behaviour to Usable Models*. Genova, Italy: 38-42.

Allwood, J., Grönqvist, L., Ahlsén, E. & Gunnarsson, M. (2003). Annotations and tools for an activity based spoken language corpus. In van Kuppevelt, J. & Smith, R. (Ed.). *Current and New directions in Discourse and Dialogue*. Kluwer Academic Publishers, Dordrecht: 1-18.

Allwood, J., Nivre, J. & Ahlsén, E. (1992). "On the semantics and pragmatics of linguistic feedback" *Journal of Semantics,* 9(1):1-26.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. & Weinert, R. (1991). "The HCRC Map Task Corpus". *Language and Speech,* 34: 351-366.

Argyle, M. (1988). *Bodily Communication* (2nd edn.). London: Methuen.

Argyle, M. & Cook, M. (1976). *Gaze and Mutual Gaze.* Cambridge University Press.

Banse, R. & Scherer, K.R. (1996) "Acoustic profiles in vocal emotion expression". *Journal of Personality and Social Psychology,* 70(3), 614-636.

Bateson, G. (1972). *Steps to and Ecology of Mind.* New York, Ballantine Books.

Batliner, A., Fischer, K., Huber, R., Spilker, J. & Nöth, E. (2000). Desperately seeking emotions or: Actors, wizards, and human beings. In *Proc. of the ISCA Workshop on Speech and Emotion.* Northern Ireland: 195-200.

Bavelas, J.,B. et al. (1992). Interactive Gestures. *Discourse Processes*, 15: 469-489.

Bazzanella, C. (1994). *Le facce del parlare*. Firenze, La Nuova Italia.

Beckman, N. (1968). *Svensk språklära*. Stockholm, Bonniers.

Bell, L., Gustafson, J. (2000). Positive and Negative User Feedback in a Spoken Dialogue Corpus. In *Proc. of the International Conference of Spoken Language Processing (ICSLP),* Beijing, China, China Military Friendship Publishing*:* 1 589-592.

Bergstöm, A., Cerrato, L., Cordeiro, C. & Svaerke Hansen, N. (2002). A comparison of verbal feedback strategies across languages. Final report of the NORFA Course on "Using Spoken Language Corpora". Dept. of Linguistics, Gothenburg University, Sweden.

Berry, A. (1994). Spanish and American Turn-taking Styles: A Comparative Study. *Pragmatics and Language Learning*, 5, University of Illinois, Urbana-Champaign: Division of English as an International Language: 180-190.

Bertrand, R., Boyer, J., Cavé, C., Guaitella, I. & Santi, S. (1995). Relationship between gestures and voice in verbal interaction: prosodic and kinesic aspects of back-channel signals. *International Conference of Phonetic Sciences (ICPhS),* Stockholm, 2: 746-749.

Beskow, J. (2003). *Talking Heads - Models and Applications for Multi-modal Speech Synthesis*. Speech Music and Hearing. KTH, Stockholm, Sweden.

Beskow, J. & Cerrato, L. (2004). Evaluation of the expressivness of a Swedish talking head in the context of human-machine interaction. In *Atti del convegno del Gruppo di Studio sulla Comunicazione Parlata (GSCP)* Padua, Italy (in press).

Beskow, J. & Cerrato, L (2007) Synthesis and Evaluation of Expressiveness in a Swedish Talking Head Trained on a Multi-modal Expressive Speech Corpus. Submitted to Special issue of the International *Journal of Language Resources and Evaluation:* Multi-modal Corpora for modelling human multi-modal behaviour.

Beskow, J., Cerrato, L., Cosi, P., Costantini, E., Nordstrand, M., Pianesi, F., Prete, M. & Svanfeldt, G. (2004a). Preliminary Cross-cultural Evaluation of Expressiveness in Synthetic Faces. In André, E., Dybkjaer, L., Minker, W. & Heisterkampf, P. (Eds.), *Proc. of Tutorial and Research Workshop on Affective Dialogue Systems* (*ADS*), Kloster Irsee, Tyskland: 240-243.

Beskow, J., Cerrato, L., Granström, B., House, D., Nordstrand, M. & Svanfeldt, G. (2004b). The Swedish PF-Star Multimodal Corpora. In *Proc. of LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multi-modal Input and Output Interfaces,* Lisboa, Portugal: 34-37.

Beskow, J., Engwall, O. & Granström, B. (2003). Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements. In Solé, M. J., Recasens, D. & Romero, J.(Eds.) *Proc. of the International Conference of Phonetic Sciences (ICPhS)* Barcelona, Spain: 431-434.

Beskow, J., Granström, B. & House, D. (2006). Visual correlates to prominence in several expressive modes. In *Proc. of Interspeech,* Pittsburg, PA: 1272–1275.

Beskow, J., Granström, B., House, D. & Lundeberg, M. (2000). Verbal and visual prosody in multi-modal speech perception. In *Proc. of Nordic Prosody VIII*: 77-87.

Beskow, J., Granström, B., & Spens, K-E. (2002). Articulation strength - Readability experiments with a synthetic talking face. In *Proc. of Fonetik* Stockholm, Sweden Speech, Music and Hearing, Quarterly Progress and Status Report 44(1): 97-100.

Beskow, J. & Nordenberg, M. (2005). Data-driven Synthesis of Expressive Visual Speech using an MPEG-4 Talking Head. In *Proc. of Interspeech,* Lisboa, Portugal: 792-796.

Birdwhistell, R. L. (1970). *Kinesics and Context: Essays on Body Motion Communication.* Philadelphia, University of Pennsylvania Press.

Blairy, S., Herrera, P. & Hess, U. (1999). "Mimicry and the judgment of emotional facial expressions". *Non-Verbal Behaviour,* 23: 5-41.

Bolinger, D. L. (1989). *Intonation and Its Uses. Melody in Grammar and Discourse*. Stanford University Press.

Boyle, E., Anderson, A. & Newlands, A. (1994). "The effects of visibility on dialogue and performance in a cooperative problem solving task". *Language and Speech,* 37: 1-20.

Brennan, S. E. & Hulteen, E. A. (1995). Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems,* 8: 143-151.

Brunner, L. J. (1979). "Smiles can be back-channels". *Journal of Personality and Social Psychology,* 37: 728-734.

Bull, P.E. (1987). Gesture and posture, Oxford, England: Pergamon Press

Campbell, N. (2001). Building a corpus of natural speech - and tools for the processing of expressive speech. In *Proc. of EUROSPEECH*, Aalborg, Denmark: 1525-1528.

Campbell, N. (2003). Modelling affect in Speech Communcation. In *Proc. of International Conference of Spoken Language Processing (ICSLP)*, Beijing China, China Military Friendship Publishing, 4: 468-471.

Campbell, N. (2004). Accounting for voice-quality variation. In *Proc. of the 2nd International Conference on Speech Prosody*, Nara, Japan: 217-220.

Carletta, J. (1996). "Assessing agreement on classification tasks: the Kappa statistics". *Computational Linguistics*, 22(2): 249-254.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. & Anderson, A.H. (1997). "The Reliability of a Dialogue Structure Coding Scheme". *Computational Linguistics,* 23(1): 13-31.

Caspers, J. (2000). Melodic characteristics of back-channels in Dutch Map Task dialogues. In *Proc. of the International Conference on Spoken Language Processing (ICSLP*), Beijing: China, China Military Friendship Publishing, 1: 565-568

Caspers, J. (2003). Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics,* 31: 251-276.

Cassell, J. (2000). Nudge Nudge Wink Wink: Elements of Face-to-Face Conversation for Embodied Conversational Agents, in Cassell, J. et al. (Eds.), *Embodied Conversational Agents*: 1-27. Cambridge, MA: MIT Press. (2000): 1-27.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H. & Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Proc. of ACM Computer Human Interaction (CHI)*, Pittsburgh, PA: 520-527

Cassell, J. & Thórissón, K. (1999). "The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents". *Applied Artificial Intelligence,* 13: 519-538.

Castagneto, M. & Ferrari, G. (2003). Influence of regional features on Map-Task dialogues. In Kruijff-Korbayová, I. & Kosny C. (Eds.) *Proc. of the 7th Workshop on the Semantics and Pragmatics of Dialogue,* Saarbrucken, Universität des Saarlandes: 22-25.

Cerrato, L. (1999). Il feedback verbale nei dialoghi elicitati con la tecnica del map task. In *Atti del GFS* Naples, Italy: 25-36.

Cerrato, L. (2002a) A comparison between feedback strategies in Human-Human and Human-Machine communication. In *Proc. of International Conference of Speech and Language Processing (ICSLP),* Denver, Co: 557-560.

Cerrato, L. (2002b). Some characteristics of feedback expressions in Swedish. In *Proc. of Fonetik* Stockholm, Sweden Speech, Music and Hearing, Quarterly Progress and Status Report: 41-44.

Cerrato, L. (2002c). A Study of Verbal Feedback in Italian. In Henrichsen P.J. (Ed.). *Proc. of the NORDTALK Symposium on Relations between Utterances,* Copenhagen, Danmark Copenhagen Working papers in LSP: 80-97.

Cerrato, L. (2003). On the acoustic, prosodic and gestural characteristics of "m-like" sounds in Swedish. Feedback in Spoken Interaction. *Nordtalk Symposium,* Gothenburg, Sweden, *Gothenburg Papers in Theoretical Linguistics:* 18-31.

Cerrato, L. (2004). A coding scheme for the annotation of feedback phenomena in conversational speech. In Martin, J.C. (Ed.), *Proc. of LREC Workshop on Models of Human Behaviour for the Specification and Evaluation of Multi-modal Input and Output Interfaces,* Lisboa: 25-28.

Cerrato, L. (2004). A comparative study of verbal feedback in Italian and Swedish map-task dialogues. In Henrichsen (Ed.) *Proc. of the Nordic Symposium on the Comparison of Spoken Languages*, Copenhagen, Danmark, Fredriksberg Bogtrykkeri: 99-126.

Cerrato, L. (2005a). The communicative function of "sì" in Italian and "ja" in Swedish: an acoustic analysis. In *Proc. of Fonetik* Göteborg, Sweden: 41-44.

Cerrato, L. (2005b). Linguistic functions of head nods. In Allwood, J., Dorriots, B. & Nicholson, S. (Eds.) *Proc. of the 2nd Nordic Symposium on Multi-modal Communication* Gothenburg, Sweden: 137-152.

Cerrato, L. & D'Imperio, M. (2003). Duration and Tonal Characteristics of Short Expressions in Italian. In Solé, M. J., Recasens, D. & Romero, J.(Eds.) Proc. of the *International Conference of Phonetic Sciences (ICPhS)* Barcelona, Spain: 1213-1217.

Cerrato, L.& Skhiri, M. (2003). A method for the analysis and measurement of communicative head movements in human dialogues. *In Proc. of Audio Visual Speech Processing (AVSP, ITRW)*, St. Jorioz, France: 251-256.

Cerrato, L. & Ekeklint, S. (2004). Evaluating users reactions to human-like interfaces: Prosodic and paralinguistic features as new evaluation measures for users' satisfaction. In Ruttkay, Zs. & Pelachaud, C. (Eds.) *From Brows to Trust Evaluating Embodied Conversational Agents*. Kluwer's Human-Computer Interaction Series 7: 101-125.

Cerrato, L. & Svanfeldt, G. (2005). A method for the detection of communicative head nods in expressive speech. In Allwood, J., Dorriots, B. & Nicholson, S. (Eds.), *Gothenburg papers in Theoretical Linguistics 92: Proc. from The Second Nordic Conference on Multi-modal Communication,* Gothenburg University, Sweden: 153-165.

Chovil, N. (1992). "Discourse-Oriented Facial Displays in Conversation". *Research on Language and Social Interaction,* 25: 163-194.

Clark, H. & Brennan, S. (1991). Grounding in communication. In Resnick, L.B., Levine J.M. & Teasley S.D. (Eds.). *Perspectives on Socially Shared Cognition* Washington, USA, APA Books: 127-149.

Clark, H. & Schaefer, E. (1989). Contributing to Discourse. *Cognitive Science* 13: 259-294.

Cohen, M.M., Massaro, D.W. (1993). Modelling Coarticulation in Synthetic Visual Speech. In Magnenat-Thalmann N. & Thalmann D. (Eds.), *Models and Techniques in Computer Animation,* Springer Verlag, Tokyo:139-156.

Cooper, B., Brna, P. & Martins, A. (2000). Effective Affective in Intelligent Systems - Building on Evidence of Empathy in Teaching and Learning. In Paiva, A. (Ed.), *Affect in Interactions: Towards a New Generation of Computer Interfaces*, Berlin, Springer: 21-34.

Cosi, P., Fusaro, A. & Tisato, G. (2003). LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model. In *Proc. of Eurospeech*, Geneva, Switzerland: 127-132.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Kollias, S., Fellenz, W. & Taylor, J. G. (2001). "Emotion recognition in human-computer interaction". *IEEE Signal Processing Magazine,* 18 (1): 32-80.

D'Imperio, M. (2002). "Italian intonation: An overview and some questions". *Probus,* 14 (1): 37-69.

Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York, USA, Grosset-Putnam.

Darwin, C. (1872, reprinted in 1999). *The Expression of Emotions in Man and Animals*, Fontana Press.

De Carlo, D., Revilla, C., Stone, M. & Venditti, J. (2002). "Making discourse visible: Coding and animating conversational facial displays". *Computer Animation*: 11–16.

Dittmann, A. T. (1972). "Development factors in conversational behavior". *Journal of Communication*, 22, 404-423.

Dittmann, A. T. & Llewellyn, L. G. (1968). "Relationship between vocalizations and head nods as listener responses". *Journal of Personality and Social Psychology*, 9: 79-84.

Duncan, S. (1972). "Some Signals and Rules for Taking Speaking Turns in Conversations". *Journal of Personality and Social Psychology,* 23 (2): 283-292.

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society* 2: 161-180.

Duncan, S. & Fiske, D. (1977). *Face-to-Face Interaction*. Erlbaum, Hillsdale, NJ.

Edlund, J. & Heldner, M. (2005). "Exploring prosody in interaction control" *Phonetica, 65*: 215-226.

Edlund, J., Heldner, M. & Gustafson, J. (2005). Utterance segmentation and turn-taking in spoken dialogue systems. In *Proc. of Sprachtechnologie, Mobile Kommunikation und Linguistische Ressourcen* Frankfurt am Main, Germany, Peter Lang: 576-587.

Edlund, J. & Nordstrand, M. (2002). Turn-taking Gestures and Hourglasses in a Multi-modal Dialogue System. In *Proc. of ISCA Workshop on Multi-Modal Dialogue in Mobile Environments*, Kloster Irsee, Germany, paper 31.

Eibl-Eibesfeldt, I. (1970). *Ethology. The Biology of Behavior*. New York, USA, Holt, Rinehart and Winston, Inc.

Ekman, P. (1979). About brows: Emotional and conversational signals. Human ethology: Claims and limits of a new discipline. In Foppa M., Lepenies W. & Ploog, D. (Eds.) *Contributions to the Colloquium von Cranach*. Cambridge University Press: 169-248.

Ekman, P. (1982). *Emotion in the Human Face.* New York, Cambridge University Press.

Ekman, P. (1993). Facial Expression and Emotion. *American Psychologist,* 48(4): 384-392.

Ekman, P. & Friesen, W.V, (1978). *Facial Action Coding System. Manual*. Palo Alto, CA., Consulting Psychologists Press.

El Emam, K. (1999). “Benchmarking Kappa: Interrater agreement in software process assessments”. *Empirical Software Engineering*, Kluwer Academic Publishers, 4:113-133.

Fabri, M., Moore, D.J. & Hobbs, D.J. (2002). Expressive Agents: Non-verbal Communication in Collaborative Virtual Environments *AAMAS Workshop on Embodied Conversational Agents: "Embodied conversational agents - let's specify and evaluate them”*, Bologna, Italy.

Ferrer, L., Shriberg, E. & Stolke, A. (2002). Is the speaker done yet? Faster and more accurate end-of-utternace detection using prosody in human-computer dialogue. *International Conference of Spoken Language Processing (ICSLP),* Denver, Co: 2061-2064.

Fries, C. C. (1952). *The structure of English.* New York, Harcourt Brace.

Gardner, R. (2001). *When Listeners Talk.* Philadelphia USA, John Benjamins Publishing Company.

Goffman, E. (1955). “On face, work I”. *Psychiatry,* 18: 213-231.

Goldin-Meadow, S. (2003). *Hearing Gesture: How our Hands Help us Think.* Cambridge, MA: Harvard University Press.

Goodwin, C. (1981). *Conversational Organization: Interaction Between Speakers and Hearers*. New York, NY Academic Press.

Graf, H.P., Cosatto, E., Strom, V. & Huang F.J. (2002). Visual prosody: Facial movements accompanying speech. In *Proc. of the Fifth International Conference of Automatic Face and Gesture Recognition*: 397-401.

Granström, B., House, D. & Beskow, J. (2002). Speech and gestures for talking faces. In House D., Granström B., Karlsson I. (Eds.) *Multi-modality in Language and Speech Systems* Dordrecht, Kluwer Academic Publishers 19: 209-241.

Granström, B., House, D. & Swerts, M. G. (2002). Multi-modal feedback cues in human-machine interactions. In Bel, B. & Marlien, I. (Eds.) *Proc. of Speech Prosody,* Aix-en-Provence: Laboratoire Parole et Langage: 347-350.

Grice, H. P. (1975). *Logic and conversation. Syntax and Semantics*. New York, Academic Press, 3: 41-58.

Grönnum, N. (1991). Terminality and completion in Danish, Swedish and German. In *Proc. of the International Conference of Phonetic Sciences (ICPhS)* Aix en Provence, France 4: 270-273.

Gullberg, M. (1998). *Gesture as a communication strategy in second language discourse. A study of learners of French and Swedish*. Lund: Lund University Press.

Gustafson-Čapková S. (2005). *Integrating Prosody into an Account of Discourse Structure*, Stockholm University.

Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D. & Wirén, M. (2000). AdApt – a multi-modal conversational dialogue system in an apartment domain. In *Proc. of International Conference of Spoken Language Processing* (*ICSLP*), Beijing, China, China Military Friendship Publishing, 2:134-137.

Gybbon, D., Mertins, I. & Moore, R. K. (2000). *Handbook of multi-modal and spoken dialogue systems.* Kluwer Academic Press.

Haggard, E. A. & Isaacs, F. S. (1996). Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In Gottschalk, L.A. & Auerback, A.H. (Eds.) *Methods of Research in Psychotherapy*. Appleton Century Crofts. New York: 154–165.

Harrison, R. P. (1974). *Beyond Words*. Englewood Cliffs, NJ, Prentice Hall.

Helgason, P. (2002). *Preaspiration in the Nordic Languages: Synchronic and Diachronic Aspects*. Linguistics, Stockholm University, Sweden.

Hirshberg, J. (2002) "Communication and prosody: Functional aspects of prosody". *Speech Communication,* 36: 31-43.

Hirshberg, J. & Nakatani, C. (1996). A prosodic analysis of discourse segments in direction-giving monologues. In *Proc. of Association of Computational Linguistics (ACL)*, Santa Cruz, CA: 286-293.

Hjalmarsson, A. (2002). *Evaluating AdApt, a multi-modal conversational dialogue system, using PARADISE.* Speech Music and Hearing, KTH, Stockholm, Sweden.

House, D. (2005). "Phrase-final rises as a prosodic feature in wh-questions in Swedish human-machine dialogue". *Speech Communication,* 46: 268-283.

Hällgren, Å. & Lyberg, B. (1998). Visual speech synthesis with concatenive speech. In *Proc. of Audio-Visual Speech Processing (AVSP),* Terrigal Australia: 181-190.

Höök, K. (2002). Evaluation of Affective Interaction. *AAMAS Workshop on Embodied Conversational Agents: "Embodied conversational agents - let's specify and evaluate them",* Bologna, Italy.

Izard, C. (1977). *Human Emotions*. New York, Plenum.

Jefferson, G. (1973). "A case of precision timing in ordinary conversation: Overlapped tag-positioned. Address terms in closing sequences". *Semiotica,* 9: 47–96.

Johnson-Laird, P. N. & Oatley, K. (1989). "The language of emotions: An analysis of a semantic field". *Cognition and Emotion,* 3: 81-123.

Jurafsky, D., Shriberg, E., Fox, B. & Curl, T. (1998). Lexical, Prosodic, and Syntactic Cues for Dialogue Acts. *ACL/COLING Workshop on Discourse Relations and Discourse Markers*, Montreal, Canada.

Katagiri, Y., Sugito M. & Nagano-Madsen Y. (1999). The Forms and Prosodic Characteristics of Backchannels in Tokyo and Osaka Japanese. In P*roc. of the International Congress on Phonetic Sciences (ICPhS),* San Francisco, CA: 2411-2414.

Kendon, A. (1967). Some functions of gaze direction in social interactions. *Acta Psychologica,* 26: 22-63.

Kendon, A. (1975) Gesticulation, speech and the gesture theory of language origins. *Sign Language Studies,* 9: 349-373.

Kendon, A. (1993) Human gesture. In Ingold T. & Gibson, K. R. (Eds.) *Tools, Language and Cognition*. Cambridge: Cambridge University Press: 43-62.

Kipp, M. (2001). Anvil - A Generic Annotation Tool for Multi-modal Dialogue. In *Proc. of EUROSPEECH,* Aalborg, Danmark: 1367-1370.

Knapp, M. L. (1972) *Non-verbal Communication in Human Interactions*. Reinhart and Winston Inc., New York.

Knapp, M. L. & Hall, J. A. (2002). *Non-verbal communication in Human Interactions*.Wadsworth.

Kohler, K. J. (2004). Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions. In Fant, G. et al. (Eds.) *From traditional phonology to modern speech processing*, Beijing Foreign Language Teaching and Research Press: 205-214.

Koiso, H., Horiuchi, S., Tutiya, A., Ichoikawa, I. & Den, Y. (1998). "An analysis of turn-taking and back-channels based on prosodic and syntactic features in Japanese map-task dialogues". *Language and Speech,* 41 (3-4): 295-321.

Krahmer, E., Swerts, M., Theune, M. & Weegels, M. (2002). "The dual of denial: two uses of disconfirmations in dialogue and their prosodic correlates". *Speech Communication,* 36(1-2): 133-145.

Krahmer, E., van Buuren, S., Ruttkay, Zs. & Wesselink, W. (2003). Audio-visual Personality Cues for Embodied Agents: An experimental evaluation. *AAMAS Workshop on "Embodied Conversational Characters as Individuals*", Melbourne, Australia.

Krauss, R. M., Garlock, C. M., Bricker, P. D. & McMahon, L. E. (1977). "The role of audible and visible back-channels response in interpersonal communication". *Journal of Personality and Social Psychology,* 35: 523-529.

Krippendorff, K. (1980). *Content Analysis: an introduction to its analysis.* Sage Publications.

Laban, R. (1976). *The Language of Movement. A Guidebook to Choreutic* Boston, Plays Inc.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia, PA, University of Pennsylvania Press.

Landis, J. R. & Koch, G. G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics,* 33: 159-174.

Larsson, S. (2002) *Issue-based Dialogue Management,* Linguistics, Gothenburg University, Sweden.

Laukka, P. (2004). *Vocal Expression of Emotion: Discrete-emotions and Dimensional Accounts*, Uppsala University.

Lindahl, I. (2001). *Språklig återkoppling i spontana dialoger*. Linguistics. Lund University, Sweden.

Lippmann, R. (1997). "Speech recognition by machines and humans". *Speech Communication,* 22: 1-15.

Lundberg, M. & Beskow, J. (1999). Developing a 3D-Agent for the August dialogue system. In *Proc. of Audio-Visual Speech Processing (AVSP)*, Santa Cruz, USA: 151-156.

Lundqvist, L. O. (1995). “Facial EMG reactions to facial expressions: a case of facial emotional contagion”. *Scandinavian Journal of Psychology,* 36: 130-141.

Magno Caldognetto, E., Cosi, P., Drioli, C., Tisato, G. & Cavicchio, F. (2003). Coproduction of Speech and Emotion: Bi-Modal Audio-Visual Changes of Consonant and Vowel Labial Targets. In *Proc. of Audio-Visual Speech Processing (AVSP) ISCA Tutorial and Research Workshop on Audio Visual Speech Processing,* St Jorioz, France: 209-214.

Magno Caldognetto, E., Cosi, P., Drioli, C., Tisato, G. & Cavicchio, F. (2004). “Visual and acoustic modifications of phonetic labial targets in emotive speech: Effects of the co-production of speech and emotions”. *Speech Communication,* 44: 173-185.

Magno Caldognetto, E. & Zmarich, C. (1999). Visual spatio-temporal characteristics of lip movements in defining Italian consonantal visemes. In *Proc. of the International Conference of Phonetic Sciences (ICPhS),* S.Francisco, USA: 881-884.

Massaro, D. W. (1998). Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. MIT Press Cambridge, MA.

Massaro D. W., Cohen, M. M., Beskow, J. & Cole, R. A. (2001). Developing and Evaluating Conversational Agents. In Cassell, J. et al (Eds.), *Embodied Conversational Agents*. Cambridge, MA, MIT Press: 287-318.

Massaro, D. W., Ouni, S., Cohen, M. M. & Clark, R. (2005). A Multilingual Embodied Conversational Agent. In *Proc. of the 38th Annual Hawaii International Conference on System Sciences.* Los Alimitos, CA, IEEE Computer Society Press (Cd-Rom).

Maynard, S. K. (1986). “On back-channel behavior in Japanese and English casual conversation”. *Linguistics,* 24: 1079-1108.

Maynard, S. K. (1987). “Interactional functions of a non-verbal sign. Head movement in Japanese dyadic casual conversation”. *Journal of Pragmatics*, 11(5): 589-606.

McClave, E. Z. (2000). “Linguistic functions of head movements in the context of speech”. *Journal of Pragmatics,* 32: 855-878.

McNeill D. (1992). *Hand and Mind. What gestures reveal about thought.* Chicago, The University of Chicago Press.

McNeill, D. (2000). Growth point, Catchments and Contexts. Cognitive Studies. *Bulletin of the Japanese Cognitive Science Society* 7(1): 22-36.

Morency, L. P. & Darrell, T. (2006) Gestural input: Head gesture recognition in intelligent interfaces: the role of context in improving recognition. In *Proc. of the 11th international conference on Intelligent user interfaces (IUI),* Sydney, Australia: 32-38.

Morency, L., Sidner, C., Lee, C. & Darrell, T. (2005). Contextual Recognition of Head Gestures. In *Proc. of International Conference on Multi-modal Interactions, (ICMI),*Trento, Italy: 18-24.

Morsella, E. & Krauss, R. M. (2004). "The role of gestures in spatial working memory and speech". *American Journal of Psychology*, 117: 411-424.

Mozziconacci, S. (1988). *Speech Variability and Emotion: Production and Perception*. Eindhoven, The Netherlands, Technische Universiteit Eindhoven.

Muller, P. & Prévot, L. A (2003). An empirical study of acknowledgment structures. *Diabruck 7th workshop on the semantics and pragmatics of dialogue*, Saarbrücken, Germany.

Murray, I. R. & Arnott, J. L. (1993). "Towards the Simulation of Emotion in Synthetic Speech: A Review of the Literature of Human Vocal Emotion". *Journal of the Acoustical Society of America,* 93(2): 1097-1198.

Nakano, Y., Reinstein, G., Stocky, T. & Cassell, J. (2003). Towards a Model of Face-to-Face Grounding. *Annual Meeting of the Association for Computational Linguistics,* Sapporo, Japan: 553-561.

Nickerson, R. S. (1976) On conversational interaction with computers. In Baecker, R. M. & Buxton, W.A. (Eds.) *Readings in Human computer interactions.* Los Atos, CA, Morgan Kaufman: 681-693.

Nivre, J. (1999). *Transcription Standard*, version 6.4. Gothenburg, Dept. of Linguistics, Internal Report.

Nivre, J. & Richthoff, U. (1988). Återkoppling och turtagning i två typer av samtal. En jämförande studie av samtal ansikte mot ansikte och telefonsamtal. *Gothenburg Papers in Theoretical Linguistics S10.*

Noor, C. (2004). Empirical evaluation methodology for embodied conversational agents, On conducting evaluation studies. In Ruttkay Zs. & Pelachaud C. (Eds.) *From Brows to Trust Evaluating Embodied Conversational Agents*. Kluwer's Human-Computer Interaction Series 7: 67-99.

Nordstrand, M., Svanfeldt, G., Granström, B. & House, D. (2004). "Measurements of articulatory variation in expressive speech for a set of Swedish vowels". *Speech Communication - Special Issue on Audio Visual Speech Processing*, 1-4(44): 187-196.

Novick, D.G., Hansen, B., Rubesh, K. D. & Ward K. (1996). Coordinating Turn-Taking with Gaze. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, PA: 1888-1891.

Ohala, J. J. (1983). "Cross language use of pitch: an ethological view." *Phonetica*, 40: 1-18.

Okato, Y., Kato, M., Yamamoto, M. & Itashi, S. (1998) System-User Interaction and Response strategy in Spoken Dialogue System. In *Proc. of the International Conference on Spoken Language Processing (ICSLP),* Sydney, Australia, 2: 495-498.

Pandzic, I. S. & Forchheimer, R. (2003). *MPEG-4 Facial Animation: The Standard, Implementation and Applications.* New York, NY, John Wiley & Sons Inc.

Pelachaud, C., Badler, N. & Steedman, M. (1996). "Generating facial expressions in speech". *Cognitive Science* 28 28: 1-46.

Pereira, C. (2000). Dimensions of Emotional Meaning in Speech. *ISCA Workshop on Speech and Emotion*, Belfast, Ireland: 25-28.

Picard, R. (1997). *Affective Computing*. MIT Cambridge.

Poggi, I. (1981). *Le interiezioni: studio del linguaggio e analisi della mente* Boringhieri, Roma.

Poggi, I. & Magno Caldognetto, E. (1996). A score for the analysis of gesture in multi-modal communication. *Workshop on the Integration of Gesture in Language and Speech,* Newark, Delaware and Wilmington.

Poggi, I. & Pelachaud, C, (2000). Performative Facial Expressions in Animated Faces. In Cassel, J., Sullivan, J., Prevost S. & Churchill, E. (Eds.) *Embodied Conversational Agents*. MIT press: 155-187.

Poyatos, F. (2002). *Non-verbal communication across disciplines.* J. Benjamin Publishing Company.

Rajan, S., Craig, S. D., Gholson, B., Person, N. K. & Graesser, A. C. (2001). "AutoTutor: Incorporating back-channel feedback and other human-like conversational behaviors into an intelligent tutoring system". *International Journal of Speech Technologies*, 4: 117-126.

Ruttkay, Zs., Dormann, J. C.& Noot, H.,(2004) ECAs on a Common Ground - A Framework for Design and Evaluation. In Ruttkay, Zs. & Pelachaud, C. (Eds.) *From Brows to Trust Evaluating Embodied Conversational Agents*. Kluwer's Human-Computer Interaction Series 7: 27-66.

Sacks, H., Schegloff, E. A. & Jefferson, G. A. (1974). "Simplest Systematics for the Organization of Turn-Taking for Conversation". *Language,* 50: 696-735.

San-Segundo, R., Montero, J. M., Colàs, J., Guitiérrez, J., Ramos, J. M. & Pardo J. M. (2001) Methodology for Dialogue Design in Telephone-Based Spoken Dialogue Systems: a Spanish Train Information System. In *Proc. of the 2nd Sigidal Workshop on Discourse and Dialogue*: 140-148.

Sanders, G. A. & Scholtz, J. (2000). Measurements and Evaluation of Embodied conversational agents. In Cassel, J., Sullivan, J., Prevost S.& Churchill, E. *Embodied conversational agents*. MIT press: 346-373.

Scherer, K. R. (1981). Speech and Emotional States. In Darby, J. K. (Ed.) *Speech Evaluation in Psychiatry*. New York Grune & Stratton.

Schlegoff, E. (1982). Discourse as an interactional achievement: Some uses of `uh huh' and other things that come between sentences. In Tannen D. *Analyzing Discourse: Text and Talk*. Washington, D.C. Georgetown University Press: 71-93.

Schiffrin, D. (1994). A*pproaches to discourse.* Oxford UK, Blackwell.

Schiffrin, D. (1987). *Discourse Markers.* Cambridge University Press.

Schröder, M. (2001). Emotional Speech Synthesis - A Review. In *Proc. of EUROSPEECH,* Aalborg, Danmark, 1: 561-564.

Shimojima, A., Katagiri, Y., Hanae, K. & Swerts M. (2002). "Informational and dialogue-coordinating functions of prosodic features of Japanese echoic responses". *Speech Communication*, 36(1/2): 113-132.

Sidner, C., L., Kidd, C., D, Lee, C. & Lesh, N. (2004) Where to look. A study of human-robot engagement. In *Proc. of Intelligent User Interfaces*, Portugal: 78-84.

Sjölander, K. & Beskow, J. (2000). WaveSurfer - an open source speech tool. In Yuan, B., Huang, T. & Tang, X. (Eds.), *Proc. of International Conference of Speech and Language Processing (ICSLP)*, Beijing, China: China Military Friendship Publishing, 4: 464-467.

Sjölander, K. & Heldner, M. (2004) Word level precision of the NALIGN automatic segmentation algorithm. In *Proc. of Fonetik,* Stockholm, Sweden: 116-119.

Strassel, S. (2004). *Simple Metadata Annotation Specification* V6.2. Linguistic Data Consortium.

Stubbe, M. (1998). "Are you listening? Cultural influences on the supportive verbal feedback in conversation". *Journal of Pragmatics,* 29: 257-289.

Sumby, W. H. & Pollack, I. (1954) "Visual contribution to speech intelligibility in noise". *Journal of the Acoustical Society of America,* 26: 212-215.

Summerfield, A. Q. (1979) "Use of visual information for phonetic perception". *Phonetica,* 36: 314-331.

Surakka, V. & Hietanen, J. K. (1998). "Facial and Emotional reaction to Duchenne and non-Duchenne smiles". *International Journal of Psychophysiology,* 29: 23-33.

Swerts, M., Koiso, H., Shimojima, A. & Katagiri, Y. (1998). On different functions of repetitive utterances. *International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia: 4:1287-1290.

Takeuchi, A. & Nagao K. (1993). Communicative Facial Displays as a New Conversational Modality. *In, Proc. of the 32nd annual meeting of the Association for Computational Linguistics (INTERCHI),* Las Cruces, New Mexico: 102-109

Tannen, D. (1984). *Conversational Style: Analyzing Talk Among Friends*. Ablex, Norwood, NJ.

Teston, B. (1988). L'observation et l'enregistrement des mouvements dans la parole: Problèmes et methods. In Santi S., et al. *Oralitè et Gestualitè* L'Hartmattan: 1988-1992.

Thórissón, K. R. (1997). Gandalf an embodied humanoid capable of real time multi-modal dialogue with people. *First ACM International Conference of Autonomous Agents,* California.

Thórissón, K. (2002). Natural Turn-taking Needs no Manual: Computational Theory and Model, from Perception to Action. In Granström, B., House, D. & Karlsson, I. *Multi-modality in Language and Speech Systems.* Dordrecht, The Nederlands, Kluwer Academic Publishers: 173-208.

Thórissón, K. R. & Cassell J. (1996). Why put an agent in a human body: the importance of communicative feedback in human-humanoid dialogue. In *Proc. of Lifelike Computer Characters*. Snowbird, Utah: 44-45.

Tomkins, S. S. (1962). *Affect, Imager,. Consciousness*. New York, Springer Publishing.

Traum, D. R. (1994) *A Computational Theory of Grounding in Natural Language Conversation.* Dept. of Computer Science, Rochester University, NY.

Walker, M. A., Litman, D., Kamm, C. A., Abella, A. (1997) PARADISE: A general framework for evaluating spoken dialogue agents. In *Proc. of the 35th Annual Meeting of the Association of Computational Linguistics, ACL/EACL*: 271-280.

Walker, M. A., Kamm, C. A. & Litman, D. J. (2000). "Towards Developing General Models of Usability with PARADISE". *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems,* 6: 271-280.

Wallbott, H. G. & Scherer, K. R. (1986). "Cues and channels in emotion recognition". *Journal of Personality and Social Psychology,* 51: 690-699.

Ward, K. & Heeman P.A. (2000) Acknowledgments in human-computer interaction. *In Proc. of the first conference on North American chapter of the Association for Computational Linguistics*, Seattle, Washington: 280-287.

Ward, N. & Tsukahara W. (2000). "Prosodic features which cue back-channel responses English and Japanese". *Journal of Pragmatics,* 23:1177-1207.

Wiener, N. (1948). *Cybernetics, or control and communication in the animal and the machine.* Cambridge, Massachusetts. The Technology Press, John Wiley & Sons, Inc.

Yehia, H.C., Kuratate T. & Vakiotakis-Bateson, E. (2002). "Linking Facial Animation, Head Motion and Speech Acoustic". *Journal of Phonetics,* 30: 556-568.

Yngve, V. H. (1970). On getting a word in edgewise. *Sixth Regional Meeting of the Chicago Linguistics Society*: 567-577.

Appendix A: Categories and Labels of the Coding Scheme

Table A1 Categories and labels for Speech Acts

SPEECH ACT	
QUESTION	Q
HESITATION	H
STATEMENT	St
FEEDBACK	FB

Table A2 Categories and labels for Verbal and Non-verbal feedback expressions

VERBAL FEEDBACK:	
WORDS	W
PHRASES	Ph
SENTENCES	S
NON-VERBAL FEEDBACK	
FACIAL DISPLAYS	FD
HEAD MOVEMENTS	HM
OTHER	OTHER

Table A3 Categories and labels for facial displays

FACIAL DISPLAYS		
GENERAL FACE	SMILE	Sm
	SCOWL	Sc
	LAUGH	L
EYEBROWS MOVEMENTS	EYEBROW FROWNING	EBFR
	EYEBROW RAISING	RBRA
GAZE DIRECTION	GAZE TOWARDS INTERLOCUTOR	GZTI
	GAZE UP	GZUP
	GAZE DOWN	GZDO
	GAZE SIDEWAYS	GZSW
HEAD MOVEMENTS	SINGLE NOD (DOWN)	S-NOD
	REPEATED NODS (DOWN)	R-NOD
	SINGLE JERK (BACKWARDS UP)	S-JERK
	REPEATED JERKS (BACKWARDS UP)	R-JERK
	SINGLE SLOW BACKWARDS UP	BACKUP
	MOVE FORWARD	FORWARD
	MOVE BACKWARD	BACK
	SINGLE TILT (SIDEWAYS)	S-TILT
	REPEATED TILTS (SIDEWAYS)	R-TILT
	SIDE-TURN	SIDE-TURN
	SHAKE (REPEATED)	SHAKE
	WAGGLE	WAGGLE
	OTHER	OTHER

Table A4 Categories and labels for multi-modal relationship

MULTI-MODAL RELATIONSHIP	
Dependent Complement	DepCmp
Dependent Contradict	DepCnt
Independent	InDep

Table A5 Categories and labels for feedback semantic-pramatic functions

FEEDBACK GIVE	
CONTINUATION **I** GO ON	**FB**Gi**CI**
CONTINUATION **Y**OU GO ON	**FB**Gi**C**Y
ACCEPTANCE	**FB**Gi**A**
NON-**A**CCEPTANCE (**R**EFUSAL)	**FB**Gi**R**
EXPRESSIVE	**FB**Gi**Ex**
FEEDBACK ELICIT	
CHECK **A**TTENTION	**FB**El**ChA**
REQUIRE **A**CCEPTANCE	**FB**El**RA**
MORE INFORMATION	**FB**El**M**

Appendix B: Materials for the Reliability Test

Comments on annotation

The annotation of the four MT dialogues was carried out with Wavesurfer. In this appendix only the transcription and annotation of FEEDBACK is shown for the materials used for the reliability test, which consist of the first 36 contribution of MP-IT Dial 1 and of the contributions 5 to 42 of MP-SW Dial 2.

While in Wavesurfer the annotation of the verbal feedback expressions signalling feedback and the specific semantic-pragmatic function of the identified feedback expressions are shown on two different tiers, here they are shown on the same level.

The labels are shown in angle brackets after the identified **FEEDBACK**, which is emphasised in bold text. Each feedback is numbered successively from 1 to 22. So for instance for the identified feedback expressions *sì* in contribution F12 of the Italian Map Task dialogue here shown, the annotation is: <6;FB;W;Gi;CY>, where 6 is the feedback number in sequence, FB stands for FEEDBACK (the speech act), W stands for word (the type of verbal expression), Gi stands for GIVE (the direction of FEEDBACK) and CY stands for CONTINUATION YOU GO ON (the specific semantic pragmatic function assigned to FEEDBACK).

Italian materials

MP-IT Dial 1
Total number of contributions: 133
Contributions here shown: 1-36 (These 36 contributions include the first 22 FEEDBACK with GIVE direction identified by the expert annotator in her second annotation. This annotation is considered as the "golden standard" in the accuracy test in chapter 5, section 5.2.3).
Scenario: Map-Task between 2 male speakers: G, the giver, and F, the follower.
Original name of the dialogue in the CLIPS database: DGmtA01N.

$G1: sei pronto Paolí?

$F2: certamente!

$G3: kay a+ io suppongo che abbiamo / vabbè / ehm disegni<+> diversi se no sarebbe // non lo [1$F4so]

$F4: e [1$G3vabbè] proviamo a vedere

$G5: **proviamo <1;FB;W;Gi;CI;repetition>//** allora a me / eeh / ti posso di' da dove parte

$F6: **sì vai <2;FB;W;Gi;CY>**

$G7: parte da<+> / alla<+> sinistra di una televisione

$F8: **alla sinistra della televisione** <**3;FB;Ph;Gi;A;repetition**> [1$G9 **vai**] **<4;FB;W;Gi;A>**

$G9: [1$F8 che] **tu non hai televisioni? <FB;S;El;RA>**[34]

$F10: **no, no, ce l'ho la [1G$11 televisione] <5;FB;S;Gi;A>**

$G11: [$F10 **ah!] <6;FB;W;Gi;Ex>** poi eeh passa da sotto, diciamo

$F12: **sì <7;FB;W;Gi;CY>/**

$G13: sale un pò come una curva diciamo<+> {u}na campana e scende al di sotto della<+> torta//

[34] This is a FEEDBACK with ELICIT direction, which has not been considered in the reliability test.

$F14: **al di sotto della torta <8;FB;Ph;Gi;A; repetition**>

$G15: e ci gira in senso [1$F16 antiorario]

$F16: [1$G15 **eh**] **<9;FB;S;Gi;CI>** aspetta un momentino perché non è così semplice come può sembrare /allora scende giù / fa una curva / tipo [1$G17 campana]

$G17: [1$F16 **curva] / in alto**, **sì <10;FB;Ph;Gi;CI;repetition>,** insomma , può anche non farla / voglio dire /comunque la<+> la fa [1$F18 sale]

$F18: [1$G17 curva **eeh aspetta] un attimo <11;FB;S;Gi;CI>** Fabri' ragiona / se parte da / sinistra della televisione come fa a farci una curva in alto se va da sotto?

$G19: ci passa / sotto la televisione passa / una [1$F20 volta passata] sotto la televisione

$F20: [1$G19 **sì**] **<12;FB;W;Gi;CY>** [whispering]# /

$G21: sale /

$F22: **sì <13;FB;W;Gi;CY**>

$G23: la curva

$F24: **ah okay! <14;FB;W;Gi;Ex>**
[1$G25sale]<15;FB;W;Gi;A;repetition>

$G25: [1$**F24 mh]<16;FB;W;Gi;CY)** /

$F26: **poi? <FB;W;El;M>**[35]

$G27: e poi scende fino a<+>/ a passare sotto la torta

$F28: **ottimo <17;FB;W;Gi;A>** così/ **vai <18;FB:W;Gi;A>**

$G29: e ci // e ci gira in senso antiorario

[35] This is a FEEDBACK with ELICIT direction, which has not been considered in the reliability test.

$F30: **ci gira in senso / antiorario <19;FB;W;Gi;A;repetition>** [1$G31 **sì] <20;FB;W;Gi;A>**

$G31: [1$F30 e poi prende] diciamo così / una<+> / una<+> via verso sinistra / che è orizzontale /

$F32: **sì <21;FB;W;Gi;CY>/**

$G33: e in ultimo tratto eeh si alza / e va a girare in senso /orario attorno alla macchina / da sinistra verso destra / la aggira /

$F34: attorno alla macchina rossa o quella blu? **<FB;S;El;M>**

$G35: ah! eh! eeh / la macchina rossa , diciamo , quella che sta più a sinistra

$F36: **si <22;FB;W;Gi;A>**

Swedish materials

MP-SW Dial 2
Total number of contributions: 246
Contributions here shown: 5-42. These 37 contributions include the first 22 FEEDBACK with GIVE direction identified by the expert annotator in her second annotation. This annotation is considered as the "golden standard" in the accuracy test in chapter 5, section 5.2.3. (The first 4 contributions are not shown since they include the personal information about the two dialogue participants).
Scenario: Map-Task between 2 female speakers: G, the giver, and F, the follower.
Original name of the dialogue Ckgt01rl

$G5: okej nu så ska vi börja eeh på kartan där du ser ett ankare

$F5: eeh **jaha <1;FB;W;Gi;CY>**

$G6: det är rätt långt ner i norr

$F7: i sydvästra hörnet

$G8: nor..alltså nord nordväst

$F9: **nordväst? <FB;W;El;RA>**

$G10: **sydväst< 2**;**FB;W;Gi;A>**

$F11: **aha just de{t} <3;FB;Ph;Gi;CI>**sydvästra det är nå{go}n bukt där

$G12: **[1**$F**13 ja// ok] <4FB;W;El;RA>**

$F13: [1$G12 **ja // okej] <5;FB;W;Gi;A>**

$G14: så om du bör börjar vid det där vid den nedre delen av ankaret

$F15: **ja <6;FB;W;Gi;CY>**

$G16: och sen ska du då gå eeh aningen nordöst upp så ska du uppåt land

$F17: **ja <7;FB;W;Gi;CY>**

$G18: eeh nu går lite öst först

$F19: hur långt upp då?

$G20: **ja** <**8;FB;W;Gi;CI>**vänta om du om du går eeh börjar å gå till öster en liten bit så att du kommer inåt land aningen en

$F21: **mm** <**9;FB;W;Gi;CY>**

$G22: en centimeter ungefär och sen börjar du gå i en böjd eeh riktning norrut eeh och du måste undvika då en nå{go}t fågelliknande djur /

$F23: jag har en säl på stranden där på

$G24: **ja** <**10;FB;W;Gi;CI>**det kanske är en säl

$F25: i bukten

$G26: **ja** <**11;FB;W;Gi;CI>**det är det vad ja den ska du gå utanför så att säga /

$F27: **mm <12;FB;W;Gi;CY>**

$G28: och sen följer du då den här bukt / alltså du håller din den väg du går på land /

$F29: **mm** <**13;FB;W;Gi;CY>**

$G30: följer buktens kurva [1$F31kan man väl säga]

$F31: [1$G30**mhmm**] <**14;FB;W;Gi;Ex>**

$G32: men då håller dig i mitten på där du ser att till hö.. till höger om dig

$F33: **mm** <**15;FB;W;Gi;CY>**

$G34: så finns det en en å eller nå{go}nting så{da}nt

$F35: **ja** <**16;FB;W;Gi;A>**

$G36: så du håller dig emellan den å eeh konturen och buktens kontur

$F37: **mm** <**17;FB;W;Gi;CI>**hur gör jag med krabborna där då

$G38: **ja** {+}<**18;FB;W;Gi;CI>** då du ska hålla dig utanför dom

$F39: **{j}a** {+}<**19;FB;W;Gi;CI>** men det kan jag inte för dom är hela vägen ända fram till ån där

$G40: **jaha**<**20;FB;W;Gi;Ex>//** [1$F41 **okej men**] <**21;FB;W;Gi;CI>**

$F41: [1$G40 jag får kliva över dom]

$G42: ja då får du helt enkelt kliva över dom **<22;FB;S;Gi;A;repetition>**

Appendix C: Example of FEEDBACK annotation

Comment on the annotation

The annotation of the ten PF-Star Dialogues was carried out with by using Wavesurfer. The focus of the annotation was on the production of head movements related to specific semantic pragmatic functions (see chapter 8, section 8.3.2).

In this appendix only the transcription and annotation of FACIAL DISPLAYS related to FEEDBACK is shown for one of the PF-Star dialogues. While in Wavesurfer the annotation of the verbal and non-verbal expressions co-occurring to signal feedback and the specific semantic-pragmatic function of the identified feedback expressions are shown on three different tiers, here they are shown on the same level.

The labels are shown in angle brackets after the identified **FEEDBACK**, which is emphasised in bold text.

So for instance for the identified feedback expressions *{j}a precis* in contribution $S12 of the dialogue shown in this appendix, the labels are </FB;Ph;R-Nod;Gi;CY>, where FB stands for FEEDBACK (the speech act), Ph stands for PHRASE (the type of verbal expression), R-Nod stands for REPEATED-NOD (the type of non-verbal expressions, in particular FACIAL DISPLAYS), Gi stands for GIVE (the direction of feedback) and CY stands for CONTINUATION YOU GO ON (the specific semantic pragmatic function of FEEDBACK).

Sample of an annotated dialogue

PF-Star dialogue 4
Number of contributions: 36
Scenario: Subject S is buying an all-inclusive trip to Italy, Subject M is the travel agent that helps him.
Original name of the dialogue: 3003

$M1: hej

$S2: ja hej hejsan det är jag som är Johan Andersson jag ringde tidigare angående

$M3: **{j}a just det <FB;Ph;Gi;A>**

$S4: **ja resa till Italien <FB;Ph;FD,R-NOD;Gi;A>**

$M5: angående till Italien

$S6: **ja <FB;W;S-Nod;Gi;CY>,**

$M7: **ja precis<FB;W;S-Nod;Gi;A)**

$S:8: du skulle titta på några paket för mig och min fru

$M9: **ja precis <FB;Ph;EBRa;CI)**

$S10 **mm <FB;W;S-Nod;Gi;CY>**

$M12 det var lite olika alternativ där med all-inclusive och ehm

$S13 **{j}a precis <FB;W;S-Nod;Gi;CY>**

$M14 **{j}a <FB;W;S-Nod;Gi;CI>//** och det beror lite grann på vilken // vad ni kommer att välja då om ni vill åka till Torino) eller om ni vill åka ner kanske till södra delen mot Rom

$S15: **{j}a <FB;W;R-Nod;Gi;A>**vi var lite sugna på på Rom precis och det

$M16: **{j}a just det <FB;Ph;R-Nod;Gi;A>**

$S17: och det så det enda krav vi har är att vi menar // att man bor hygglig centralt något rökfritt ställe självklart /

$M18: **{j}a<FB;W;S-Nod;Gi;CY>**

$S19: eftersom min fru är allergisk

$M20: **jaha {j}a naturligtvis <FB;Ph;R-Nod;Gi;A>**

$S21: **mh <FB;W;S-Nod;Gi;CY>**ehmm och och luftkonditionering /<FB;R-Nod;El;RA>

$M22: **{j}a <FB;Ph;S-Nod;Gi;A>** ju det det är standard på på dom flesta

$S23: **ja <FB;W;Gi;A>**

$M24: så det är så vi har det

$S25: ja men vad har du för alternativ då

$M26: **ja[+] <FB;W;S-Nod;Gi;CI>** / /det finns den här tio dagarsresan då/ **<FB;R-Nod;El;RA>**

$S27 **mm <FB;W;S-Nod;Gi;CY>**

$M28: som / den kostar åttatusen per person

$S29: {**j}a just de{t}<FB;W;Gi;A>**

$M30: och /

$S31: och det ingick utflykter med den eller?

$M32: **{j}a precis <FB;Ph;R-Nod;Gi;A>**

$S33: ah för det lät intressant faktiskt <FB;R-Nod; El;RA>

$M34: **{j}a,ja visst <FB;Ph;R-Nod;Gi;CI>** och det det är guider hela veckan också som man har tillgång till

$S35: det låter jättebra nä men alltså då tar vi den