UNIVERSITY OF TRENTO - Italy

CIMeC

Center for Brain/Mind Studies

Doctoral School in Cognitive and Brain Sciences

Language, Interaction & Computation

Doctoral Thesis Title:

Computational Modeling of (un)Cooperation:

The Role of Emotions

Candidate:

Federica Cavicchio

Supervisor:

Prof. Massimo Poesio

Computational Modeling of (un)Cooperation:

The Role of Emotions

DOCTORAL THESIS

To obtain the degree of Doctor at the University of Trento

on account of the decision of the graduation committee

to be publicly defended on January 18th, 2010

by

Federica Cavicchio

Supervisor:

Prof. Massimo Poesio

Acknowledgments

Humani nil a me alienum

Terenzio, *Heautontimorumenos*, v.77

After this three-year journey, I'd like to thank the people who I met along this experience and made it special. First of all, I wish to thank my tutor, supervisor and occasionally mentor Massimo Poesio for his support and encouragement during this research. I wish to thank Marco Baroni for his comments on the statistics and for his useful math and statistic lessons. Thanks also to Francesco Vespignani for his help in setting up the psychophysiological recordings and his useful insights on the world of physiology. Thanks to Prof. Valentina D'Urso for introducing me to the wonderful world of emotion studies, for her friendship and humanity and for her hints on the experimental design of this work. Also, thanks to Prof. Emanuela Magno Caldognetto for the nice and useful discussions on communication multimodality. I wish to thank Emanuela also for introducing me, several years ago, to the exciting world of research, for the chats on life and boyfriends and family and above all for the book crossing. I would never thank her enough for introducing me to Ian McEwan's and Cormac McCarty's books. Thanks to my family for loving and supporting me, though they do not totally grasp why at the age of 31 I'm still studying. Thanks to my beloved Fernando for his love and tolerance of my ups and downs. A huge hug to all my friends in Rovereto, my dearest Sara, my two office mates Luigi and Federico and to my flatmate Stella. Also, thanks to Stella for her help in improving my English and all the nice chitchats on life and work. Finally, thanks to all my experiment participants. I hope you have forgiven me for getting you upset…

Abstract

The philosopher H. P. Grice was the first to highlight the extent to which our ability to communicate effectively depends on speakers acting cooperatively. This tendency to cooperation in language use, recognized since Grice's William James lectures, has been a key tenet of subsequent theorizing in pragmatics. Yet it's also clear that there are limits to the extent to which people cooperate: theoretical and empirical studies of the Prisoner's Dilemma have shown that people prefer to cooperate if the other party cooperates, but not otherwise. This would suggest that in language use, as well, the level of cooperation depends on the other person's cooperativeness. So far, however, it has proven remarkably difficult to test such prediction, because it is difficult to analyze cooperation and communicative style objectively, and the schemes proposed so far for, e.g., non-verbal cues to cooperation tend to have low reliability. In this study the existence of a negative correlation between emotions and linguistic cooperation is demonstrated for the first time, thanks to newly developed methods for analyzing cooperation and facial expressions. The heart rate and facial expressions of the participants in a cooperative task were recorded after uses of cooperative and uncooperative language; facial expressions and the level of linguistic cooperation in each utterance were classified with high reliability. As predicted, very high negative correlations were observed between heart rate and cooperation, and the facial expressions were found to be highly predictive of her level of cooperation. Our results shed light on a crucial aspect of communication, and our methods may be usable to research in other aspects of human interaction as well.

Keywords: cooperation, pragmatics, emotion theory, computational modelling, multimodal communication, facial expressions.

Contents

Chapter 1 Introduction……………………………………………………………….…….….1

1.1 Cooperation and Emotion in Dialogues……………………………………….1

1.2 Challenges…………………………………………………………….………4

1.2.1 Challenges in Data Acquisition and Annotations……………………..…5

1.2.2 Challenges in Reliability Studies…………………………….…….…...6

1.3 Goals and Research Questions…………………………………….…………8

1.4 Outline………………………………………………………………………….10

Chapter 2 Background…………………………………………………………….….…12

2.1 Cooperation in Dialogues……………………………………………….….12

2.1.1 A Computational View on Cooperation………………………………15

2.1.2 Uncooperation in Dialogues……………………………………….…...25

2.2 Emotion Studies: Categorical *vs* Appraisal Theories……………………...28

2.3 Emotion Annotation Reliability: Percent Agreement, Kappa and α …………..34

2.3.1 Measuring Category Judgment Performance in Non verbal Communication Studies: Hit Rate, Index of Accuracy and Unbiased Hit Rate…………………………………………………………………….......43

2.4 Summary and Conclusion……………………………………………………48

Chapter 3 The Rovereto Emotion and Cooperation Corpus (RECC): A New Resource to Investigate Cooperation and Emotion……………………………….....52

3.1 Multimodal Corpora collection and validation - Problem statement………..52

3.2 The RECC Corpus Design……………………………………………….....56

3.3 Recording and eliciting procedure……………………………………….....61

3.4 Corpus transcription…………………………………………………….…..66

3.5 Conclusions………………………………………………………………....69

Chapter 4 The RECC Coding Scheme and its Validation…………………………..71

4.1 Emotion Coding Scheme…………………………………………………...71

4.2 RECC Cooperation Coding Scheme ……………………………………….77

4.3 Turn Taking and Gaze Direction Annotation…………………………….....81

4.3.1 Turn Management Annotation Scheme………………………………...81

4.3.2 Gaze Direction Annotation……………………………………………85

4.4 RECC Coding Scheme Reliability…………………………………………..86

4.4.1. Annotation Procedure…………………………………………………..87

4.4.2. Kappa Statistic Results………………………………………………94

4.5 Corpus Public Releases………………………………………………………95

4.6 Conclusion……………………………………………………………………..96

Chapter 5 The Predictors of (un)Cooperation……………………………………….97

5.1 Emotion and Cooperation…………………………………………………97

5.1.1 Heart Rate and Cooperation………………………………………..97

5.1.2 Facial Expressions and Cooperation…………………………………99

5.2 Gaze, communicative role, gender and cooperation……………………….102

5.2.1 Gaze and cooperation……………………………………………...102

5.2.2 Conversation Role, Gender and Cooperation……………………...104

5.3 Discussion and Conclusion………………………………………………..107

Chapter 6 Conclusion…………………………………………………………………..111

Bibliography……………………………………………………………………………...115

Appendix…………………………………………………………………………………..132

Chapter 1

Introduction

1.1 Cooperation and Emotion in Dialogues

The concept of cooperation in pragmatics can be traced back to H. P. Grice's William James lectures. He stated that felicitous conversations rest on a Principle of Cooperation requiring speakers to '*make conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange*'(Grice, 1975: 45). However, there is a general difficulty in pinning down the term cooperation. Many authors assume that people are very good at cooperating to *reduce misunderstanding* in the process of communication. Therefore, cooperation is often assumed to be the equivalent of being totally explicit, which in turn leads to a folklinguistic notion of the term cooperation (Davies, 2006). Indeed linguists often describe how good people are at cooperating and how natural it is. However, as Coupland, Giles and Wiemann (1991) pointed out, linguists always describe how good humans are at language without explicitly discuss that people (and occasionally themselves) sometimes fail to correctly interpret something or make poor linguistic choices. Grice’s use of the term cooperation is rooted in rationalism. His work comes from a philosophical tradition based on intuition and reflection and it was never explored by its originator in an empirical framework. As a consequence, any interpretation of Grice’s Principle of Cooperation can be arguable. However, clearly there are limits to the extent to which people cooperate. Grice himself knew that people do not *always* follow cooperation maxims as they communicate, and he identified four ways in which discourse participants regularly break, or fail to fulfill, maxims in conversation: violating, opting out, clashing, and flouting (Grice, 1989).

What happens, though, when cooperation breaks down? In an interaction, a speaker’s level of cooperation seems to depend on the other speaker’s cooperativeness. Classic game theory has shown that people prefer to cooperate if the other party cooperates, but not otherwise. Besides, emotions are found to be important predictors of cooperation. Specifically, negative emotions are positively correlated with the tendency to not cooperate. For instance, in a novel experiment Sanfey, Rilling, Aronson, Nystrom and Cohen (2003) used functional magnetic resonance imaging to monitor the brain activity of responders while playing economic games. They found that after receiving unfair proposals, participants having a greater activation of negative emotion brain areas were more likely to not cooperate. On the other hand, participants receiving the same unfair proposal but having a stronger activation of brain areas linked to problem solving and cognitive conflict were more likely to cooperate. So far it has proven remarkably difficult to measure the effects of cooperation and non-cooperation on other dialogue partners through, e.g., an analysis of the partner's communicative style.

In order to investigate whether emotions are predictors of cooperation in dialogues we design the data collection - i.e. the corpus - eliciting negative emotions and expecting a decrease of cooperation in the speaker. The interactions have been audio and video taped and the psychopsysiological data have been collected and aligned with the audiovisual ones. Beside linguistic aspects of interaction, we take into account other modalities of face-to-face communication such as gaze direction and facial expressions. Both of them are useful to study cooperation and emotions. The decision of using emotion to elicit uncooperative behavior raised a number of questions regarding the nature of emotions. The theoretical concept of emotions we adopted is based on component appraisal. We will investigate physiological and motivational changes affecting the autonomic nervous system (e. g., cardiovascular and skin conductance changes) and the somatic nervous system – changes in motor expression in face, voice, and body. The way these different modalities interact with each other to convey an emotion was the focus of investigation since Mehrabian (1971). He investigated the importance of verbal and non verbal messages in the expression of feelings and attitudes. He stated that words, vocal expressions and facial expressions have a different weight in determining an emotion. He quantified that the weight of words was 7%, the one of vocal expression was 38% and the facial expressions was 55%. This rule is particularly valid when verbal and non verbal communications are not congruent such as when a speaker verbally says "I*t's OK, I don't mind*" but non-verbally avoids eye contact, looks anxious etc. (Truong, 2009). Subsequently, the receiver of the message is more likely to trust the non verbal message. Voice realizations, in particular affect burst and laugh, can lead to incongruity as well. Indeed, one can laugh to alleviate a stressful situation. As a consequence, voice production alone can be misleading as it is not possible to disentangle the difference between a real and a fake laugh.

Emotion can be measured and expressed in different modalities. In this work we will focus on psychophysiological data and facial expressions. Physiological measures are the most reliable signal of arousal and when combined with the pleasant/unpleasant assessment of the situation, they can give us a general indication of emotion (Scherer, 1993, 2009). Facial expressions and gaze are likely to express emotions and other communicative functions transcending simple indications of one’s current feelings (Scherer, 1992). Cooperation and emotion will be investigated in the context of higher order planning. The method we will adopt is dialogue annotation and annotation reliability assessment- i.e. we will investigate whether other annotators would agree with the researcher's assessment. Moreover, as we are interested in finding out facial predictors of cooperative and uncooperative behaviour, we will investigate whether mutual gaze predicts cooperation. Hopefully, our findings will be taken into account when designing the future Human Computer Interfaces.

1.2 Challenges

In this section we identify the challenges involved in collecting a multimodal corpus with the aim of investigating cooperation and emotions. The challenges are divided into two main groups: data acquisition and annotation and reliability of annotation studies.

1.2.1 Challenges in Data Acquisition and Annotations

In the last years, many corpora have been collected and annotated with the aim of studying emotions in daily interactions. A large number of these emotion oriented corpora is multimodal, that is to say they incorporate the recordings and annotations of several communication modalities such as speech, hand gesture, facial expression, body posture, etc. In many of these corpora emotive expressions are produced by expert or semi-expert actors. It is often taken for granted that these expressions are the "gold standard" to study facial display of emotions (Ekman, 1992). This is not completely true; as Wagner (1993) pointed out, each actor's production should be validated when assessing the real closeness to the "standard" emotion representation that a group of judges has in mind. Another method to collect emotion corpora is using task-oriented games, such as the Map Task (Anderson et al., 1991) or Multiparty dialogues (Carletta, 2007). These corpora are mainly focused on face-to-face verbal (and non verbal) interactions. Even if task-oriented and Multiparty corpora are specifically produced to analyze linguistic features, non verbal behavior (such as gaze, gesture and even emotion displays) has also an important role. Ecological corpora are usually recorded from TV shows, news and interviews and feature a wide range of verbal and non verbal behaviors (Douglas Cowie et al., 2005). In a good number of cases the resulting data are difficult to classify and analyze (Martin, Caridakis, Devillers, Karpouzis & Abrilian, 2006).

As our main interest is investigating the relationship between cooperation and emotion, adopting a task-oriented methodology is quite straightforward. Task-oriented methods are likely to elicit unscripted dialogues and at the same time to allow control of the context. One of the main issues for researchers interested in conversational analysis is the problem of the Observer's Paradox (Labov, 1972). Although "casual" conversation analysis is often seen as the "gold standard" in pragmatics, there is an increasing interest in other sort of dialogues and methodologies (Bargiela-Chiappini & Harris, 1997; Connor & Upton, 2004),

Our main objective is how participants manage, transfer and negotiate information in order to achieve a common goal. There are interaction aspects to this – not answering questions or directly refusing the suggestions of the interaction partner would probably lead to a partial or a total breakdown of cooperation in the conversation. Elements as the above are carefully considered in our analysis, since our primary interest is the interpersonal aspects of talk. Another important aspect in our corpus collection is the emotion elicitation method and the psychophysiological data recordings. One can argue that psychophysiological measures such as heart rate and skin conductance require more effort and are usually intrusive for the speakers. Therefore, we used wearable measuring equipments in order to reduce both the amount of effort and the obtrusiveness.

1.2.2 Challenges in Reliability Studies

The large use of corpora in linguistics and engineering has raised questions on coding scheme reliability. Collecting a multimodal audiovisual corpus may be computationally demanding but its analysis is far more challenging. The collection of such a large amount of multimodal data has raised a debate on corpora analysis and consequently on schemes to code multimodal communication and the scheme reliability. When the concept of dialogue coding system was introduced, most people assumed that its concern was the identification and labeling of overall dialogue structures (e.g. Carletta et al. 1997; Houghton & Isard 1987; Kowtko et al. 1992; Sinclair & Coulthard 1975) or of structures within a dialogue (e.g. Conversation Analysis). Beyond this, in our opinion coding schemes established a way to identify the presence of certain types of discourse strategies – such as cooperation in our case. The aim of testing coding scheme reliability is to assess whether a scheme is able to capture observable reality and eventually allow some generalizations. Multimodal coding schemes are mainly focused on dialogues (dialogue acts, topic segmentation) but recently the "emotional area" has started to be included into annotations. At the moment, the weakest point of multimodal studies and in particular, of multimodal studies of emotions is the lack of annotation scheme reliability. This could be due to the nature of emotion data. Indeed, annotation of mental and emotional states is a very demanding task. Moreover, context is inescapably linked to modality. For example, facial expressions are subject to a wide range of influences. These include culture-specific display rules and interactions with speech, involving both the lips and the eyebrows. On the other hand, the low coding scheme reliability can be due to the nature of Kappa[1]

[1] The literature is full of terminological inconsistencies. Carletta (1996) called the coefficient of agreement she argues for "kappa," referring to Krippendorff (1980) and Siegel and Castellan (1988), and using Siegel and Castellan's terminology and definitions. However, Siegel

statistic - which is, basically, the standard statistic performed to assess coding scheme reliability. Kappa requires annotation categories to be clearly separable of each other. As a consequence of the poor results found so far in categorical emotion annotation, we do not analyze emotions using categorical schemes. Having the opportunity to record the psychophysiological data, emotion typology was investigated trough facial expressions (Scherer & Ceschi, 2000). We also investigated gaze direction and turn management

1.3 Goals and Research Questions

In our study, uncooperative communication and cooperation was investigated by recording cooperative and uncooperative dialogues between subjects. Previous studies have shown that visual access to each other's non verbal behavior fosters a dyadic state of rapport that facilitates mutual cooperation (Argyle, 1990*;* Tickle-Degnen & Rosenthal, 1990) stablishing whether facial cues are predictive of cooperation. Our hypothesis is that the negative emotion elicitation would lead to a reduced level of cooperation in the other participant. To test this, the level of cooperation of the utterances following the elicitation was measured. To investigate the relationship between cooperation and emotions in dialogues,

and Castellan's statistic, which they called K, was actually Fleiss's generalization to more than two coders of Scott's *π* (1955), not of the original Cohen's *k* (1960). Fleiss (1971) proposed a coefficient of agreement for multiple coders and called it *k,* even though it calculates expected agreement based on the cumulative distribution of judgments by all coders and is thus better thought of as a generalization of Scott's *π*. This unfortunate choice of name was the cause of much confusion in subsequent literature: Often, studies which claimed to give a generalization of *k* to more than two coders actually report Fleiss's coefficient (e.g., Bartko & Carpenter, 1976; Siegel & Castellan, 1988; Di Eugenio & Glass, 2004). Since Carletta introduced reliability to the Computational Linguistics community based on the definitions of Siegel and Castellan, the term "kappa" has been usually associated in this community with Siegel and Castellan's K, which is in effect Fleiss's coefficient, that is, a generalization of Scott's *π*. To confuse matters further, Siegel and Castellan used the Greek letter *k* to indicate the parameter which is estimated by K. In what follows, I call Kappa the coefficient discussed by Siegel and Castellan that I used to calculate reliability in this work.

we collected a new corpus: the Rovereto Emotion and Cooperation Corpus (RECC). Following the appraisal theory of emotion, this corpus will allow us to have a broader view of emotions in a dialogue setting based on psychophysiological recordings, the assessment of the situation pleasantness/unpleasantness, and the corresponding facial expressions. A further condition was added: in half of the interactions a screen divided the two speakers, so that they could not see each other's face. The psychophysiological measures and the facial expressions were used as predictive variables of cooperation. Therefore, our hypothesis was translated into the two following research questions:

- **research question 1:** *Are psychophysiological measures, specifically heart rate, predictors of cooperation?*
- **research question 2:** *Is facial expression a predictor of cooperation?*

As in half of the interactions were without eye contact, an additional question addressed was the following:

- **research question 3:** *Is eye contact a predictor of cooperation?*

Based on the sociolinguistic perspective, we examined two aspects of cooperation: dominance and gender. Research on dominance is basically focused on how asymmetry of knowledge held by the two speakers reflects on cooperation (Markova & Foppa, 1991; Drew & Heritage, 1992). Another important variable leading to differences in cooperative behavior is gender. In many studies on both western and non-western cultures, women are found more cooperative and men more competitive. A fourth and fifth research question was explored:

- **research question 4:** *Does dominance affect cooperation?*
- **research question 5:** *Is gender a predictor of cooperation?*

1.4 Outline

In the next chapters, we address the research questions previously mentioned. Firstly, in Chapter 2 we gave an overview of the previous studies on cooperation and on emotions and further on the attempts made so far to investigate them with reliable techniques.

In Chapter 3, we described how we collected RECC and how we overcame the limits confronted in previous multimodal database collecting both audiovisual and psychophysiological data. In Chapter 4, RECC coding scheme reliability was validated via Kappa statistics. We reported and discussed the scores of each label we used to annotate facial expressions and cooperation features along with the turn taking and gaze ones. The RECC coding scheme can aid to explore how different emotive sets (positive or negative) modify cooperation; how turn management and sequencing strategies are expressed; how gaze can enhance or disrupt cooperation; how emotions modify the multimodal communicative channels.

In Chapter 5, we addressed our five research questions. A linear regression and a logistic regression model (i. e. a particular flavor of general linear model) were run. The results were discussed in the light of componential appraisal theory of emotions and other communicative functions of facial expressions and gaze. Finally, in Chapter 6 we drew conclusions from the experiments performed and discussed them in the light of the research questions.

Chapter 2

Background

2.1 Cooperation in Dialogues

Cooperation is a central concept in pragmatics and dialogue studies. Many observations led to the belief that humans have a pan-specific pattern of interaction marked by specific rules separable from language (Levinson, 1995, 2006). Interactions are by and large cooperative. The concept of cooperation in pragmatics can be traced back to H. P. Grice's William James lectures. He stated that felicitous conversations rest on a Principle of Cooperation requiring speakers to '*make conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange*' (Grice, 1975: 45). Grice generalized the Cooperative Principle from the four *maxims* he found that discourse participants seem to follow: *Quantity* (give as much information as required, and no more than that), *Quality* (do not say what is false or that for which you lack adequate evidence), *Relation* (be relevant), and *Manner* (be clear, orderly and avoid ambiguity) (Grice, 1989: 28). Grice’s proposals have proved to be enormously influential, and Grice’s Principle of Cooperation has been a key tenet of subsequent theorizing in pragmatics (Allen

& Perrault, 1980; Cohen, Morgan & Pollack, 1990; Clark & Wilkes-Gibbs, 1986; Clark, 1996). There is a number of accounts and interpretations of dialogue performance that requires some notion of cooperation or collaboration as part of the explanatory mechanism of communication. For instance, Searle used general principles of cooperation in a conversation to account for indirect speech acts and implicatures (Searle, 1979). Levinson referred to an implicature relying on '*some very general expectation of interactional cooperation*' (1983: 50). More recently Levinson (2006: 48) claimed that there is an asymmetry between what are called *preferred response* and *dispreferred responses*. His main point is that interaction is biased towards a cooperative direction because the interaction system is set up so that it is just easier to comply with requests or accept invitations than decline them. This remark is probably a fair reflection of the difficulty in pinning down the term cooperation: it does nothing to signal to the reader that there is a difference between an everyday usage of the term cooperation - the 'folklinguistics notion' of the term (Davies, 2006: 30) - and the Gricean use of it, rooted in rationalism. Grice's work comes from a philosophical tradition based on intuition and reflection, and was never explored by its originator in an empirical framework. Therefore, testing the Cooperation Principle is not straightforward. Grice's primary focus was on rationality and communication efficiency, so dialogue is interpreted as a rational and efficient activity. Davies argued that the Cooperation Principle can be "translated" in the following norms (2006: 51):

1.Speakers will avoid unnecessary effort:

Although speakers should have a moral commitment to doing the work necessary to the task, they are not expected to do any *more* than that. The Cooperation Principle makes it possible for speakers to decrease their effort, and thus meet this ideal.

2.Speakers will improve at tasks:

Speakers should have the ability to *learn*. This could be seen as the application of reason (i.e. rationality) to a particular problem set. In terms of task-oriented dialogues, we would expect speakers to produce better task results over time. This can also be linked to notions of efficiency: the agent learns the minimum that is required to do the task.

3.Speaker effort will decrease:

This hypothesis is linked to the previous two. As speakers learn, they will determine what effort is absolutely necessary to the task, and what is extra. They can then adjust their behavior accordingly. Therefore, they can minimize their effort for the task.

Summarizing, Grice Cooperation Principle states that successful dialogues rely on effort saving strategy and in the idea of a general decrease of the total amount of effort needed in that dialogue during time. This interpretation of Grice's Principle of Cooperation can be arguable. Nevertheless, Davies' primary intention was to demonstrate the clear distinction between Gricean Cooperation and other folklinguistic notion of the term cooperation rather than produce an inarguable interpretation of Grice.

2.1.1 A Computational View on Cooperation

Two alternative views on cooperation are given by Clark and Wilkes-Gibbs (1986) and Shadbolt (1984). In their computational view, Clark and Wilkes-Gibbs claimed that each utterance should be considered as a presentation (Brennan & Clark, 1996; Clark & Brennan, 1991; Clark & Krych, 2004; Clark & Schaefer, 1987; Schober & Clark, 1989) to be accepted by the addressee before it can be deemed to be added to the speakers' common ground. Common ground is the open stockpile of assumptions shared by interacting speakers and fueling implicative inference (Grice, 1989; Levinson, 1995, 2000; Enfield, 2006). As common ground increases, speakers can be less explicit because a certain level of shared knowledge is being assumed. Clark and Wilkes-Gibbs found out that when accomplishing a task speakers used shorter referring expressions and took fewer turns as long as the common ground increased. This decrease in words and turns was interpreted as a decrease in effort: speakers said less because they saw the opportunity to conserve effort. So, as collaboration is a joint effort, Clark et al. argued that the minimization of effort is a joint activity, hence the *Principle of Least Collaborative Effort*. On the other hand, Shadbolt claimed that communication is based on a *Principle of Parsimony*. In Shadbolt's view, a speaker can choose between two conversational risks: the low or the high. The high risk approach makes the assumption that speakers share knowledge before starting the interaction. If it is not the case, there will be the risk that a potentially effortful repair sequence must be take place. On the contrary, the low risk approach takes more effort initially, as it settles down a larger common ground but is more likely to succeed at the first attempt. The trade-off between the two strategies is the opportunity to save some effort against the possibility of having to engage in a potentially more effortful repair sequence. So, the risk-effort trade-off is the judgment that the speaker makes in terms of the likelihood of a particular risk being worthwhile (for a computational application of this model see Carletta, 1992, Carletta & Mellish, 1996). In other words, a speaker will try to choose the approach which will be the least effortful – and therefore the most risky – that is still likely to succeed. Therefore, a speaker during the interaction computes for each utterance the best risk-effort trade off – though facing the risk of receiving an inadequate pay-off. The Principle of Parsimony is translated with the following norms (Davies, 2006: 49):

1- Risks would be taken – some failures should be encountered.

If speakers are trying to work out where the best trade-off occurs, then they are bound to take risks which do not pay off. Otherwise they would never find out which actions are necessary and which aren't. Such risky behavior is bound to lead to some task failures.

1a. Speakers with equal commitment (whether high or low) should be associated with more task success: where the commitment level of participants is mismatched, the needs of the participants (particularly those with more commitment) will not be met, which leads to less effective dialogue. Thus a poorer task will result.

1b. There is no relationship between increased collaboration and task success.

2- The need for equal commitment takes precedence over the input of individuals: the effort of one cannot replace the lack of effort by another.

3- Risks would decrease over time – fewer failures: as speakers work out what behavior is acceptably risky, and what behavior isn't, then we would expect the *bad risks* to decrease.

4- Task success would improve as speakers negotiate trade-off more successfully over time: if speakers work out the best point on the risk-effort trade-off, then their failure rate should also be minimized and they should produce better task results.

5- Behavior would modify as speakers try out different risk-effort combinations, and eventually settle on a set of useful combinations: speakers should try out various strategies until they find one which satisfies their constraints. This also makes the assumption that the behavior found later in the task (as participants gain experience) would better represent their 'best-fit' on the risk-effort scale. It should be noted that the risk-effort approach would suggest that speakers are equally likely to start from either a high or low risk posture, and they may adjust down or up respectively. This is in contrast to Grice, where the speakers are predicted to decrease risk levels as their experience of the task increases.

Davies tested four models of cooperation - Grice's Cooperative Principle, the 'folklinguistic' notion of cooperation often confused with the Gricean Principle, Clark and Wilkes-Gibbs' *Principle of Least Collaborative Effort*, and Shadbolt's *Principle of Parsimony* - on the HCRC Map Task Corpus. HCRC Corpus is made up of task-oriented dialogues (Anderson et al., 1991) which consists of 128 task-oriented dialogues collected from 64 speakers. The task they undertook involved one speaker (the Instruction Giver) describing the route on their map to the other (the Instruction Follower). The maps showed the same fictional location, but they were not identical. The route (which only the Giver had) was based around a number of small named pictures (known as features or landmarks), but not all of these are on both maps: about eight out of eleven are shared. In her analysis, Davies focused on 32 dialogues (16 speakers). The aim was to compare the explanatory power of each of the cooperation principles on real language data. She translated each principle into a dialogue coding scheme and tested it on the dataset. In particular, the coding schemes tried to distinguish among the levels of effort that participants embark on their utterances. This is reflected in a weighting system (see Table 1) that takes into account the effort invested by each speaker, providing a basis for the empirical testing of dialogue principles.

The use of this system provides a positive and negative score for each dialogue move with respect to the effort involved. So, when an instance of a particular behavior is found, a positive coding is attributed with respect to the involved effort level. Instead, a negative coding is attributed when an instance where a particular behavior should have been used is not found. Moreover, regarding negative scorings, the lowest value (-4 i.e. the higher negative weighting) is attributed when a behavior requiring a minimum quantity of effort is not used, while the highest negative value (-1) is attributed when a high effort behavior is not engaged. Vice versa, the lowest positive weighting value (+1) is attributed for a low effort utterance, while the highest positive weighting attribute (+4) is scored when a high effort behavior is observed. This system provides a positive or negative score coming from the sum of all codings for a dialogue. In turn, this score is an account for the effort invested in the dialogue. Davies' attempt to estimate cooperation from a narrow set of indicators to a sort of data-driven set provided for the first time a basis for the empirical testing of dialogue principles. This data driven set is called coding scheme. In linguistics and engineering the large use of corpora has raised questions on coding scheme. In particular there are still many open questions on the coding scheme reliability and the generalization of collected data. Analyzing a multimodal corpus is very demanding. The aim of testing coding scheme reliability is to assess whether a scheme is able to capture observable reality and to eventually allow some generalizations. Following, we report the instructions coded in Davies' coding scheme, with a brief description of each move and the consequent positive or negative attributed weight (Table 2 and Table 3).

Table 1. Effort levels and weightings (from Davies, 2006: 43)

Effort Level (Least first)	Positive Weighting	Negative Weighting
Level 1 – Minimum Effort	+1	-4
Level 2 – Moderate Effort	+2	-3
Level 3 – Medium Effort	+3	-2
Level 4 – High Effort	+4	-1

Table 2. Summary of Positive Codings (From Davies, 2006: 43-44)

SUMMARY OF POSITIVE CODINGS		
INSTRUCT		**Positive Weighting**
+NEW-QUESTION	Asks question *not directly prompted* by previous utterance	+4
+RELEVANT-INFO	Introduces new, unsolicited information (‘new’ in terms of focus, potentially relevant to route section)	+4
+NEW-SUGGESTION	Makes unsolicited suggestion about where route might go nest (need not be *correct*)	+4
+QUERY	Question (function not form) prompted by previous utterance either because of information problem or checking *self* understanding (check if +KNOWLEDGE-MISMATCH is appropriate)	+3
+OBJECTION	Statement (function not form) prompted by previous utterance, concerned with information problem (check if +KNOWLEDGE-MISMATCH is appropriate)	+3
+CHECK	Question which solicits *other* understanding of information already offered	+2

RESPONSE		
+REPLY-MIN –REPLY-FULL	Insufficient or inappropriate information	+1 & -3
+REPLY-YN	Yes-No reply to Yes-No question	+1
+REPLY-FULL	Reply to WH-question, or full reply to Yes-No question	+2
(+INFO-INTEG)	Additional information offered (Move should be coded as REPLY-FULL) [RARE]	+4
FOLLOW-UP		
+ACK-SHORT	Appropriately brief follow-up	+1
+ACK-FULL	Full follow-up	+2
(+INFO-INTEG)	Additional information offered (Move should be coded as ACK-FULL) [RARE]	+4
FEATURE-SPECIFIC CODINGS		
+FEATURE-INTRO	Highlighted (re-)introduction of a feature	+2
+FEATURE-LOC	Attempt to locate position of feature	+3
+FEATURE-UNIQUE	Attempt to uniquely identify feature (e.g. in terms of location)	+3
HIGHER-LEVEL CODINGS		
+KNOWLEDGE-MISMATCH	Move points out mistaken assumption (should be move-coded as +QUERY	+3

Table 3. Summary of Negative Codings (from Davies, 2006: 44-45)

SUMMARY OF NEGATIVE CODINGS		
INSTRUCT		**Negative Weighting**
-NEW-QUESTION	Not applicable	N/A
-RELEVANT-INFO	Failure to introduce useful knowledge when necessary	-1
-NEW-SUGGESTION	Failure to make a suggestion (This behavior is potentially helpful rather than necessary, and therefore failure is rare)	-1
-QUERY	Failure to indicate information problem.	-2
-OBJECTION	Not applicable: defined on difference in function which can only be identified if strategy is realized – use -QUERY	N/A
-CHECK	Failure to check other's understanding of information offered (mainly at topic/segment boundaries)	-3
RESPONSE		
–REPLY-FULL	No response given when required	-3
+REPLY-MIN –REPLY-FULL	Reply too short, or inappropriate	+1 & -3
(-INFO-INTEG)	More information necessary [RARE]	-1

FOLLOW-UP		
-ACK-SHORT	No follow-up given when necessary	-4
-ACK-FULL	Inappropriately brief follow-up. (can occur with +ACK-SHORT)	-3
(-INFO-INTEG)	More information necessary [RARE]	-1
FEATURE-SPECIFIC CODINGS		
-FEATURE-INTRO	New feature introduced, but not highlighted (i.e. treated as shared information)	-3
-FEATURE-LOC	Failure to start negotiation process for unshared (typically) feature	-2
HIGHER-LEVEL CODINGS		
-KNOWLEDGE-MISMATCH	Move fails to point out mistaken assumption (should be move-coded as -QUERY)	N/A

It should be pointed out that negative codings are annotated when a particular dialogue behavior that should have been used is absent. Negative codings are independent from task success.

In the following we focus on two of the four cooperative principles tested by Davies: Gricean *Cooperation Principle* and *Principle of Parsimony* which yielded the best statistical evidence with respect to all the others. Regarding Gricean cooperation, two effects were investigated. The first one is the *avoidance of unnecessary effort.* This effect can only be investigated by looking for changes in dialogue behavior. To do so, the annotators judge what is necessary or unnecessary. In order to investigate *avoidance of unnecessary effort* two particular attributes were used: the checking routines [CHECK] and checking shared knowledge of new landmarks were introduced into the conversation [FEATURE-INTRO]. Although these attributes were chosen because they are an indication of shared knowledge between the two speakers, they are not directly prompted by the task. [CHECK] instruction controls for given instructions. As map features don't have to be checked in advance, the speakers can choose whether to check the route or the feature before moving to the next instruction. As regards [FEATURE-INTRO] instruction, Davies explored the effect of experience on cooperative behavior. She performed unrelated test (Wilcoxon Mann Whitney) to investigate whether there is a significant difference between the first time each Giver instructed a map and the second time each Giver gave the map. The hypotheses developed for Grice's *Cooperation Principle* are reasonably well supported. Davies found evidence of changes in dialogue behavior, improvement in task success and a refocusing of effort (which appears to have been effective). So, rational/efficient speakers learn to invest effort effectively.

As regards *Principle of Parsimony*, the first investigated effect was the *total negative score* (the sum of all the negative scores in each dialogue) as it is an indication of the overall amount of risk taken during the interaction. The second effect analyzed was the change of dialogue behavior, measured again with [CHECK] and [FEATURE-INTRO]. As a result, good support was found for this principle. Task success improved over time, which is associated with both changes in dialogue behavior and a decrease in risks taken as speakers work on a risk-effort trade-off. This is also associated with the refocusing of the effort: overall effort may not decrease and speakers seem to be using it more effectively. They manage to find an optimum point in the risk-effort trade-off and therefore improve their performance on the task. Further findings include the lack of relationship between absolute effort and task success and the positive correlation between equal commitment by the two participants and task success. Non parametric statistical tests were run to test the predictions of the four principles on the information generated by Davies' coding system. The strongest support was found for the Principle of Parsimony. There is evidence that speakers try to minimize effort on an individual basis rather than on a cooperative basis. The findings further endorsed the Gricean Cooperation Principle, although the predictions of this principle were weakened by the difficulty of transforming the underspecified nature of Grice's work into a univocal set of predictions.

2.1.2 Uncooperation in Dialogues

Based on Davies' research, it is clear that there are limits to the extent to which people cooperate. Grice knew that people do not *always* follow cooperation maxims as they communicate, and he identified four ways in which discourse participants regularly break, or fail to fulfill, maxims in conversation: violating, opting out, clashing, and flouting (1989).

What happens, though, when cooperation breaks down? Current models have little to say about this kind of dialogues. Typically, differences in cooperation have been investigated in pragmatics and sociolinguistics usually measuring two aspects of dialogue: dominance and gender. Research on dominance is basically focused on how asymmetry of knowledge held by the two speakers reflects on cooperation (Markova & Foppa, 1991; Drew, 1992). In particular, in the case of the Map Task, the Giver is more knowledgeable than the Follower as the Giver has the route marked on his map while the Follower has not. Another important variable resulting in differences in cooperative behavior is gender. Although the relationship between language and gender is complex, and the relevant evidence on that relationship is difficult to develop and interpret, it seems clear that in both western (Zimmerman & West, 1975; West & Zimmerman, 1983; Fishman, 1983; Holmes, 1983, 1984, 1986; Coates, 1988; Nordenstam, 1992) and non-western (Brown, 1980; Ide, 1982; Smith, 1992) cultures, women are more cooperative and men are more competitive.

Traum and Allen (1994) were the first to propose a model of coordinated dialogue that did not rely on cooperation. In their view, obligations, which are imposed by norms of social interaction, could be ignored and cooperation might or might not be adopted in dialogues. This suggests that in language use a person's level of cooperation depends on the other person's cooperativeness. This statement is further corroborated by theoretical and experimental studies of the Prisoner's Dilemma (Axelrod & Hamilton, 1981; Dreber, Rand, Fudenberg & Nowak, 2008). This classic game theory's play has shown that people prefer to cooperate if the other party cooperates, but not otherwise.

Computational works on Dialogue Agents have mostly modeled cooperation using notions such as *Joint Intentions* (Cohen & Levesque, 1990) and *Shared Plans* (Grosz & Sidner, 1990). These notions are used to automate the kinds of pragmatic reasoning described by Grice and Searle and furthermore to compute speaker meaning using contextual knowledge as well as compositional semantics in order to engage into more flexible dialogues. However, so far it has been proven remarkably difficult to assess the effects of cooperation and non-cooperation on communication exchanges. Davies (1998) tried to analyze cooperation relying on the analysis of the interaction partner's communicative style. Unfortunately, the method she proposed had low reliability. Moreover, there is no dataset addressing explicitly the cooperative and uncooperative communication.

Given that, in our study the first step in order to investigate uncooperative and cooperative communication was collecting a corpus of cooperative and uncooperative dialogues between speakers. The chosen task to elicit conversations was the Map Task (Anderson et al., 2001). The Map Task is considered a default cooperative task and accordingly we use it as the baseline to study cooperation. In attempt to elicit uncooperative utterances, we used a negative emotion elicitation method in carefully controlled circumstances (Anderson, Linden & Habra, 2005). The idea of inducing uncooperative behavior using emotions came from economic game theory, in which emotions were found to be important predictors of cooperation. For example, Pillutla and Murnighan (1996) measured the feelings of respondents when confronted with unfair offers in order to predict their tendency to reject the offer. Anger was positively correlated with the tendency to not cooperate. Other researchers showed that when respondents were treated unfairly, they felt not only anger, but sadness, irritation, and contempt (Bosnam, Sonnemans & Zeelenberg, 2001; Xiao & Hauser, 2005). In a novel study, Sanfey et al. (2003) used functional magnetic resonance imaging to monitor the brain activity of respondents while playing economic games. The results showed that the participants that had a greater activation of negative emotion brain areas when receiving unfair proposals were more likely not to cooperate. On the other hand, the participants that had a stronger activation of brain areas linked to problem solving and cognitive conflict were more likely to cooperate.

Based on these findings, we designed the data collection eliciting negative emotions and expecting a decrease of cooperation in the speaker to whom the negative emotion was addressed. Besides linguistics aspects of interaction, we take into account other modalities of face-to-face communication such as gaze direction and facial expressions. Both of them are useful to study cooperation and emotions.

2.2 Emotion Studies: Categorical *vs* Appraisal Theories

The decision of using emotion to elicit uncooperative behavior has raised a number of questions regarding the nature of emotions. In the past decade, the theoretical concept of emotions has shifted from the discrete or basic emotion theory to dynamic architectural frameworks. While the former were based on the study of a limited number of innate, hard wired affect programs for basic emotions (such as anger, fear, joy, sadness, and disgust; Ekman, 1984, 1992; Izard, 1977, 1993; Tomkins, 1984) the latter are built on appraisal (Frijda, 1986, 2009) and motivational changes. Specifically, the abovementioned foundations of appraisal theories of emotion, explicitly affect the autonomic nervous system (e. g., cardiovascular and skin conductance changes) and the somatic nervous system (motor expression in face, voice, and body). It should also be noted that the appraisal process often occurs in an automatic, unconscious, and effortless fashion (Scherer, 2009). The appraisal theory is quite far from the basic emotion theories, in the tradition of Ekman (1992) and Izard (1993). Basic emotion theory addresses a small number of emotions and has a rigid notion of affect programs, leading to prototypical response patterns (see Scherer & Ellgring, 2007). However, these differences are much less decisive than they appear at first sight. Particularly, basic emotion theorists have stressed in their recent writings that they:

1. consider complex emotions in addition to basic emotions;
2. postulate emotion families that allow for many gradations within each family;
3. assume affect programs to be flexible.

At the date, there is an increasing consensus on a componential approach to emotion and the need to consider appraisal as one of the central underlying mechanisms.

One of the theoretical models in the tradition of appraisal theories of emotions is the Component Process Model (CPM) that is focused on the dynamic unfolding the emotions. As shown in the flow diagram (Fig. 1), the CPM suggests that an event and its consequences are appraised with a set of criteria on multiple levels of processing (i. e. the appraisal component).

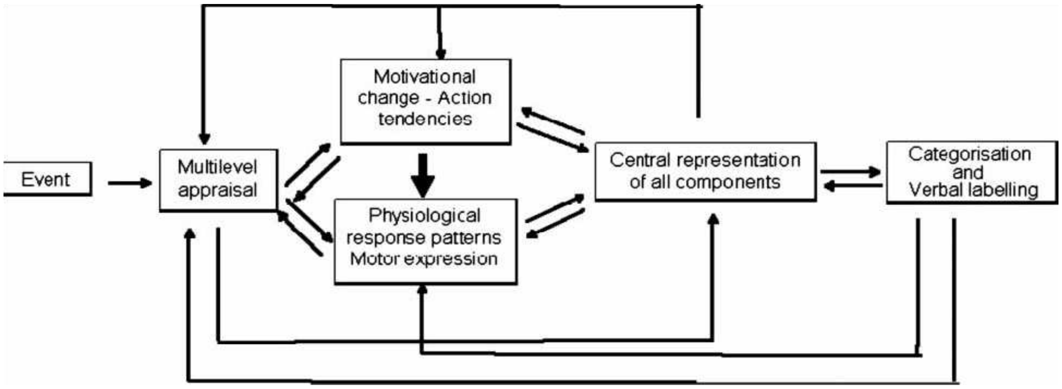

Event
Multilevel appraisal
Motivational change - Action tendencies
Physiological response patterns Motor expression
Central representation of all components
Categorisation and Verbal labelling

Figure 1: The dynamic architecture of the component process model (From Scherer, 2009: 1308)

Based on the appraisal results and the concomitant motivational changes, changes will occur in the autonomic nervous system (e.g., in the form of cardiovascular and respiratory changes) and in the somatic nervous system (in the form of motor expression in face, voice, and body). The fundamental assumption of the CPM is that the appraisal results drive the response patterning in other components. In fact, the appraisal results trigger outputs designed to produce adaptive reactions that are in line with the current appraisal results. Thus, the variety of emotions is the result of all the subsystem changes. These subsystem changes are theoretically predicted on the basis of a componential patterning model, which assumes that the different organism subsystems are highly interdependent and that changes in one subsystem will tend to elicit related changes in other subsystems. The appraisal mechanism requires interaction between many cognitive functions (i. e. to compare the features of stimulus events; to retrieve representation in memory; to respond at motivational urges) and their underlying neural circuits. In addition, CPM controls attention deployment and relies heavily on implicit or explicit computation of probabilities of consequences, coping potential, and action alternatives. As shown in Figure 2, the CPM architecture assumes bidirectional influences between appraisal and various cognitive functions. For example, minimal attention needs to be given for appraisal to start, but a relevant outcome will immediately deploy further attention to the stimulus. Stimulus features are compared with schemata in memory but strongly relevant features are stored as emotional schemata in memory. Event consequences are compared with current motivational states, but particular appraisal outcomes will change motivation and produce adaptive action tendencies. These bidirectional effects between appraisal and other cognitive functions are illustrated by the arrows in the upper part of Figure 2.

A very interesting question in the field of emotion studies concerns the emotion labeling mechanism. Languages differ with respect to emotion vocabulary. Furthermore, the nature and the origin of the differences between the semantic fields of emotion terms in different languages is still an open question. Russell and Barrett (1999) claimed that it is possible to investigate psychological ‘‘primitive’’ of the affective feeling called ‘‘core affect’’. The “primitive” is a point in a low dimensional valence/arousal space and the basis for the construction of a specific emotion category. On the contrary, the appraisal theorists deny the existence of such well defined categories and consider emotive terminology a part of the dynamic representation of an emotion. When an emotion becomes conscious, it can be assigned to a fuzzy

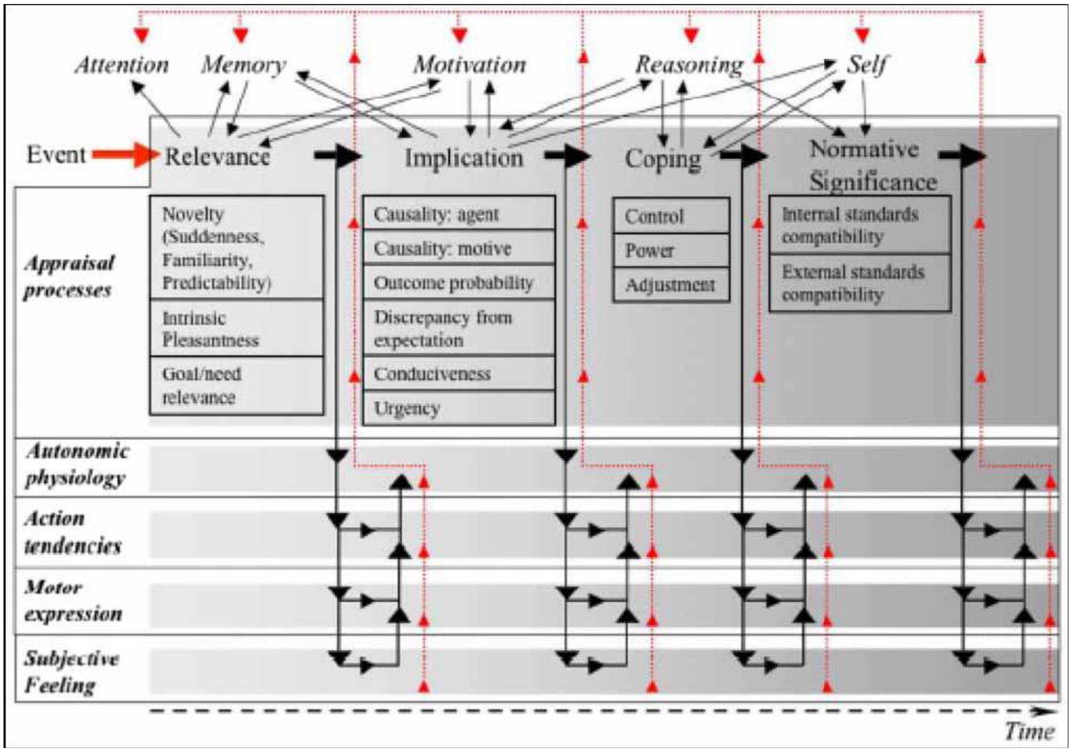


Figure 2: Schematic summary of the component process model (Sander, Grandjean, & Scherer, 2005: 321).

emotion category or be labeled with words, expressions, or metaphors. In appraisal theorists' view, the emotion process is considered as a varying pattern of change in several subsystems of the organism that is integrated into coherent clusters (Scherer, 1984, 2001). The first results suggested that four dimensions are necessary to define the affective space onto which the meaning of major individual emotion terms can be projected: valence, power/control, arousal and unpredictability (Fontaine, Scherer, Roesch & Ellsworth, 2007).

Results on emotion annotation give consistency to appraisal theory's notion of emotion categorization (Truong, 2009). In fact, although the appraisal models of emotions are generally accepted as the prevalent model by the scientific community, the corpora and dataset focused on the study of verbal and non verbal aspects of emotions deal with a limited number of stereotypical emotive expressions. The first example of database focused on emotions is the collection of pictures by Ekman and Friesen (1975), which is based on an early version of the basic emotion theory. This early work has been a source of inspiration for many other data set on emotions, either pictures or audio and/or video. A big issue in emotion corpora collection is the naturalness of the databases. Many databases are made up with skilled actors performing basic emotions. This practice jeopardizes the chances that the resultant face and voice activations are consistent with real everyday life emotions. Indeed database focused on basic emotions have scarce application to real-life emotions. In many cases the speakers involved in an interaction show emotions quite different from the basic ones, as they are often filtered out by social display rules. Given that, in order to collect and annotate emotion database and corpora, one should consider the dynamic aspects of emotions as well as the specific context in which the interaction is taking place.

Corpora research has not generally paid much attention to the questions about context, but the issue is gaining recognition. Context is inescapably linked to modality and emotions are strongly multimodal since they may appear in various different channels. For example, facial expressions of emotions are potentially available in a wide range of contexts, but they are also subject to a wide range of influences. These include culture-specific display rules and also interactions with speech, involving both the lips and the eyebrows.

The emotion level that has rarely been considered is discourse analysis. Lee and Narayanan (2003) classified utterances into rejections, repeats, rephrases, ask-startovers, showing that they relate systematically to emotional state. Beyond that, emotion influences the classical linguistic variables of syntax and vocabulary. A considerable amount of studies provided a rich source of ideas about the way emotion might influence choice of vocabulary—immediacy of expression, concreteness of terms, uses of expletives, and so on (Berger & Bradac, 1982). Recently, Athanaselis et al. (2005) showed that the verbal content of emotional speech can be better recognized when the language model is based on a corpus biased towards utterances with some emotional content.

In the last years, many corpora have been collected and annotated with the aim to study emotions in daily interactions. A large number of these emotion-oriented corpora are multimodal. Usually, multimodal corpora target the recording and annotation of several communication modalities such as speech, hand gesture, facial expression, body posture, etc. In a considerable number of cases these multimodal emotive corpora have been collected in a natural setting and often the resulting data are difficult to classify and analyze. In the following section, I report some of the most remarkable problems on emotion annotation.

2.3 Emotion Annotation Reliability: Percent Agreement, Kappa and *α*

The three main coding scheme sets used so far to describe the emotional content of a database are the following: categorical, continuous, and appraisal-based. Categorical schemes assign terms like angry, ashamed, jealous, and so on to an utterance, a facial expression or a gesture. Numerous teams have tried to label relatively ecological material (Douglas Cowie et al., 2003) using schemes based on established psychological lists of emotions. For example, Craggs and Wood (2004) used a list of emotive terms derived from Ortony and Turner (1990) to annotate emotions. The result of categorical coding scheme annotation was considered unacceptable because of their very low inter-rater agreement. Usually, in order to validate a coding scheme the Kappa statistic is run (Siegel & Castellan, 1988). Kappa is a suitable statistic technique only for clearly separate categories, which is hardly the case of emotions in an ecological setting. Moreover, categorical coding schemes are based on few archetypal emotions which often do not appear in ecological data. In some studies, as a solution to categorical annotations poor reliability, emotions have started to be annotated into *cover classes*. For example, Callejas and Lopez-Cozar (2008) and Abrilian, Devillers, Buisine and Martin (2005) used three categories - positive, neutral and negative- to annotate emotions. Again, the main problem of these coding schemes is their low reliability.

In multimodal annotation many different methods have been proposed to study the reliability of non verbal behavior. Usually these methods are borrowed from computational linguistics such as pairwise agreement (Litman & Hirschberg, 1990; Kowtko, Isard & Doherty, 1992) and percent agreement (Passonneau & Litman, 1993). In pairwise agreement, after the annotation is made by two independent coders, a third coder is asked to confirm the annotation and finalize it in case of disagreement (Poggi & Vincze, 2008; Colletta, Venouil, Kunene, Kaufmann & Simon, 2008; Douglas-Cowie et al., 2005). Percent agreement is calculated as the number of items identified with the same label by two or more independent coders. The total amount is then divided by the total number of labels identified. In many annotation schemes (Allwood, Cerrato, Jokinen, Navarretta & Paggio, 2006, 2007; Cerrato, 2004; Kipp, Neff & Albrecht, 2006; Martel, Osborn, Friedman & Howard, 2002; Martell & Kroll, 2006) percent agreement has been used as a validation method. Nevertheless, neither pairwise agreement nor percent agreement assures reliability and generalization of annotated results because these methods are not corrected for agreement due by chance. In 1993 Wagner (1993) claimed that validation processes based on percent agreement and the hit rate are meaningless. To better understand the problem of agreement among rates due to chance I report in the following an example from Wagner (1993: 4). One could assume to annotate 100 items (e.g. facial expressions) using 3 category labels: neutral, pleasant and unpleasant (see Table 4). These 100 items are pre-selected and categorized by the experimenter: 45 of them are pleasant, 30 neutral and 25 unpleasant. The experimenter asks an annotator to attribute one of the three categories (neutral, pleasant, unpleasant) to each of the 100 items. In Table 4 the fictitious ratings of one annotator on the 100 emotional items are reported.

Overall, 48 items match with the category attributed by the experimenter. The agreement scores are .800 for pleasant, .200 for neutral and .240 for unpleasant. Thus, pleasant items seem to be very recognizable. But we can’t be sure that these scores are more than what we could expect by chance. We can expect consequences from inappropriate measure of performance or failure to correct chance rates which are not easily predictable.

In particular, Wagner noticed that inaccuracies are risky for data interpretation when:

1.accuracy rates are low in comparison to chance agreement so that there is a greater proportional error on (true) chances frequencies;

2.variation in ratings is very high (so that, again, proportional error is greater on true chances frequencies);

3.there is a high number of observations, that have as a result an increase in errors due to the incorrect estimation of binomial and chi-squared distributions.

Conditions 1 and 2 frequently occur in studies on emotions expressions. For example, when accuracy rate is high (as with emotional expressions produced by actors) choosing an unsuitable chance level would lead to totally different reliability results. With regards to condition 3, it might be avoided by ensuring that the number of items to be annotated is not greater than the number of annotators, whatever the unit of analysis may be. Some attempts to solve the third condition have been made in studies on facial expressions of embarrassment (Ekman & Friesen, 1986) and in facial expressions of contempt (Hall & Levin, 1980); these studies focus only on one particular response category and compare its use with different items. Although being sophisticated, these types of procedures are clearly specific and unsuitable for all types of non verbal rating studies.

Table 4. Fictitious ratings of one annotator on 100 emotive items.

Items	**Annotator Ratings**			
	Pleasant	**Neutral**	**Unpleasant**	**Total**
Pleasant	36	9	0	45
Neutral	6	6	18	30
Unpleasant	14	5	6	25
Total	56	20	24	100

.

Following Wagner, for a widest application in non verbal behavior, a measure must fulfill 4 properties (Wagner, 1993: 10):

1.it should be insensitive to annotators' bias;

2.it should be insensitive to differences in the presented items of different type;

3.it should allow separate analysis of accuracy for each item type;

4.It should allow comparison of performance between studies even with different numbers of categories.

The first two requirements correct raw hit frequencies while the third and fourth ones assure generalization and reliability of results.

Carletta (1996: 252) suggested applying Kappa for measuring agreement in corpora annotation. The kappa coefficient (K) measures pairwise agreement among two coders correcting for chance agreement:

$$K= \frac{P(A)-P(E}{1-P(E)}$$

where P(A) is the proportion of times the coders agree and P(E) is the proportion of times we expect them to agree by chance. For more than two coders chance agreement is calculated as the agreement expected on the basis of a single distribution, reflecting the combined judgments of all coders (Fleiss, 1971). Thus, expected agreement is measured as the overall proportion of items assigned to a category k by all coders n. Two important problems are present in reliability studies: 1) annotator bias, underlining the need of increasing the number of annotators when annotators' marginal distributions are widely divergent (Artstein & Poesio, 2005); 2) the prevalence problem, which concerns the difficulty in reaching significant agreement values when most of the items fall under one category. For the latter problem, Artstein and Poesio (2008) argued that, as reliability is the ability to distinguish between categories, in case of skewed data (which means that a category is very common with respect to the others) we must focus on agreement on rare categories if these are the category of interest. Thus, when we are facing with skewed data, the agreement test on rare categories turns to be the significant one.

Another debated point on agreement studies is the interpretation of Kappa scores. There is a general lack of consensus on how to interpret these values. Some authors (Allwood et al., 2006) considered to be reliable the values between .67 and .8. Others, as a "rule-of-thumb", accepted as reliable only scoring rate over .8 (Krippendorff, 2002). Some attempts have been made to apply Kappa to non verbal or emotion annotation but with little success. Usually, for emotions, gesture and gaze annotation Kappa score is quite low. As an example, in the MUMIN coding scheme, gaze direction and head have scored respectively .54 and .20 (Allwood et al., 2007). In the AMI meetings corpus annotation, Carletta (2007) has reported a Kappa score of .54 referred to the so-called “socio-emotional area”, comprising emotions and other non verbal features such as gaze direction; Pianesi, Leonardi and Zancanaro (2006) in “socio-emotional area” have scores ranging from .40 to .60. In emotion annotation, Douglas-Cowie et al. (2005) showed for emotion annotation in three different conditions - audio, visual and audiovisual - Kappa scores ranging from .37 to .54. It should be noted that most of the multimodal corpora cited henceforth are considered ecological. They are recorded from TV shows, interviews or storytelling. As most of them are not task-oriented the coders have to face a large variety of non verbal behavior and linguistics features. It should be considered that the assumption underlining annotation is that coding scheme categories are mutually exclusive and equally distinct from one another (Carletta, 1996). This is often not the case of multimodal and affective coding schemes. As a consequence, multimodal coding schemes are too general or too narrow in categorization in order to successfully annotate the multimodal communication. So far, what is clear is that it seems inappropriate to propose a general cut off point, especially for multimodal annotation, as very little literature on multimodal scheme validation has been reported up to now.

An additional interesting coefficient for corpora validation is *α* (Krippendorff, 1980, 2002). *α* is a weighted measure which calculates the expected agreement by looking at the overall distribution of judgments without telling which coder has produced which judgments. It can be applied to multiple coders’ agreement study and it allows for different magnitudes of disagreement. Moreover, *α* is useful to assess chance agreement when categories are not clearly distinguished from one another. Initially, *α* emerged in content analysis but currently it is widely applied in anaphoric relation annotation. It is calculated as:

$$\alpha = 1 - \frac{D_\mathrm{o}}{D_\mathrm{e}}$$

where Do is the observed disagreement and De is the expected disagreement when the coding is due to chance. It should be noted that for the same data set it is possible to have very different *α* scores as a result of different chosen boundaries. In multimodal annotation, *α* has been applied as a reliability measure only by Reidsma, Heylen & Op den Akker (2008). They use *α* to validate “addressing” annotation data in AMI multimodal corpus. AMI corpus is made up of Multiparty dialogues on problem solving task. “Addressing” is annotated not only linguistically (as for example, when one of the participants asks to another to clarify his/her point or is addressing to the group) but also with gaze direction. Reidsma, Heylen & Op den Akker founded an *α* score of .57 and .87. As pointed out by Artstein and Poesio (2008), it should be noted that *α* score interpretation is sometimes even more problematic than Kappa score. As *α* is a weighted measure, its score interpretation is unpredictable with the Kappa “rule-of-thumb”. This is because one can report very different scores for the very same experiment. New task and distance-metric specific interpretation methods should be assessed. Apart from this difficulty, *α* seems a promising reliability measure for multimodal data annotation. In fact, as the nature of multimodal data usually leads to overlapping categories annotation, *α* can help in estimating data reliability.

In a recent effort to resolve the issue of agreement score interpretation, Reidsma and Carletta (2008) propose to solely rely on machine learning in order to generalize computational linguistics data. They showed how highly reliable annotated data could produced patterns difficult to be generalized, whereas poorly reliable data could be successfully generalized when disagreement does not have learnable patterns. They concluded that agreement coefficients seem not to be suitable indicators of success in machine learning. However, it should be noted that the aim of annotation is not only producing a set of data so as to be implemented in machine learning studies but also to assess whether they can capture some kind of reality. Even if disagreement patterns would lead to generalization, it is not automatically assumed that the generalization will be meaningful. Instead, poor agreement in annotated data should lead researchers to rethink the categorization typologies they are using in their coding scheme or to speculate which is the nature of the data they are confronted with. As the decision whether a coding scheme is reliable is a qualitative one, then we still have to rely on coding agreement.

2.3.1 Measuring Category Judgment Performance in Non verbal Communication Studies: Hit Rate, Index of Accuracy and Unbiased Hit Rate

In the previous section we discussed the Kappa-like statistics and their use for multimodal emotion categorization. Particularly, two main problems were raised: 1) multimodal data may be difficult to be captured by Kappa and 2) there is a (generalized) difficulty to interpret Kappa scores. In this section we will focus on measures currently used to assess data reliability in non verbal behavior and non verbal communication studies. These studies mainly investigated emotion recognition (Goeleven, De Raedt, Leyman & Verschuere, 2008), cultural differences in emotion expression and gesture regulation (Hietanen, Leppänen & Lehtonen, 2004) which are popular topics in multimodal annotation as well. Usually, in non verbal behavior studies the experimenter selects two or more labels from a set of discrete categories (such as, for example, emotion categories or gesture typology) that have to be assigned to a group of stimuli by a group of judges. Items (which are, for example, emotional faces to be annotated) are pre-annotated by the experimenter. The aim of the experimenter is finding out whether his/her labeling is the same with the annotators'. In order to achieve this objective, hit rate (H) is usually used. H is calculated as the proportion of stimuli identified by the judges (Woodworth, 1938). H calculation is very similar to percentage agreement. In fact, in the given data in Table 1, where 48 items match with the category attributed by the experimenter, the H is .800 for pleasant, .200 for neutral and .240 for unpleasant. Again, chance agreement is not taken into account in H calculation. Therefore, Wagner (1993) proposed the application of two alternative statistics: the index of accuracy (IA) and the Unbiased Hit Rate (Hu). In the following, we sketched these two statistical measures currently used in non verbal behavior studies where chance probability is taken into account.

A measure which considers misses or false alarms is the index of accuracy (IA). The IA is the difference between the probability of a hit p(H) and chance probability p(C) (which the proportion of all ratings of that response category). The difference is then divided by p(C) [40]:

$$IA= \frac{p(H)-p(C)}{p(C)}$$

which can also be written as:

$$IA= \frac{p(H)}{p(C)} - 1$$

The resulting index of accuracy is the extent to which rating performance is to be expected better or worse than chance. Using values in table 1, for pleasant items p(H) is 36/45 and p(C) is 56/100, thus IA is calculated as (36/45-56/100)/(56/100)=.429, for neutral p(H) is 6/30 and p(C) is 20/100, so IA is (6/30-20/100)/(20/100)= .000. Finally, for unpleasant items p(H) is 6/25 and p(C) is 24/100, so IA is (6/25-24/100)/(24/100)=.000. Thus, with this analysis we determine that neutral and unpleasant recognition are rated by the annotators below chance level compared to the labels attributed by the experimenter. However, this measure is not suitable for certain types of stimuli. In particular, IA fails in recognition of rarely used categories. Moreover, as IA depends on the size of chance probability, it does not allow comparison between different studies or even different categorical classes.

An example of how IA is applied is in Wagner and Smith (1991). They investigated the effect of expression of positive and negative emotions under social conditions. Pairs of friends and pairs of strangers were unobtrusively videotaped while they viewed and rated (individually) a number of emotional stimulus slides. In a second time, a group of separate annotators tried to code from videotapes the emotions reported by each subject. IA was performed to assess the latter ratings reliability. Expressions were more often correctly identified for participants videotaped with friends than for those recorded with strangers. These results support the hypothesis that the degree to which emotions are expressed depends on the role of an accompanying person. Unfortunately, as this result relies entirely on p(C) size (that is to say on number of ratings and the number of the response categories), it is impossible to compare this study with others having a different number of emotion categories. Thus, IA does not fulfill the fourth property suggested by Wagner as a wider application of a measure in non verbal behavior categorization.

An interesting measure to overcome chance agreement and allowing comparison between different studies is unbiased hit rate (Hu). This measure has been proposed by Wagner (1993) for the analysis of agreement on non verbal behavior. As explained before, hit rate takes into consideration only ratings. In order to consider the annotators' rating performance, Hu is proposed as the answer since it includes false alarms and annotator biases.

Hu is obtained by multiplying the conditional probabilities of a hit p(H) and the differential accuracy p(A) (the latter is the conditional probabilities that the category selected by the annotator matches the same one chosen for that very same item by the experimenter):

$$Hu= p(H) \times p(A)$$

when none of the items is identified by the annotator and none of the selected category matches the experimenter ones, Hu has a value of 0. When an item is always identified by an annotator and the chosen category always matches the experimenter ones, Hu value is 1. Given data in Table 1, for pleasant items Hu is (36x36)/(45x56)=.514, for neutral Hu is (6x6)/(30x20)= .060 and for unpleasant Hu is (6x6)/(25x24)=.060. Hu is insensitive to biases and to the number of the used categories. In addition, Hu can capture not only the sensitivity but also the accuracy with which categorization task is executed, resulting in a precise estimation of the annotators' performance.

As for Kappa, Hu score must be tested for significance (in contrast, as Krippendorff pointed out, *α* does not require to be tested for significance). Thus, once Hu is computed for each annotator and each category, a within-subject ANOVA and a pair t-test must be performed on Hu results to check if raters selected a category higher than chance. As Hu is a proportion, it must be arcsin transformed to perform an ANOVA and a T-test. The within-subject ANOVA should be conducted with each category as within-subject and the arcsin Hu value for each annotator as dependent variable. Furthermore, pair t-tests are conducted between arcsin Hu and chance probability for each annotation category (i. e. for every emotion category such as pleasant, unpleasant and neutral). The aim is to discover if the selection of categories by the group of annotators is above chance. Chance scores are obtained by multiplying the two condition probabilities of both item and annotator performance into a joint probability p(c). p(c) will check for each annotator if the stimulus is correctly identified and the selected category is correctly used. Given data in Table 1, the chance proportion for item/annotation combination is (.45x.56)=.252 for pleasant, (.30x.20)=.060 for neutral and (.25x.24)=.060 for unpleasant. It should be noticed that the chance values for neutral and unpleasant categories are the same as the Hu values. The performance of a group of annotators may be compared with chance by pairing observed values of Hu with corresponding values of p(c). If both within-subject ANOVA and pair t-test have significant p-values, then the selection of the intended category will be above chance level.

In the following section, we will summarize an application of Hu to validate a corpus of emotive faces, the *Karolinska Directed Emotional Faces* (Goeleven et al., 2008). The 490 acted affective facial pictures belonging to the six basic emotions (angry, fearful, disgusted, happy, sad and surprised) and the neutral expression were validated using Hu. After the ratings, Hu values per each emotion ranged from .29 to .89. This indicated that some emotive facial displays were really near to the annotators' representation of the corresponding emotion while others were not. Then, an arcsin transformation of Hu score per annotator for each emotion was separately performed. To validate Hu performance, a within-subject ANOVA were conducted, with emotion categories as within-subject variable and Hu values as dependent variable. The results revealed a main effect of emotion with respect to subject (p<.0001). The chance proportions were calculated per annotator for each emotion so as to check if annotators rated each emotion higher than chance. A paired t-test was performed between arcsin Hu and chance scores for each emotion. As results are significant (p<.0001), then emotion ratings were far above chance. Another recent study used Hu to validate a corpus of sign language gestures and emotions (Hietanen, Leppänen & Lehtonen, 2004).

2.4 Summary and Conclusion

As a consequence of the poor results found so far in categorical emotion annotation, we do not analyze emotions using categorical schemes. As regards appraisal theory, there are few researchers that have analyzed emotion with appraisal coding scheme. Among these few, Scherer and Ceschi (2000) analyzed 45 videotaped interactions between passengers and the airline agents processing their claims for lost baggage. In that research, they claimed that the predictive validity of appraisal can be extended to smile, going beyond verbal report. Similarly, in a series of experiments Kuppens, Mechelen, Smits, De Boeck and Ceulemans (2007) showed that anger can occur in combination with different patterns of appraisals. Furthermore, on the basis of the component patterning model, it can be predicted which motor expressions, action tendencies, and physiological changes can be expected to underlie a negative experience (Scherer, 1987, 2000). Alternatively, observing particular motor expressions of an individual in a given situation, it should be possible to infer the results of the person’s specific appraisal of an event and predict the likely emotion (Scherer, 1992).

On the basis of the previous findings, we collected and annotated a new corpus in which physiological measures and facial expressions were investigated with the intention to analyze emotion typology. Therefore, following the appraisal theory, we elicited the emotion through an anger provoking script and we measured the physiological output of the script, in combination with the pleasantness assessment. Facial expressions are annotated with a coding scheme focused on the facial configurations (e.g. smile, grimace, frowning) and without using emotion categories. Lower and upper face configurations are annotated following mouth and eyebrows shape alone. In this way annotators do not directly take into account aspects such as valence and arousal which can produce an overlapping of two dimensions over a single category. Facial expressions will also be checked as predictors of cooperation.

In Table 5 we summed up the pros and cons of the reliability measures and methods we have described so far. As in multimodal annotation field very few Kappa and *α* studies have been reported so far, it appears to be absolutely necessary that researches clearly report the methodology applied to validate their annotations (e. g. the number of coders, if they code independently or not, if their coding is only manual etc.).

In summary, we can conclude that at the moment the weakest point of multimodal studies and in particular of multimodal studies of emotions is the lack of coding scheme annotation reliability. This can be due to the nature of emotion data. Indeed, annotation of mental and emotional states is a very demanding task. Furthermore, it can be due to the nature of Kappa statistic - which is basically the standard statistic performed to assess coding scheme reliability - requiring categories to be clearly separable from each others. To overcome these limits, we collected a novel corpus, Rovereto Emotion and Cooperation Corpus (RECC). In Chapter 2 we will describe RECC collection. RECC is a task-oriented corpus with registered psychophysiological data and aligned with audiovisual data. This corpus allows to clearly identifying emotive events through psychophysiological data. As a consequence, RECC coding scheme has mutually exclusive and distinct categories as the first interest is not annotating mental states but cooperation, turn management, facial expressions and gaze.

Table 5. Pros and Cons of corpora validation measures.

	Pros	Cons
Kappa	- It is corrected by chance agreement; - It allows comparisons among different coding schemes and different annotation conditions; - it has been widely used in literature for natural and task driven corpus.	- It needs a consistent number of annotators to avoid annotator's bias; - Difficult in reaching significant agreement values for skewed data; - Interpretation of Kappa scores is not so straightforward.
α	- It is useful to assess chance agreement and disagreement patterns when categories are not clearly distinguished from one another, such as in natural multimodal corpora.	- The interpretation of its score values is sometimes even more problematic than with Kappa; - to date it has been applied only once in multimodal corpora validation.
Machine Learning Techniques	- Highly reliable annotated data can give patterns difficult to be generalized. On the contrary, poorly reliable data can be successfully generalized when disagreement does not have learnable patterns.	- The aim of annotation is not only producing a set of data to be implemented in machine learning but also to assess if they can capture some kind of reality; - Poor agreement should lead to rethink categorization typologies used in the coding scheme.
IA	-It takes into account misses or false alarms; -it does not have agreement coefficients or cut off points to rely on for assessing reliability.	- It does not take into account annotator's bias - it fails in recognition of rarely used categories; - it depends on the size of chance probability, therefore it does not allow comparison between different studies or different categorical classes
Hu	- It takes into account chance agreement and the annotator rating performance (bias); -it does not rely on agreement scores difficult to be interpreted or cut -off points.	- It is only suitable for corpora in which item values are pre-selected by the experimenter

Chapter 3

The Rovereto Emotion and Cooperation Corpus (RECC): A New Resource to Investigate Cooperation and Emotion

3.1 Multimodal Corpora collection and validation - Problem statement

In the last years, there has been a growing interest in the investigation of multimodality has resulted in the collection of an increasing number of multimodal corpora. One of the main goals of the collection and analysis of multimodal corpora is clarifying the aspects of speech production. The main research questions addressed are how language and gesture correlate with each other (Kipp, Neff & Albrecht, 2006) and how emotion expression modifies speech (Magno Caldognetto, Poggi, Cosi, Cavicchio & Merola, 2004) and gesture (Poggi, 2007). Other aspects investigated in many studies are multimodal cues of irony (Poggi, Cavicchio & Magno Caldognetto, 2008), persuasion (Guerini, Stock & Zancanaro, 2007) or motivation (Sosnovsky, Brusilovsky, Lee, Zadorozhny & Zhou, 2008). The corpus elicitation method is a crucial independent variable that should be taken into account. Multimodal corpora can be roughly divided into three main categories: acted, task-oriented or ecologically recorded. Acted corpora are mainly focus on facial displays of emotions recordings. The emotive facial expressions produced by expert or semi expert actors are considered the “gold” standard for studying facial display of emotions. This is not completely true, as for example each actor’s production should be validated assessing the real closeness to the “standard” emotion representation with the one that the group of annotators has in mind. Task-oriented corpora, such as Map Task and Multiparty dialogues (Carletta, 2007), are mainly focused on face to face verbal (and non verbal) interactions. Thus, these types of corpora are specifically produced to analyze linguistic features such as turn management and feedback. Additionally, non verbal behavior (such as gaze, gesture and even emotion displays) is often analyzed. Ecological corpora are usually recorded from TV shows, news and interviews. Ecological corpora usually comprise a wide range of verbal and non verbal features. The potential of developing coding schemes for annotating such a variety of verbal and non verbal features depends on the researcher’s resources.

As I mentioned before, the large use of corpora in linguistics and engineering has raised questions on coding scheme reliability. Collecting a multimodal corpus might be computationally demanding because of the large amount of space needed to store the audiovisual recordings but it even more challenging to be analyzed. Consequentially, the collection of such a large amount of multimodal data has raised the question of corpora analysis and therefore the problem of coding scheme reliability. The aim of testing coding scheme reliability is to assess whether a scheme is able to capture the observable reality and eventually allow some generalizations. Multimodal coding schemes are mainly focused on dialogues (dialogue acts, topic segmentation, emotion and attention) and can include in their analysis the so called "emotional area" (e.g. the EMOTV annotation scheme, Abrillers et al., 2005) or the relationship between gesture and speech (e.g. FORM coding scheme, Martel et al., 2002; Martell & Kroll, 2006; and CoGEST annotation scheme, Gut, Looks, Thies & Gibbon, 2003). The Multimodal Score annotation scheme (Magno Caldognetto et al., 2004; Magno Caldognetto & Poggi, 2001, 2002) is strictly focused on communicative goals of prosody, facial movements and posture analysis. Other multimodal schemes such as MUMIN (Allwood et al., 2007) analyze turn management, gesture and face movements. To the date, many coding schemes had been settled to capture multimodal aspects of communication. However, very few of these coding schemes are reliable. Testing the coding scheme reliability is a key asset for an annotated corpus. Reliability measures assess whether a coding scheme is able to capture in some way the observable reality allowing some generalizations.

Since the mid-Nineties, in natural language processing and computational linguistics studies, Kappa has found application in validating coding scheme reliability. Basically, Kappa is a statistical method that assesses agreement among a group of observers. Kappa is currently applied to assess agreement on corpora annotation. In multimodal communication it is also very important assessing the coding scheme reliability. Thus, in order to validate some multimodal coding schemes, often Kappa has been used. Presently many multimodal coding schemes have a very low Kappa score (Carletta, 2007; Douglas Cowie et al, 2005; Pianesi, Leonardi & Zancanaro, 2007; Reidsma, Heylen & Op Den Akker, 2008). This could be due to the nature of multimodal data. In fact, some authors (Coletta et al., 2008) argued that annotation of mental and emotional states is a very demanding task. The low annotation agreement which affects multimodal corpora validation could also be due to the nature of the Kappa statistics. In fact, the assumption underlining the use of Kappa as a reliability measure is that coding scheme categories are mutually exclusive and equally distinct from one another. This is clearly difficult to be obtained in multimodal communication as communication channels (i.e. voice, face movements, gestures and posture) are deeply interconnected into one another and contribute (case by case) to the final meaning of the multimodal “sentence”. In the following section, I review the methods currently used to validate multimodal corpora. The reliability studies are focused on categorical judgments and the validation of non verbal behavior such as gaze, gesture and emotions, which are the main focus of the investigation in most multimodal coding schemes.

In order to overcome both the limitations caused by the nature of data and the nature of Kappa, we collected a new corpus, Rovereto Emotion and Cooperation Corpus (RECC). RECC has been collected to shed light on the relationship between cooperation and emotions in dialogues. in accordance with the appraisal theory of emotion, we believe that RECC will allow us to obtain a broader perspective regarding emotions during face-to-face interactions.

Further, using psychophysiological state of the speakers, their facial expressions and pleasantness/unpleasantness assessment of the situation will enable us to reliably evaluate an emotive state in each situation.

3.2 The RECC Corpus Design

RECC is an audiovisual and psychophysiological corpus of dialogues elicited with a modified Map Task. The Map Task is a cooperative task used for the first time at the University of Glasgow by the HCRC group at Edinburg University (Anderson et al., 1991). The HCRC Map Task Corpus was produced in response to one of the core problems of natural language studies. Despite most language use takes the form of unscripted dialogue, much of our knowledge of language is based on scripted materials and carefully selected examples. As a consequence, there is no evidence that inn naturally occurring speech a certain phenomena of theoretical interest will appear with any frequency. Even huge corpora might fail to provide *sufficient instances* to support any strong claims about the phenomenon under study. Moreover, HCRC Map Task addressed the problem of *context*: critical aspects of both linguistic and extra linguistic context may be either unknown or uncontrolled. The HCRC research group’s intention was to elicit unscripted dialogues to boost the likelihood of occurrence of certain linguistic phenomena, and to control some of the effects of context. To this extent while Map Task dialogues are spontaneous, the HCRC corpus is largely and carefully, elicited. One of the issues researchers interested in the analysis of conversational data have to face is the problem of the Observer’s Paradox: how can one record natural data when ethical and legal requirements demand that the subjects know that they are being observed? Labov's answer was to engage the subject by getting them to talk about a near-death situation. Obviously, the talk genre produced in such a situation is primarily a transactional rather than an interaction type. Instead, our main interest is how participants manage, transfer and negotiate information. There are interactional aspects to this – specifically, not answering questions or directly refusing the interaction partner's suggestions would probably lead to partial or total breakdown in the conversation. Our analysis considers aspects as the abovementioned since our primary interest is the interpersonal aspects of talk. Nevertheless, the use of task-oriented data could be unsuitable to study cooperation and emotions. Then, again, although “casual” conversation analysis is often seen as a gold standard in pragmatics, there is an increasing interest in both other sort of dialogues and other methodologies. For instance, business talk is a type of task-oriented dialogue concerned with the transferal and negotiation of information (e.g. Bargiela-Chiappini & Harris 1997; Connor & Upton 2004). The above studies argued that task-oriented data is of legitimate interest to linguists, provided that their aims have considered the constraints of the data. Besides, as Davies (1998, 2006) argued, when analyzing a conversation there is the need to know the speakers' state of knowledge and their likely goals. Such a degree of insight is rarely possible when observing casual conversations. It is the transactional nature of the Map Task which makes our approach possible.

The HCRC Map Task dialogues involved two participants. The two speakers sit opposite one another and each has a map which the other one cannot see. One speaker – designated as the *Instruction Giver* - has a route marked on his/her map while the other speaker - the *Instruction Follower* - has no route. The speakers are told that their goal is to reproduce the Instruction Giver's route on the Instruction Follower's map. The maps are not identical and the speakers are told this explicitly at the beginning of their first session. All maps consist of landmarks – also called *features* - portrayed as simple drawings and labeled with a name. The differences in the maps result from the systematic manipulation of the design variable *sharedness*: it is the extent to which features are contrasted or shared between pairs of maps. A number of the features were *common* in both maps, while other features varied for absence/presence on the two maps, name change and position (relevant vs. irrelevant, depending of their closeness to the road). Another variable taken into account was eye contact. The option of placing a barrier between Map Task participants to prevent them from seeing each other's faces allowed controlling the availability of the visual channel for communication. Therefore, half of the participants who took part in the task were able to make eye-contact with their partner, while the other half had no eye-contact. Another variable controlled for by the HCRC Map Task is the familiarity of the two speakers. Participants were divided into groups of four called “quads”. Each person was involved in four dialogues in their quad, and each quad generated eight dialogues in total. Each participant is a Giver twice and a Follower twice; they give the same route twice (to different Followers) and are Followers on two different maps. The main advantage of the Map Task dialogues over many others task (e.g. Grosz & Sidner, 1986; Clark & Wilkes-Gibbs, 1986; Clark & Schaefer, 1987; Schober & Clark, 1989; Clark & Brennan, 1991) is that participants do not have all the necessary information. It is not sufficient for the Giver to *describe* the route to the Follower: without the knowledge of the unshared features (on either map), it is likely that the route drawn by the Follower will describe some aspects of the route incorrectly. This means the dialogue is an equal enterprise. The input of the Giver and Follower is equally important, and there is joint responsibility for a good task result, rather than being the main responsibility of the Giver. Following the collection of HCRC Map Task corpus, the annotation of dialogue structures was a straightforward task.

When the concept of a dialogue coding system was introduced, most people assumed that its purpose was the identification and labeling of overall dialogue structures (e.g. Carletta et al., 1997; Houghton & Isard, 1987; Kowtko, Isard & Doherty, 1992; Sinclair & Coulthard, 1975) or of structures within a dialogue (e.g. Conversation Analysis) rather than a scheme which attempts to identify the presence (or *absence* in Davies' occurrence*)* of certain types of discourse strategies. Davies produced a subset of HCRC Map Task coding scheme to analyze cooperation. The author’s goal was to annotate the presence or absence of the same strategies so as to code cooperation. Therefore, she annotated what was not there as well as what was. Annotation is a totally evaluative methodology, and this contradicts what is often perceived as a basic tenet of linguistics: *describe* not *prescribe*. Accordingly, evaluation could be seen as a type of prescription. However, Davies replied to this critique that she did not investigated language in the way that concerns a descriptive linguist. On the contrary, the linguistic strategies she evaluated with her coding scheme were essentially realizations of higher order planning, and did not involve issues of ‘standards’ (i.e. grammaticality, lexical choice or register). Moreover, the main point in annotating dialogues is that other annotators would agree with the researcher's assessment. This essentially demands that the coding scheme must be reliable, well-defined, rigorous and usable by other coders (Carletta 1996; Isard & Carletta, 1995). To this end, we undertake a reliability study on the annotated data.

Our Map Task has some similarities with respect to the HCRC one. In front of them, the participants had both a map with a group of features. A number of them are in the same position and with the same name, but the majority of them is in different positions or has names that sound similar to each other (e. g. Maso Michelini vs. Maso Nichelini, Fig. 1). The Giver must drive the other participant (the Follower) from a starting point (the bus station) to the finish (the Castle of Rovereto). Giver and Follower were both native Italian speakers and they did not know each other before the task. As in HCRC Map Task, our corpus interactions have two conditions: screen and no screen. In the screen condition a barrier was present between the two speakers. In the no screen condition a short barrier was placed between the speakers allowing Giver and Follower to see each others' face. Screen conditions were counterbalanced. The two conditions enabled us to test whether seeing the speakers face during interactions influences facial emotion display and cooperation (for the relationship between gaze/no gaze and facial displays see Kendon, 1967; Argyle & Cook, 1976; for the influence of gaze on cooperation and coordination see Brennan, Chen, Dickinson, Neider & Zelinsky, 2008). Previous studies have shown that visual access to each others' non verbal behavior fosters a dyadic state of rapport that facilitates mutual cooperation (Argyle, 1990; Tickle-Degnen & Rosenthal, 1990; Hendrick, 1990). However, these findings do not establish whether facial cues actually are predictive of cooperation. A further condition, *emotion elicitation*, was added. In emotion elicitation conditions the Follower or the Giver can alternatively be a confederate, with the aim of getting the other participant angry[2].

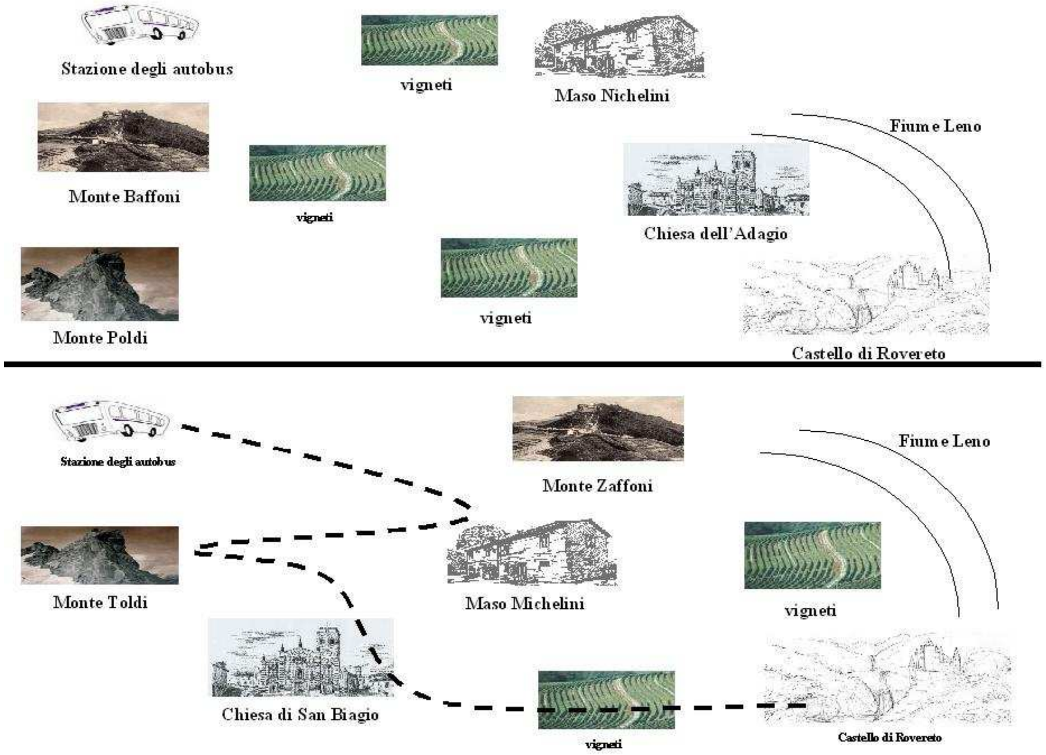


Figure 1: Giver and Follower Maps of the RECC corpus

3.3 Recording and eliciting procedure

RECC is made up of 20 interactions, 12 with a confederate, for a total of 240 minutes of audiovisual and psychophysiological recordings such as the

[2] All the participants had given informed consent and the experimental protocol was approved by the Human Research Ethics Committee of University of Trento.

electrocardiogram, the derived heart rate value and skin conductance. During each dialogue, the psychophysiological state of non-confederate Giver or Follower is recorded and synchronized with video and audio recordings. So far, RECC corpus is the only multimodal corpus which has psychophysiological data for assessing emotive states. The psychophysiological state of each participant has been recorded with a BIOPAC MP150 system. Audiovisual interactions were recorded with 2 Canon Digital Cameras and 2 free field Sennheiser half-cardioid microphones with permanently polarized condenser placed in front of each speaker.

The recording procedure of RECC was influenced by the work of Anderson, Linden & Habra (2005). They investigated the physiological arousal due to acute anger provocation; stress reactivity and recovery were measured on participants performing a mental arithmetic task while receiving scripted comments at set intervals designed to provoke anger through harassment.

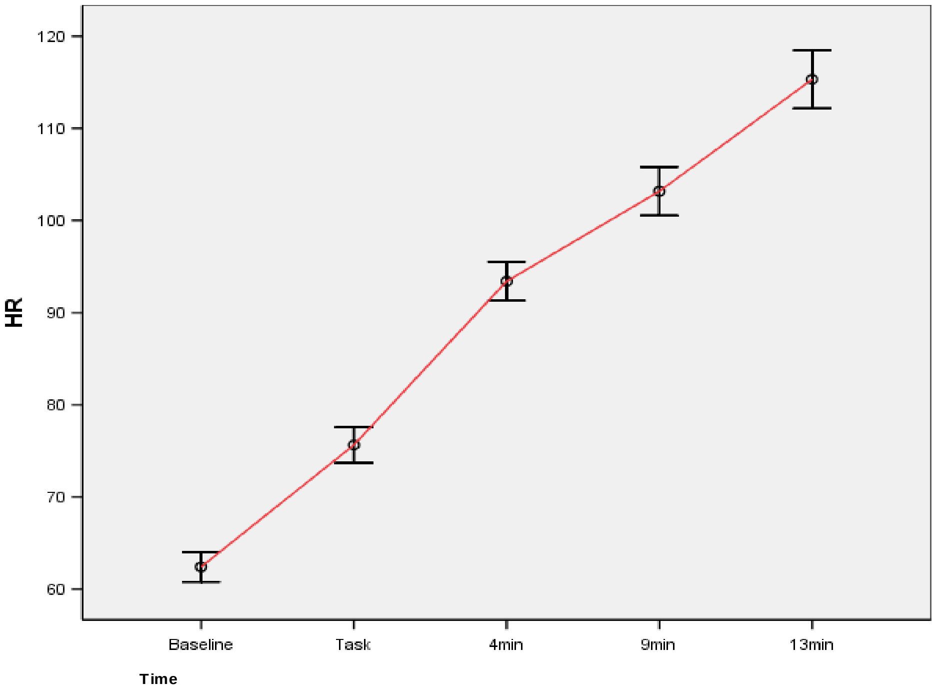
The recording procedure of RECC was the following. Before starting the task, we recorded the *baseline condition* of the participant. Specifically we recorded participants' psychophysiological outputs for 5 minutes without challenging them. Then the task started and we recorded the psychophysiological outputs during the interaction first three minutes that interaction occoured which we called the *task condition*. Soon, the confederate started challenging the other speaker with the aim of getting him/her angry. Two groups of subjects were recorded. The first group consisted of 12 Italian native speakers (average age=28.6, dv=4.36) matched with a confederate partner. During these sessions, the confederate (the same person in all the interactions) performed uncooperative utterances in carefully controlled circumstances (*14*) by acting negative emotion elicitation lines at minutes 4, 9 and 13 of the interaction.

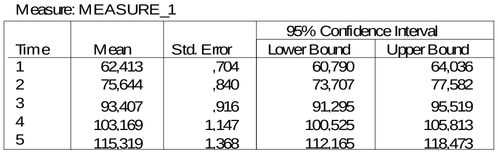
The following lines were given by the confederate when acting the Follower role:

- *"You are sending me in the wrong direction, try to be more accurate!";*
- *"It's still wrong, you are not doing your best, try harder! Again, from where you stopped";*
- *"You're obviously not good enough at giving instructions".*

A control group of 8 pairs of participants (average age=32.16, dv=2.9) were also recorded while playing the Map Task with the same maps. Eye contact, communicative role (Giver and Follower) and gender (male or female) conditions were counterbalanced.

Our hypothesis is that the confederate's uncooperative utterances would lead to a reduced level of cooperation in the other participant. To test it, we first need to check if the eliciting protocol adopted caused a change in participants' heart rate and skin conductance. In Fig. 2 we show the results of a 1x5 ANOVA executed in confederate condition. Heart rate (HR) is confronted over the five times of interest (baseline, task, after minute 4, after minute 9, after minutes 13), that is to say just after emotion elicitation with the script. A HR x Time ANOVA showed a significant effect of Time (F(4, 8)=2.48, p<.00001), meaning that HR changed between task beginning and the three sentences in the script. In the control group session, in addition to a baseline measurement, HR was measured 3 times at equal intervals dung the interaction. A HRxTime ANOVA showed the effect of Time was non-significant (F(3,5)=3,28, p<.117). So, HR is significantly different in the confederate condition, meaning that is to say that the procedure to elicit emotions allows recognition of different psychophysiological states with respect to the non confederate condition. Moreover, the indicated HR values confirmed the ones showed by Anderson, Linden and Habra (2005).





Measure: MEASURE_1

Time	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	62,413	,704	60,790	64,036
2	75,644	,840	73,707	77,582
3	93,407	,916	91,295	95,519
4	103,169	1,147	100,525	105,813
5	115,319	1,368	112,165	118,473

Figure 2: 1x5 ANOVA on heart rate (HR) over time in confederate condition in 12 participants

Furthermore, from the inspection of skin conductance values (Fig. 3) there is a linear increase of the number of peaks of conductance over time. This can be due to two factors: emotion elicitation and also an increase in task difficulty leading to higher stress and therefore to an increasing number of skin conductance peaks.

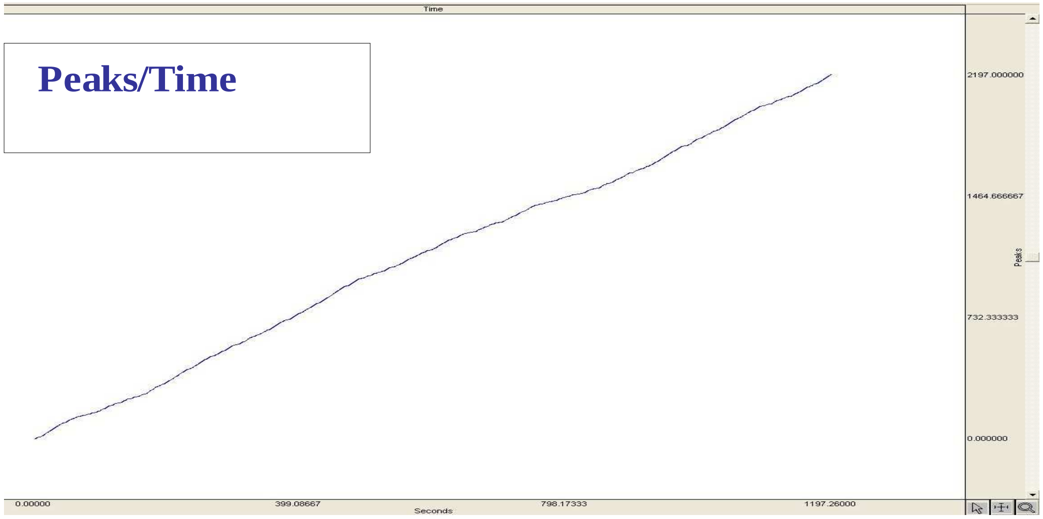


Figure 3: Number of skin conductance positive peaks over time in confederate condition in 12 participants

According to Tassinary & Cacioppo (2000) pointed out, it is not possible to assess which emotion arises based on psychophysiological data alone. In fact, HR and skin conductance are signals of arousal. So, a high arousal can be due to emotions characterized by high arousal and high valence such as happiness or high arousal and low valence such as anger. Therefore, following the appraisal theory of emotions, a 7 points Modified Self-Assessment Scale (adapted from Bradley & Lang, 1994) was completed by all participants. The aim was to obtain the emotive valence assessment measuring the polarity (positive to negative) of the emotion felt by each speaker toward his/her partner during the interaction. Subjective valence ratings were measured by having 14 participants complete a 7.5 cm visual analogue emotion rating scale form (Fig. 4). The polarities of the rating scale were counterbalanced for left and right. 43% of the participants rated the experience as quite negative, 29% rated the experience as almost negative, 14% of participants rated it as negative and 14% as neutral. Participants who reported a neutral experience were discarded from the corpus.

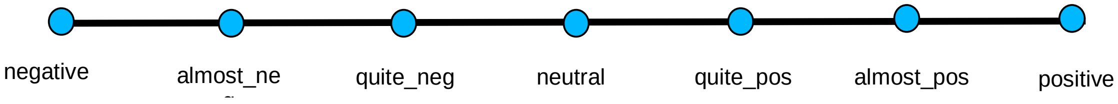


Figure 4: The emotion polarity rating scale form

3.4 Corpus transcription

RECC corpus consists of manually produced orthographic transcriptions for each speaker and in addition, it is time aligned with all the communication modalities. Videos were imported using the ANViL (ANnotation of Video and Language) software (Kipp, 2001). ANViL is a free video annotation tool offering frame-accurate, hierarchical multi-layered annotation driven by user-defined annotation schemes. ANViL can import pitch and intonation data from arbitrary tiers in PRAAT - a free software program for the analysis of speech. Originally, ANViL was developed for Gesture Research, but it has also proved suitable for many areas of research such as human-computer interaction, linguistics and anthropology. The annotation board is intuitive and it shows color-coded elements on time-aligned multiple tracks corresponding to different communicative modalities. For example, in our corpus we codified speech, cooperation, turn management, gaze, upper and lower part of the face configurations. The different modalities we wanted to code were specified in an XML file. The transcription output is an XML file which includes the dialog text and data relevant to the timing of starting and ending of each codified element.

Orthographic Transcriptions of the interactions were done adopting a subset of the conventions applied to the transcription of LUNA project speech corpus (Rodriguez et al., 2007). All possible spontaneous speech effects were transcribed such as disfluencies, hesitations, repetitions, grammatical and pronunciation errors, and filled pauses. Two transcribers converted the recordings into plain texts. Every conversation was divided into turns related to the Giver and the Follower. Firstly, in order to make the subsequent processing easier and the form of the transcribed files more uniform, we adopted the following conventions:

– Spellings are transcribed using capital letters separated by spaces and tagged with symbols e. g. [pron=SPELLED-] M I C H E L I N I [-pron=SPELLED];

– Numbers are transcribed as words;

– Only the actually spoken part of a word is transcribed. Possible truncation is marked with *lex*=_, as annotators do not interpret a word or an utterance;

– In case of mispronunciation, the correct form is transcribed with an indication that it has been mispronounced [pron=*-] Micalini [-pron=*]

– In general, the transcription does not include punctuation marks unless the F0 contour will indicate a question mark or a full stop;

– Words that cannot be recognized are transcribed with the symbol *pron*=**;

– Tag *lex*=*FIL* is used to represent pause fillers, hesitations and articulatory noises as breath, laugh, cough, etc;

– Non-human noises are annotated with the tag *noise*;

– Silence is annotated only if it lasts more than 1 second – as *sil*.

Secondly, transcribers fixed the problems in each orthographic transcription and run a validation script to find unrecognized spelling and transcription codes.

RECC is a multimodal corpus. Actions and communicative movements produced by the upper and lower part of the face, gaze and head movements were transcribed and annotated. In the first step, transcribers marked the beginning and the end of each individual action in each video segment. Once the first step was completed, a second step was performed to verify the precision of boundaries. If the beginning or the end of each event had a very large error (> 200 msec), the transcribers modified the appropriate event. If too many events had been coded previously, then the transcribers deleted the unnecessary ones. If there was a missing event, they could set it by adding the duration of the new event.

3.5 Conclusions

As already mentioned above numerous multimodal corpora have been collected so far as a result of the increasing interest toward multimodality. Furthermore, Kappa has often been used in order to validate the multimodal coding schemes. However, thus far many multimodal coding schemes have a very low Kappa score. This could be attributable to either to the nature of the multimodal communication which is multichanneled and multilayered or the specificity of the Kappa statistics that requires categories to be mutually exclusive and equally distinct one another. This is clearly difficult to be obtained in multimodal communication as communication channels (i.e. voice, face movements, gestures and posture) are deeply interconnected with one another and altogether contribute (case by case with different weight) to the final meaning of the multimodal “sentence”. RECC was collected with the goal to overcome the abovementioned drawbacks. The main goal of this corpus is to investigate the relationship between cooperation and emotion in a dialogue setting. Following the appraisal theory of emotion, in an attempt to assess that an emotive state was constant, we recorded and then aligned the psychophysiological data (HR and Skin Conductance) and the corresponding facial expression. The participants received scripted comments at set intervals designed to provoke anger through harassment. The psycophysiological recordings showed that the HR is significantly different in the confederate condition as opposed to the non-confederate condition. Thus, the emotion eliciting procedure allows recognition of different psychophysiological states. Moreover, there is a linear increase of the number of peaks of skin conductance over time. Pleasantness of the experience was measured after the interaction using a visual analogue emotion rating scale form. The scores showed that the majority of the participants rated the interaction as a negative experience. Participants that rated the interaction as neutral or positive were discarded from the corpus. In the next chapter we will describe RECC coding scheme. The coding scheme is designed to annotate cooperation and facial expressions along with other pragmatics and multimodal aspects of dialogue. Finally, a reliability study will examine the annotated data to estimate whether RECC coding scheme is well-defined, rigorous and usable by coders.

Chapter 4

The RECC Coding Scheme and its Validation

4.1 Emotion Coding Scheme

The emotion annotation coding scheme used to analyze our Map Task is different from the emotion annotation schemes proposed so far in computational linguistic literature. As we briefly sketched in Chapter 2, the state of the art on emotion annotation methods can be gathered in two major groups. For the first group, the main idea is derived from Craggs and Wood (2005). The authors proposed to annotate emotions based on a scheme where emotions were expressed at different blending levels (i. e. blending of different emotion and emotive levels). According to Craggs and Wood annotators must label the given emotion with a main emotive term (e. g. anger, sadness, joy etc.) and correct the emotional state with a score ranging from 1 (low) to 5 (very high). In the second group, a three steps rank scale based on emotion valence - positive, neutral and negative (Martin et al., 2006; Callejas & Lopez-Cozar, 2008; Devillers, Vidrascu & Lamel, 2005) - was used to annotate a variety of corpora mostly recorded from TV interviews. However, both these methods had quite poor results in terms of annotation agreement among coders. Moreover, several studies on cognitive aspects of emotions have shown how emotional words and their connected concepts influence emotion judgments and their labeling (for a review, see Feldman Barrett, Lindquist & Gendron, 2007). Thus, labeling an emotive display (e. g. a voice or a face) with one or more emotive terms is definitely not the best solution for recognizing an emotion.

Keeping this in mind, we decided not to label emotions directly but to indirectly attribute arousal and valence values. In our coding scheme we concentrate on the annotation of face configurations. According to the appraisal theory of emotion, an emotion affects the autonomic nervous system (psychophysiological recordings to measure cardiovascular system and skin conductance changes) and the somatic nervous system (motor expression in face, voice, and body). Regarding the nervous system, an ongoing debate in brain and cognitive sciences argues whether the perception of a face -and, specifically, of a face displaying emotions- is based on a holistic perception or the perception of parts. Although many studies in neuroscience are persistently examining how to determine the basis of emotion perception and decoding, it is still not clear how brains and computer might learn the parts of an object such as a face. Recognition of facial expressions and identity recognition seem to be dissociated. The strongest evidence for a possible dissociation between the identification of faces and that of facial expression recognition is demonstrated by agnosic patients. Agnosics have an impaired ability to identify individual faces (even those of their family and themselves), but preserve the ability to recognize facial expressions (Bruyer et al., 1983; Farah, O'Reilly & Vecera, 1994). The opposite can also occur; i.e., cases have been described where agnosic patients were unable to classify facial expressions, but they exhibited a intact ability to identify faces (Kurucz, Feldmar & Werner, 1979). One of the most notable findings against the independence of identity and facial expression recognition is that people are slower in identifying happy and angry faces than identifying faces with neutral expression (Etcoff & Magee, 1992). Similarly, subjects are slower in identifying pictures of familiar faces when those are shown with uncommon facial expressions (Hay, Young, & Ellis, 1991) or when some artificial deformations are added to the images.

Understanding how different characteristics can be extracted from a single facial image is central to achieving an accurate conceptual framework for all aspects of face perception. It should be noted that interesting insights on facial perception have emerged from image-based analysis techniques, such as Principal Component Analysis (PCA) algorithms, which learn holistic representations. PCA is a standard statistical technique used to identify a relatively small number of factors representing the relationships among many inter-correlated variables. As applied to the image-based analysis of faces, PCA serves a similar function: it identifies a limited number of factors that can represent the complex visual information in faces in a suitable form for face recognition. PCA-based systems have been proved to reliably extract and categorize facial cues such as identity, sex (Hancock, Burton, & Bruce, 1998), race (Valentin, Abdi, & O’Toole, 1994) and expressions (Calder, Burton, Miller, Young & Akamatsu, 2001). PCA has also been accepted as a good psychological metaphor for the structural encoding and representation of faces.

Alternatively, Donato, Bartlett, Hager, Ekman and Sejnowski, (1999) compared the performance of a number of image-based analysis techniques, including PCA, in their quest to develop an automated method of identifying facial expressions based on muscle position. Donato et al. analyzed separately the upper and lower parts of the face separately obtaining good results. On the contrary, when whole facial images were analyzed with PCA *poorer* performances were found (Padgett & Cottrell, 1995) compared to PCA analyses in which the eye and mouth regions were analyzed separately (part-based analyses). One possible explanation for this finding is that Padgett and Cottrell's part-based PCAs may have produced better classification rates because the former method would have reduced the level of noise in the analysis.

Nevertheless, researchers on emotion recognition based on face displays agree that some emotions as anger or fear are discriminated only by mouth or eyes configurations. The face seems to be evolved to transmit orthogonal signals, having a lower correlation with each other. Then, these signals are deconstructed by the "human filtering functions" - i. e. the brain - as optimized inputs (Smith, Cottrell, Gosselin & Schyns, 2005). Unlike PCA, other methods such as the non Negative Matrix Factorization (NMF) extract parts from visual data only on a positive constrains leading to part based additive representations (Lee & Seung, 2001). The goal of this technique is to find intuitive basis in training examples that can be faithfully reconstructed using linear combination of basic images which are restricted to non-negative values. Thus NMF basis images can be understood as localized features that correspond better to the intuitive notions of the parts of the images. A face can conceptually be represented as a collection of sparsely distributed parts: eyes, nose, mouth etc. NMF gains better results in facial identity and expression recognition with respect to PCA-alike algorithms (Xue, Tong & Zhang, 2007; Buciu & Pitia, 2006). On account of such findings, we decided to code facial expression based on a part-based coding scheme. It can be argued that there is no necessity in devising a new facial expression coding scheme since we can use the well known Facial Action Coding Scheme (FACS, Ekman & Friesen, 1978) to code facial expressions. FACS is a good coding scheme to annotate face expressions starting from movement of muscular units, called action units. Some of our coding scheme features to annotate face expressions are inspired by facial configurations coded in FACS. Even if accurate, FACS method has three major shortcomings. First of all it is a slightly problematic to annotate facial expression, especially the mouth ones, when the subject analyzed is speaking, as the muscular movements for speech production overlaps with the emotional configuration. Secondly, learning FACS is very time consuming and resource exhausting. Finally, the method suggested by the authors to assess reliability among coders is hardly corrected for chance agreement.

In our coding scheme facial expressions are “deconstructed” in eyebrows and mouth shapes (Table 1). The shapes have implicit emotive dimensions. For example, a smile was annotated as “)” and a large smile or a laugh is marked as “+)”. The latter markup means a higher valence and arousal than the previous one. Other annotated features are grimace “(“, asymmetric smiles (*1cornerup*), lips in normal position/closed mouth, lower lip biting and open lips (*O*). As regards eyebrows, annotators marked them in normal position, frowning (two levels: *frown* and +*frown*, the latter with an implicit higher valence) and eyebrows up (*up* and +*up*).

Table 1. Coding scheme for facial expression annotation

Upper or lower face configuration	**Annotation label**
Open mouth	O
Lips in relaxed position/closed mouth	-
Lip corners up (e.g. smile)	)
Open smile or laugh	+)
Lip corners down (e.g. grimace)	(
Lower lip biting	lbiting
1 mouth corner up (asymmetric smile)	1cornerUp
Eyebrows relaxed	- -
Eyebrows up	up
Eyebrows very up	+up
Frown	frown
Deep frown	+frown

4.2 RECC Cooperation Coding Scheme

The approach we used to analyze cooperation in dialogue task is mainly based on Davies' model (Davies, 2006). The basic coded unit is the “move”, which stands for the individual linguistic choices used to successfully complete the Map Task. The idea of evaluating the utterance choices in relation to the task success can be traced back to Anderson and Boyle (1994), who linked the utterance choices to the accuracy of the route achieved on the Map Task. Davies extended the meaning of “move” to the goal evaluation, computing the effort needed to plan and produce an utterance. In particular, Davies stressed some useful points for the computation of effort between the two interaction partners:

- *social needs of dialogue:* there is a minimum effort needed to keep the conversation going. It includes minimal answers like “yes” or “no” and feedbacks. These brief utterances are classified by Davies (following Traum, 1994) as low effort, as they do not require much planning to the overall dialogue and to the joint task;
- *responsibility of supplying the needs of the communication partner:* to keep an interaction going, one of the speakers can provide follow-ups which take more consideration of the partner’s intentions and goals in the task performance. This involves longer utterances and, of course, a larger effort;
- *responsibility of maintaining a known track of communication or starting a new one:* substantial effort is needed in considering the actions of a speaker within the context of a particular goal. Speakers mainly deal with

> situations where one of them reacts to the instruction or question given by the other participant, rather than move the discourse on another goal. Indeed, goal shifting is assumed to involve a considerable effort as it requires reasoning about the task as a whole. This effort adds up to the effort needed to produce a particular utterance.

Following Traum (1994), speakers tend to engage in lower effort behaviors than higher ones. Thus, if you do not answer to a question, the conversation will end, but you can choose whether or not to query an instruction or offer a suggestion about what to do next.

Our coding scheme to annotate cooperation is inspired by Davies', though some substantial modifications have been carried out. First of all, we reduced the number of instructions. Reliability tests run on Davies' coding scheme (Davies, 1998) had Kappa scores ranging from 0.69 to 1.0. Despite that, Davies remarked that the coder agreement was not significant for some of the markups. This means either that there was no agreement on those markups or that they are very rare. Moreover, Davies' coding scheme weighting has been modified. In her coding scheme, negative codings are annotated when a particular dialogue behavior that should have been used is absent. An annotation of such complex behaviors is very difficult to pursue. We realized that attributing negative codings to the absence of felicitous dialogue acts was too much challenging for coders without a specific training. Our cooperation typology is similar to the HCRC dialogue moves coding. In our coding scheme we used *check*, *question answering* and *giving instruction* as measures of knowledge sharing (i. e. grounding) between the two speakers. *Check* instruction covers query and objection categories, as our focus is not on dialogue move form but on its function. In a Map Task setting a question is a way to check the extent of knowledge shared. Another group of dialogue moves are related to question answering. A check move can be answered with a yes or no (*Question answering (Y/N)* or some information can be added (*Question answering + adding information)* or an instruction can be repeated as a *check* move follow up (*Repeating Instructions*). The last group of dialogue moves concerns *giving instruction*. It is considered the task baseline. Linked to this move, there are *Acknowledgment* and *Spontaneous info/description adding* dialogue moves. *Acknowledgment* move is coded for every utterance which minimally shows that the speaker has heard the move to which it answers (back-channels are also considered acknowledgments). *Spontaneous info/description adding* introduces new information relevant to the task.

As regards uncooperative dialogue moves, we coded when a speaker fails or refuses to answer a question, add information or repeat an instruction when required by the other speaker. The code instructions are *No answer to question* (no answer given when required), *Inappropriate reply* (failure to introduce useful information when required) and *No Spontaneous Add/Repetition of Instruction* (information is not added or repeated when required). Concerning the effort weighting system, we propose to attribute positive and negative weights in an ordinal scale from +2 to -2. This weighting system, called cooperation level, takes into account the level of effort gained in each move. As in Davies' coding scheme, the lowest value (-2) was attributed when a behavior requiring a minimum effort did not take place, while the highest negative value (-1) was attributed when a high effort behavior did not occur. On the other hand the lowest positive weighting value (1) is attributed when minimum effort moves take place in the dialogue while the highest positive weighting attribute (2) is scored when a high effort behavior appears. We also attribute a weight of 0 for actions which are in the area of "minimum needs" of dialogue, such as in this particular task giving instructions (Table 2).

As a result of the nature of the Map Task, where Giver and a Follower have different dialogue roles, we had two slightly different versions of the cooperation annotation scheme. In particular, *giving instruction* was present only when annotating the Giver cooperation.

4.3 Turn Taking and Gaze Direction Annotation

The other two important dialogue indexes we codify in our coding scheme are the dialogue turn management and the presence or absence of eye contact through gaze direction.

4.3.1 Turn Management Annotation Scheme

The term *turn management* denotes a system that has the purpose of managing the flow of interaction, minimizing overlapping speech and pauses (Yngve, 1970; Sacks, Schegloff & Jefferson, 1974; Goodwin, 1981). Turn management is quite systematic in Map Task dialogues, probably because there are only two participants. Duncan (1972: 283-284) suggested that the turn management mechanism is "*mediated through signals composed of clear-cut behavioral cues, perceived as discrete*"; moreover he pointed out that "*the signals are used and responded to according to rules*". The managing of

Table 2. Example of coding scheme for cooperation annotation

Instructions (Cooperation Typology)	**Cooperation Level**
No answer to question: no answer given when required	-2
Inappropriate reply : failure to introduce useful information when required	-2
No Spontaneous Add/Repetition of Instruction: information is not spontaneously added or repeated after a check	-1
Giving Instructions: task baseline	0
Acknowledgment: a verbal response which minimally shows that the speaker has heard the move to which it responds	1
Question answering (Y/N): Yes-No reply to a check	1
Check: questions (function or form) which solicit other understanding of information already offered	1
Repeating Instructions: repetition of an instruction following a check	1
Question answering + *adding information*: Yes-No reply + new information introduction	2
Spontaneous info/description adding: introduces new information relevant to the task	2

conversational turn regulates the interaction flow and minimizes speech and pauses overlapping. In a recent paper Stivers et al. (2009) tested whether turn taking is a universal system taking into account functional yes–no questions in 10 languages of different part of the world. The response time - that is to say the time elapsed between the end of the question turn and the beginning of the response turn - was calculated in both vocal and gesture modalities. The findings suggested a strong universal basis for turn-taking behavior in yes-no questions, as a minimal overlap between turns was found. In contrast to these claims of a universal system *minimal-gap minimal-overlap*, there are arguments in favor of a culturally variable turn taking system. There are many systematic reasons for the occurrence of a turn overlap, including competing for an early start for the next turn, projection of possible completion or transition-relevance places (Sacks, Schegloff & Jefferson, 1974). Yuan, Lieberman and Cieri (2007) randomly selected four conversations from the English, Japanese, and Mandarin CallHome corpora, which contain phone conversations between family members and friends. They focused on two types of speech overlaps: 1. one speaker takes over the turn before the other speaker finishes (turn-taking type); 2. One side speaks in the middle of the other side's turn (backchannel type). Even though this study was based on an automated (and extremely coarse) binary division of overlap types, Yuan, Lieberman and Cieri had interesting preliminary results. They found that females made more speech overlaps of both types than males; and both males and females made more overlaps when talking to females than talking to males. The speakers also made fewer overlaps when talking with strangers than talking with familiars, and the frequency of speech overlaps was significantly affected by the conversation topic. Moreover, the two conversation sides were highly correlated on their frequencies of using turn-taking type of overlaps but not backchannel type.

From a computational point of view, turn management is generally coded by the three general features: *Turn gain, Turn end* and *Turn hold*. An additional dimension entails whether the speakers both agree upon a change in conversation turn. Thus, a turn gain can either be classified as a Turn take if the speaker takes a turn that wasn't offered, possibly by interrupting, or a Turn accept if the speaker accepts a turn that is offered. Similarly, the end of a turn can also be achieved in different ways: there is Turn yield when the speaker releases the turn under pressure or a Turn offer if the speaker offers the turn to the interlocutor.

According to Duncan, in conversation back-channel cues are also used. In Duncan's proposal, back-channel cues are considered as an alternative to turn-taking; this is because in Duncan's perspective back-channels are reasonably 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. Therefore, we included back-channels in our turn management annotation scheme as a separate category.

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). For example, short verbal back-channel expressions such as *sí* in Italian used to show agreement can be produced with a rising intonation or with non-verbal feedback expressions, such as head nods (Cerrato, 2004). Both Yngve (1970) and Duncan (1972) included head nods as a typical example of back-channels in their descriptions. Maynard (1986) analyzed head nods in Japanese dyadic conversations. He noticed that the main function carried out by the numerous head nods produced during those conversations was back-channeling.

The annotation categories for turn taking have as follows:

- **Turn giving/offer**: the speaker gives/offers the conversational turn to the interlocutor. This is usually marked by the intonation contour or the presence of a pause;

- **Turn accept**: the speaker accepts a turn that is being offered/given and starts talking;

- **Turn yielding**: the speaker can release the turn under pressure of the other speakers;

- **Turn holding**: the speaker holds his/her conversation turn even if under pressure of the other speaker. Usually turn is held with speech sounds or word repetition;

- **Turn taking**: the speaker take a turn that wasn't offered, possibly by interrupting the other speaker.

- **Back-channel:** (includes coding of head nods or head shakes) any verbal or non verbal response which minimally shows that the listener has heard or (dis)agreed with the speaker. It should be noted that a back-channel does not change who is currently in control of the dialogue at the moment.

4.3.2 Gaze Direction Annotation

Another important cue to classify turn segments is gaze (Taylor & Cameron, 1987; Levinson, 2006). Research showed that in the western culture when a listener intended to take turn she/he pulled away her/his gaze, which was typically directed at the speaker's face up until the turn release (Duncan, 1972). Moreover, the annotation of the gaze of the participants can give us a general indication of where the attention of the speaker is focused (Carletta, 2007) in a particular moment: the map, the interlocutor or other events.

In the attempt to annotate gaze direction we adopted categories from both AMI (Carletta, 2007) and EmoTV (Martin et al., 2006) coding schemes. Gaze direction was codified as follows:

- **to the interlocutor**: one of the speakers is looking at the other, usually in the area of the face. In case, this includes eye contact;

- **to the map**: the speaker is looking at the map laying on the desk in front of her/him;

- **unfocused:** when the speaker’s glance is not focusing on anything or anybody in particular;

- **side-turn**: when the speaker looks on the side;

- **waggle:** when the speaker's glance moves from side to side.

Gaze direction was codified in both short and full screen conditions. In the latter, the two speakers can't see each other as they were separated by a screen but we hypothesized that in spite of this, the speakers would still look at each other.

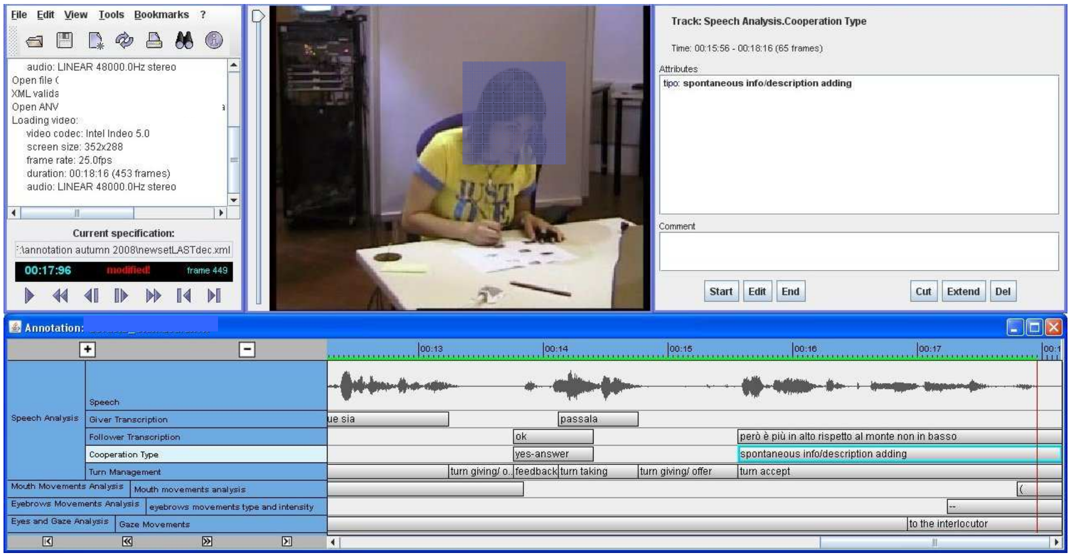
4.4 RECC Coding scheme reliability

In this section we will show and discuss the results of the RECC coding scheme validation. As we stated in Chapter 2, a widespread debate on coding scheme validation is still ongoing after the Carletta's work (1996). For the first time the author dealt with the problem of reliability and chance agreement of annotated data. Validating a coding scheme is the first, essential step to avoid subjective judgments and assure research reliability and reproducibility of the results. In Chapter 3, I argued that multimodal corpora validation is a very demanding task because of the nature of the multimodal data - multimodal communication channels such as face, gesture and body posture are deeply interconnected with each other. Further they contribute -case by case with different weights- to the final meaning of the multimodal "sentence". Emotion annotation is also a difficult task. In fact, the categories usually employed to classify emotions – e. g. overarching categories as positive, neutral and negative- are fuzzy. During the past years the debate regarding the interpretation of Kappa scores has been steadily growing. It is characterized by a lack of consensus on how to interpret this value. Some authors (Allwood et al., 2006) considered reliable the Kappa values between 0.67 and 0.8 for multimodal annotation. Other authors accepted as reliable only scoring rates over 0.8. it is clear that it seems inappropriate to propose a general cut off point, especially for multimodal annotation where very little literature on reliability measures has been reported so far. Currently, in this field it is critical that researches start clearly reporting the methods applied to validate their coding scheme - the number of coders, whether they coded independently or not, whether their coding was manual or not - and the scores for each category of their multimodal communication coding schemes.

4.4.1. Annotation Procedure

Up to now, we have annotated 10 tokens of an average length of 100 seconds each (mean 100 s, dv 5,2 s). Six of the videos were from the confederate sessions and 4 videos were from the control sessions. For the confederate session, the videos were taken at the same points that the HR was measured, i.e., at the beginning of the conversation and after the triggering of un-cooperative utterances (see Chapter 3). For the control session, videos were taken again that the HR was measured. All the videos were annotated by six independent coders following the guidelines reported in RECC Annotation Manual (see Appendix). Two coders repeated the annotation after one month to ensure that the annotation was stable across time. All the annotators were Italian native speakers. Only two of them had previous experience as coders.

RECC coding scheme is an xml file that implemented in ANViL software (ANnotation of Video and Language; Kipp, 2001). ANViL is a free video annotation tool that offers hierarchical multi-layered annotation accurate at the frame level. ANViL is driven by user-defined annotation schemes. RECC coding scheme was implemented in an xml file called *specification* file. The annotators' task was to open the video file, the associated xml file and to fill in the empty tags. A sliding menu appears with all the possible categories for a track when the annotator clicks on its tag.



```
    audio: LINEAR 48000.0Hz stereo
Open file (
XML valida
Open ANV                                  a
Loading video:
    video codec: Intel Indeo 5.0
    screen size: 352x288
    frame rate: 25.0fps
    duration: 00:18:16 (453 frames)
    audio: LINEAR 48000.0Hz stereo
```

Current specification: :\annotation autumn 2008\newsetLASTdec.xml

Track: Speech Analysis.Cooperation Type

Time: 00:15:56 - 00:18:16 (65 frames)

Attributes

tipo: **spontaneous info/description adding**

Giver Transcription	ue sia; passala
Follower Transcription	ok; però è più in alto rispetto al monte non in basso
Cooperation Type	yes-answer; spontaneous info/description adding
Turn Management	turn giving/ o..; feedback; turn taking; turn giving/ offer; turn accept
Mouth movements analysis	(
eyebrows movements type and intensity	--
Gaze Movements	to the interlocutor

Figure 1: Screenshot of RECC coding scheme implemented in ANViL

appeared. After playing the single video, the annotator can select the suitable category and press ok to see it on the screen.

4.4.2. Kappa Statistic Results

A Kappa statistic (Siegel & Castellan, 1988) was performed on the annotations to assess their reliability. We chose Kappa statistic as it is the suitable measurements when chance agreement is calculated for more than two coders. In this case, the agreement is expected as the single distribution which reflects the combined judgments of all coders. Thus, expected agreement is measured as the overall proportion of items assigned to a certain category *c* by all the *n* coders.

In Tables 3 and 4, Kappa scores of the Givers' and the Followers' cooperation analysis are reported. It should be noted that all the features adopted to annotate the Givers' cooperation had a high positive score. This means that all the features were used by all the annotators and the annotators agreed on the labeling of the same elements with the same features.

Table 3. Givers' Cooperation Kappa Scores

Raters = 6
K = 0.816
p< 0.001	

	K and p value
No answer to question	0.817; 0.001
Inappropriate reply	0.959; 0.001
No Spontaneous Add/Repetition of Instruction	0.766; 0.001
Giving Instructions	0.856; 0.001
Acknowledgment	0.797; 0.001
Question answering (Y/N)	0.881; 0.001
Check	0.797; 0.001
Repeating Instructions	0.827; 0.001
Question answering + adding information	0.784; 0.001
Spontaneous info/description adding	0.780; 0.001

While Kappa scores for Givers' cooperation annotation were high and well supported by low p values, yet we identified a high disagreement on the attribution of the categories -*Inappropriate reply* and *Repeating Instructions*- releted to the Followers' cooperation, as shown in Table 4. The high p value (p<.915) suggested that only a minority of the raters decided to assign this label to the Follower annotation. This could be attributed to the nature of the task. The Follower seemed to be more cooperative than the Giver, who has a dominant role giving the instructions. This difference could also be interpreted as the Followers' intention not to sidetrack the Givers with unrelated replies.

Table 4. Followers' Cooperation Kappa Scores

Raters = 6
K = 0.829
p< 0.001	

	K and p value
No answer to question	0.713; 0.001
Inappropriate reply	-0.005; 0.915
No Spontaneous Add/Repetition of Instruction	0.796; 0.001
Acknowledgment	0.803; 0.001
Question answering (Y/N)	0.826; 0.001
Check	0.817; 0.001
Repeating Instructions	-0.005; 0.915
Question answering + adding information	0.713; 0.001
Spontaneous info/description adding	0.878; 0.001

With reference to turn management (Table 5), turn giving had the lowest Kappa score (0.543) compared to the other annotation features whose values or score were above 0.7. This could be a result of the emotive aspect of a large part of the annotated dialogues. Emotion elicitation is likely to prompt overlapping speech. Therefore, annotators had difficulties to find out when the speaker was actually releasing the conversational turn. Observing turn taking could be very informative about how one is feeling or how one wants to be perceived. Although turn taking has been studied widely, the relation between turn taking and emotions has been much less so. In a recent study, Maat and Heylen (2009) examined at how some basic choices in the management of turns influenced the impression of personality, emotion expression and interpersonal stance modifying turn-taking strategies.

Table 5. Turn management Kappa Scores

Raters = 6
K = 0.784
p< 0.001	

	K and p values
Turn holding	0.805; 0.001
Turn yielding	0.841; 0.001
Turn accept	0.778; 0.001
Back-channel	0.887; 0.001
Turn giving	0.543; 0.001
Turn taking	0.753; 0.001

The authors tested how it was possible to create different impressions of friendliness, rudeness and arousal by varying the timing of start or end of a speakers' turn with respect to the start or end of the interlocutor's speech turn.

Gaze direction annotation (Table 6) was reduced to two features: *to the map* and *to the interlocutor*. All the other categories showed high disagreement among coders. Qualitatively, even in the full screen condition (that is to say when a screen blocks the speakers’ eye contact) the two speakers were searching for eye contact.

Finally, the Kappa scores of both eyebrows (Table 7) and mouth (Table 8) annotations were above 0.8. From a closer inspection of the single annotation feature, we discovered that eyebrows annotation features denoting a higher valence (such as +*frown*) had a lower Kappa score. The lower agreement on these features denoted a difficult in perceiving a significant difference between the valence levels or a significant difference between the upper and lower face configurations. This observation was confirmed by the other annotation schemes as well. For example, in FACS every action unit could be classified within a five point intensity scale, ranging from low intensity to extreme intensity. Six coders that analyzed intensity in 19 pictures reached an agreement on 55% of the pictures. However, the data was not corrected for chance agreement. This means that the unbiased agreement surely was much lower than the one reported.

Table 6. Gaze Direction Kappa Scores

Raters = 6
K = 0.788
p< 0.001	

	K and p value
Down / to the map	0.865; 0.001
To the interlocutor	0.889; 0.001
Side-turn	-0.006; 0.906
Unfocused	0.466; 0.001
Waggle	0.189; 0.001

As regards the mouth annotation scheme, we observed a similar pattern: the agreement was high on the annotation of both positive valence levels (smile and open smile/laugh), while asymmetric smile annotation had a high disagreement among coders. As in the previous case, this category was picked up only by few annotators.

Table 7. Eyebrows Configuration Kappa Scores

Raters = 6
K = 0.855
p< 0.001	

	K and p value
Normal position **--**	0.962; 0.001
frown	0.841 ; 0.001
+frown	0.588 ; 0.001
Up	0.788 ; 0.001
+up	0.544 ; 0.001

All the annotation categories had Kappa scores above 0.7. Thus, our coding scheme has a very high reliability. Nevertheless, some features had a negative Kappa score and a high p value. In future annotations, these features should be discarded from the coding scheme. As regards the features with low p value and Kappa scores under 0.7, we should check in future annotations whether the dataset we annotated was too small to test coder agreement for those specific features or those features were not relevant for RECC.

Table 8. Mouth Configuration Kappa Scores

Raters = 6
K = 0.805
p< 0.001	

	K and pvalue
Grimace **(**	0.904; 0.001
Smile **)**	0.750; 0.001
Open smile/ laugh **+)**	0.762; 0.001
Asymmetric smile **1up**	-0.005; 0.910
closed lips/ normal position **-**	0.867; 0.001
Lower lip biting	0.904; 0.001
Open mouth **O**	0.753; 0.001

4.5 Corpus Public Releases

To the date, RECC can be ordered by email to: federica.cavicchio@gmail.com. The psychopsysiological recordings are not publicly available, since the HR and the skin conductance data were collected and analyzed with Acknowledge®, a Biopac's licensed software. Videos are divided by recording conditions - confederate, control, high screen and low screen - and task role – Giver or Follower. At the following link www.clic.cimec.unitn.it/RECC one can find reports of the documentations on the corpus collection methodology and the coding scheme. The RECC annotation manual is available at http://www.clic.cimec.unitn.it/RECC, together with an XML file consisting of the ANViL specification file of the scheme.

4.6 Conclusion

RECC is a unique resource considering the way it was collected and the phenomena it challenged. It is the first multimodal corpus that includes audiovisual recordings aligned with psychophysiological data. RECC was built with the purpose to investigate linguistics, pragmatics and emotions in a dialogue setting. Our expectation is that researchers will obtain from the RECC elicitation method and the RECC annotation scheme a range of features that are necessary for the progress in the domain of multimodal dialogue studies.

Our coding scheme reliability was very high when compared with other multimodal annotations. This is because we analyzed cooperation and emotion using a coding scheme based on the decomposition of the several factors underlining an emotion. In particular, we did not refer to emotive terms directly. In fact every annotator had his/her own representation of a particular emotion, which could be different from the one of another coder. This representation can be a problem especially for the annotation of blended emotions, which are ambiguous and mixed by nature. As some authors argued (Colletta et al., 2008) annotation of mental and emotional states is a very demanding task. In general, the analysis of non verbal features requires a different approach compared with other linguistic tasks. This is because multimodal communication has multiple semantic levels. For instance, a facial expression can deeply modify the sense of a sentence, such as in humor or irony. Furthermore, RECC coding scheme is partially derived from cognitive and neuroscience corroborations. The reliability test we run had confirmed the usefulness of our annotation features. RECC coding scheme is an important step towards the creation of annotated multimodal resources which are crucial to investigate multimodal communication. Particularly, RECC coding scheme can aid exploring how different emotive sets (positive or negative) modify cooperation in different cultural settings; how turn management and sequencing strategies are expressed in different cultural settings; how gaze can enhance or disrupt cooperation; how emotions modifies the multimodal communicative channels.

Corpora annotated according to the RECC coding scheme represents useful resources to model back-channel, turn management and facial expressions of multimodal agents. Given these premises and the results we obtained, we consider promising the implementation of cognitive and neuroscience evidence in computational coding schemes. Our findings will be hopefully taken into account in order to guide the design of Human Computer Interfaces.

Chapter 5

The Predictors of (un)Cooperation

In the previous Chapter we showed the reliability assessment of the Rovereto Emotion and Cooperation Corpus (RECC) annotation. In this Chapter, two studies investigated the emotive and non verbal predictors of cooperative and uncooperative behaviour. In the first study, we tested via a linear regression the relationship between cooperation and Heart Rate (HR). In the second study, we tested via a logistic regression model whether facial expressions, gaze, asymmetries in knowledge and gender predict cooperation.

5.1 Emotion and Cooperation

5.1.1 Heart Rate and Cooperation

RECC is a task-oriented corpus with psychophysiological data registered and aligned with audiovisual data. RECC was collected to elucidate the relationship between cooperation and emotions. This corpus allows to clearly identify emotions and their facial expression in a dialogue setting. RECC is, to our knowledge, the first corpus having audiovisual and psychophysiological data recorded and aligned together.

One of RECC core hypotheses is that there is a negative correlation between psychophysiological measures such as HR and cooperation. So, once the data from the corpus annotation was validated and then considered reliable, this hypothesis was tested by a linear regression (Fig. 1). The mean HR of each analysed corpus segment was correlated with the normalized cooperation for that segment. To calculate it we ascribed to each cooperation type the corresponding cooperation level (see Chapter 3). Cooperation levels were normalized by dividing the sum of the cooperation “weights” for all the moves during a particular video by the number of moves in that video. The normalized cooperation (*cooperation_norm* in Fig. 1) was negatively correlated with HR (R2 = 0.85, p <0.001, S.E. 0.03). The control condition data was concentrated at the high-cooperation, low-heart rate top-left corner of the diagram, whereas each subsequent utterance in the confederate sessions was associated with higher heart rate and lower cooperation. Interestingly, cooperation levels never became negative; this suggests that a recovery of cooperation took place soon after the confederate's provocation. Qualitatively, our recordings showed that uncooperative utterances produced by the confederate had as a result a significant lack of cooperation: after triggering negative emotions the Follower did not answer to questions whereas the Giver stopped answering to questions or gave replies which were not relevant to the task.

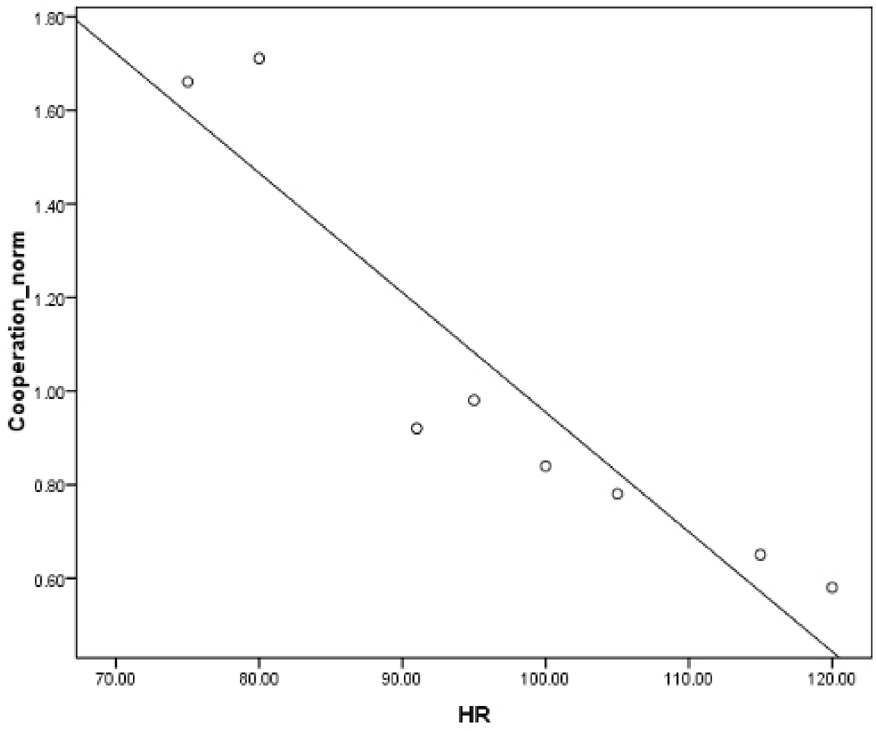


Figure 1: Negative linear correlation between HR and normalized linguistic cooperation (cooperation_norm)

5.1.2 Facial Expressions and Cooperation

The second hypothesis we tested is whether non verbal expressions of negative emotions (i. e. a particular facial expression) would predict cooperative or uncooperative communication. Previous studies showed that visual access to non verbal behaviour fostered a dyadic state of rapport which facilitated mutual cooperation (Argyle, 1990; Tickle-Degnen & Rosenthal, 1990). However, these studies did not establish which facial cues were predictive of cooperation. As cooperation is a discrete dependent variable (i. e. it takes only a limited number of values), our hypothesis was tested via a logistic regression[3].

In order to investigate which facial expression was more likely to predict cooperation, cooperative and uncooperative scores were shrunk to two levels, 0 for uncooperative behaviour and 1 for cooperative behaviour. The upper and lower face configurations observed during each cooperative and uncooperative interaction were included in the model. Cooperation levels equal to 0, assigned for *giving* i*nstruction* according to RECC coding scheme, and the corresponding facial expressions were discarded from the data set. High or short screen condition (i. e. whether the two speakers can see each other), role in the interaction (Giver or Follower) and gender (male or female) were included in the model as well. Our model used 11 explanatory variables which were considered likely to influence cooperation. Interaction effects were investigated but not found. To evaluate the model fit, we run a Wald z-statistic, establishing as significant upper ($p<0.0001$) and lower ($p<0.0001$) face configuration, screen ($p<0.02$) and gender ($p<0.005$). Results are stable across 1000 bootstraps.

[3] Logistic regression (also called *maximum entropy*) is a statistical model commonly applied to predict the probability of occurrence of an event by fitting data to a logistic curve. It is a generalized linear model used for binomial regression. Like many other forms of regression analysis, logistic regression makes use of several predictor variables that may be either numerical or categorical. The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (i. e. the dependent variable) and a set of independent variables (i. e. the predictors). Unlike in linear regression, in logistic regression parameters do not minimize the sum of squared errors but they maximize the likelihood of observing the sample values. The probability of an event will range from 0 to 1. Probability can be calculated with the logarithm of the odds ratio (log odds ratio). In the logistic function, the probability of log odds ratio can take as an input any values from negative infinity to positive infinity, whereas the output is confined to values between 0 and 1. A positive regression coefficient means that that predictor increases the probability of the outcome, while a negative regression coefficient means that the predictor decreases the probability of that result; a large regression coefficient means that the predictor strongly influences the probability of that outcome; while a near-zero regression coefficient means that that predictor has a little influence on the probability of that result.

As a results of the logistic regression model, we found that cooperation (Fig. 2) was predicted by eyebrows in relaxed position (coefficient 2.16, S.E. 1.06, p<0.04). Cooperation was also predicted by a smile (3.44, S.E. 0.78, p<0.0001) or an open smile (6.64, S.E. 1.35, p<0.0001), lips in relaxed position (2.8, S.E. 0.67, p<0.0001) and lower lip biting (3.67, S.E. 0.85, p<0.0001; Fig. 3).

Mouth configurations were found to be more likely to predict cooperation than eyebrows. Precisely, the strongest predictions for cooperation came from open smile/laugh, smile and lower lip biting.

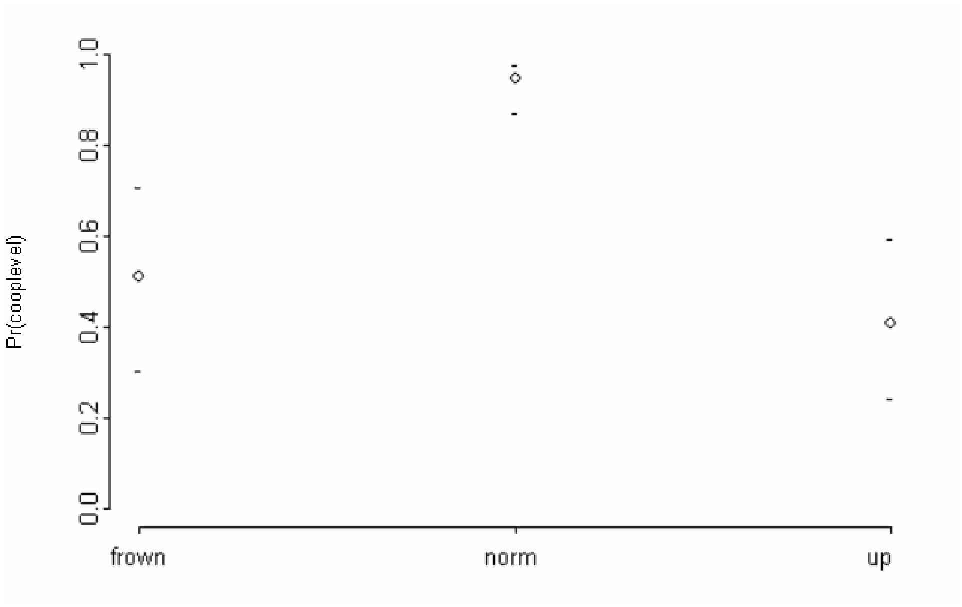


Figure 2: Effects of facial expression (eyebrows configurations) as predictors for log-odds ratios of cooperation (Pr(cooplevel)). Error bars refer to S. E.

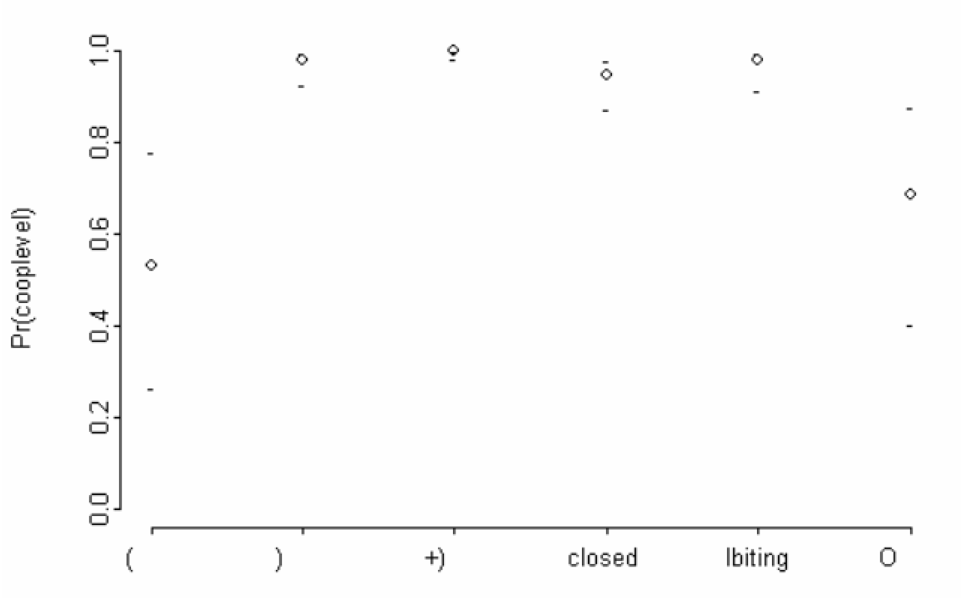


Figure 3: Effects of facial expression (mouth configurations) as predictors for log-odds ratios of cooperation (Pr(cooplevel)). Error bars refer to S. E.

5.2 Gaze, communicative role, gender and cooperation

5.2.1 Gaze and cooperation

One of the central puzzles of human evolution is when and how humans became so cooperative. Humans engage in frequent, large-scale, complex, cooperation with an unprecedented degree among all the animal species (Richerson & Boyd, 2005). Humans seem to be especially inclined, as compared with other primates, to engage with one another in collaborative interactions (Bard & Vauclair, 1984; Tomasello & Carpenter, 2005). In interactions, each participant visually monitors what the other is attending to. In order to facilitate a shared activity, it appears to be an advantage in initiating and maintaining communicative interactions that one’s eyes are easily visible to others.

In many studies, researchers emphasized the monitoring functions of gaze with respect to back-channel. A conversation can be viewed as a collaborative endeavor in which participants contribute to the discourse ensuring that their meanings are mutually understood (Clark & Schaefer, 1987; Clark & Wilkes-Gibbs, 1986). For example, smiles are one of the non verbal behaviors that participants use to indicate the achievement of a mutual understanding (Brunner, 1979; Duncan, 1973; Duncan & Fiske, 1977; Kraus, Fussell & Chen, 1995; Krauss & Morsella, 2000). For the larger part, non verbal behavior can be captured only visually. Therefore speakers would glance frequently at their listeners when information about mutual comprehension would be crucial to maintain the ongoing dialogue (i. e. a back-channel is needed). Another perspective regarding gaze functions during interactions is its role in conversational regulation. Turn taking shifting can be associated with changes in gaze direction: as speakers complete their turns, they are likely to be looking directly at their listeners, and speakers typically begin their turns with gaze averted. Among others, Kendon (1967) suggested that directed gaze informed the listener that the speaker is prepared to relinquish the conversation turn, while averted gaze indicated the opposite. Such a strong interpretation of gaze function was denied by many investigations (Duncan & Fiske, 1977; Duncan, 1972; Duncan & Niederehe, 1974) which claimed that the role of gaze in signaling the end of a conversational turn was minimal. On the other hand, there is evidence that directed gaze served as a signal that elicited back-channel responses (Brunner, 1979; Duncan, 1973; Duncan & Fiske, 1977).

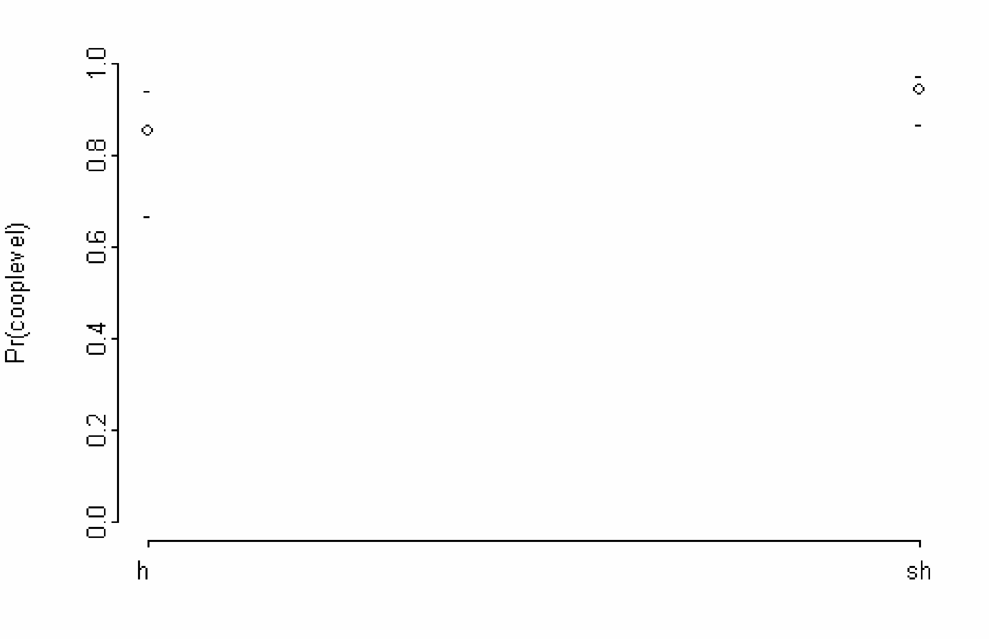


Figure 4: Effects of screen (h= high and sh=short) as predictor for log-odds ratios of cooperation (Pr(cooplevel)). Error bars refer to S. E.

We tested via logistic regression the hypothesis that mutual gaze predicts cooperation. In RECC, half of the interactions had a high screen separating the Giver and the Follower while in the other half a short barrier prevented the speakers from seeing each others' maps. As a result, the short screen condition, which allowed mutual gaze, was found to be a predictor of cooperation (b=1.19, S.E. 0.5, p<0.02).

5.2.2 Conversation Role, Gender and Cooperation

In RECC the Giver and the Follower had distinct communicative role. The Follower is the one that gave the instruction, the one with the route signed on the map. Therefore, one could argue that the Giver have a much more powerful role in the interaction compared to the Follower since the former has more information than the latter. If this is the case, than the Giver has the dominant role in the interaction. Asymmetry of information can affect cooperation (Markova & Foppa, 1991; Drew, 1992). Specifically, one could argue that the main responsibility to achieve a good task result should be attributed to the Giver as she/he holds more information than the Follower.

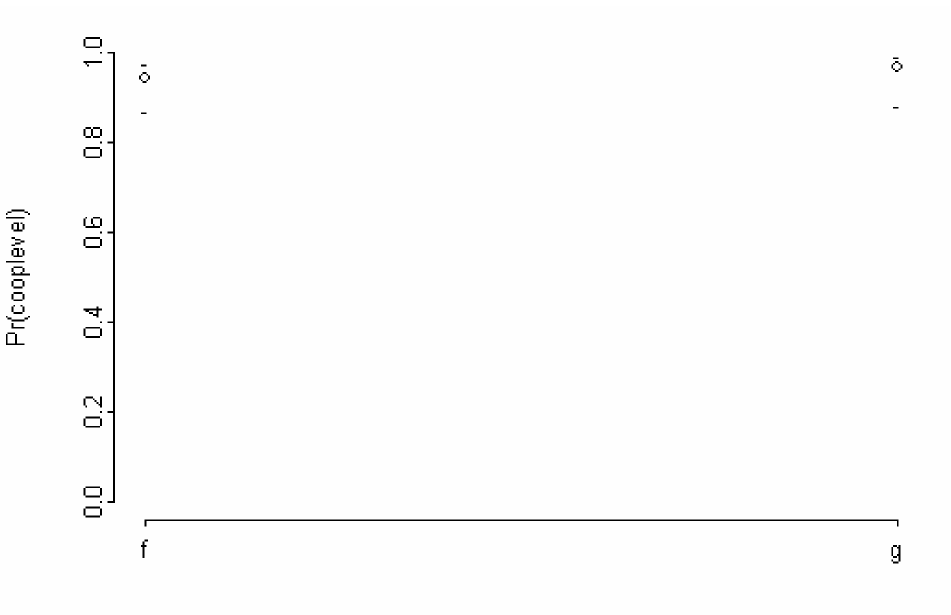


Figure 5: Effects of interaction role (f=Follower and g= Giver) as predictors for log-odds ratios of cooperation (Pr(cooplevel)). Error bars refer to S. E.

In the logistic regression model we tested whether the interaction role (Giver or Follower) is a predictor of cooperation. As a result, role was found not to be significant (0.15, S.E. 0.28, p<0.7). As regards dominance, we can conclude that asymmetry in communication roles was not a predictor of cooperation in our corpus.

Another important variable affecting cooperation is gender. In RECC the variable gender was counterbalanced in all the recordings. In the confederate condition the confederate was the same for all the interactions. Our confederate was a male, and his matched companion could be either a male or a female. We counterbalanced gender condition in all the confederate interactions so that the 12 participants that interacted with the confederate were consisted of 6 females and 6 males.

Although the relationship between language and gender is complex, and the relevant evidence on that relationship difficult to develop and interpret, it seems clear that discourse variables (e.g. tag questions, politeness formulas, patterns of turn taking, use of minimal responses, etc.) in both western (Zimmerman & West, 1975; West & Zimmerman, 1983; Fishman, 1983; Holmes, 1983, 1984, 1986; Coates, 1988; Nordenstam, 1992) and non-western (Brown, 1980; Ide, 1982; Smith, 1992) cultures often showed that women are more cooperative and men are more confrontational/competitive. Such patterns have been interpreted as an indication that women are "politer" and more concerned with the solidarity building aspects of dialogue than men. This explanation located the source of the differences in gender. However, this is not the only possible interpretation. It might be that men tend to use more non-standard forms because their social networks are tighter. This interpretation was supported in particular by Milroy's (1987) study in which women in tighter social networks are more inclined to use non-standard forms than standard ones. Also, some studies showed facilitative behavior correlated with the interaction role, and not necessarily with gender (O'Barr & Atkins, 1980; Cameron, McAlinden & O'Leary, 1988). Freed and Greenwood (1996) reported minimal differences in the use of pragmatic devices between male and female when engaged in similar activities. Such observation suggested that cooperation differences in conversation lie in the interaction role.

In the logistic regression model we tested whether gender was a predictor of cooperation. As a results, males were found to be predictive of a decrease in cooperation (-0.8, S.E. 0.27, p<0.005).

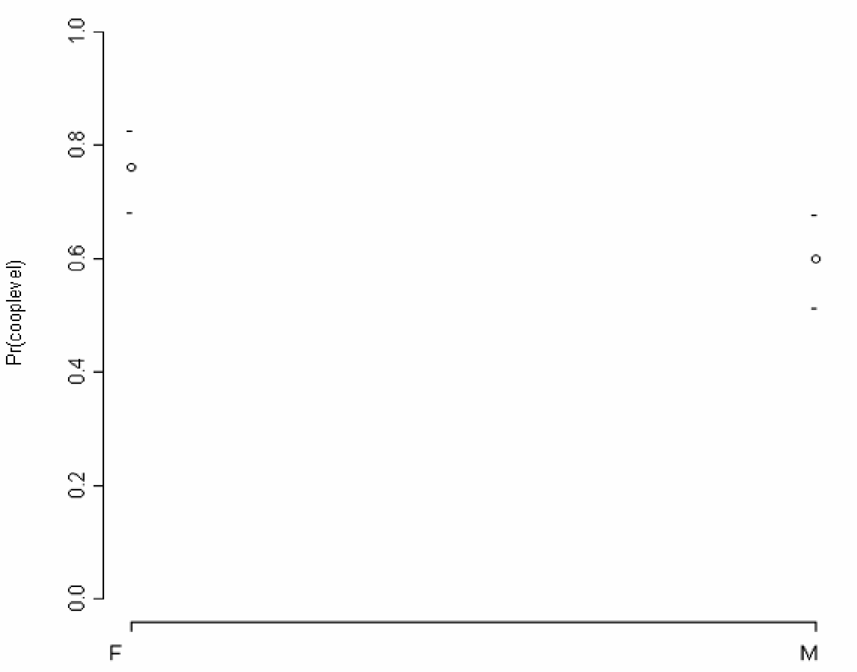


Figure 6: Effects of gender (F=female and M= male) as predictors for log-odds ratios of cooperation (Pr(cooplevel)). Error bars refer to S. E.

5.3 Discussion and Conclusion

In this Chapter we tested whether the confederate's uncooperative scripted utterances reduced the level of cooperation in the other speaker. To test this, we measured the level of cooperation in the utterances following the non-cooperative triggering. The predictive variables were HR and the facial expression. Thus, our hypothesis was translated into two following subhypotheses: HR and facial expressions could predict non-cooperation. As regards the first hypotheses, we demonstrated via a linear regression that moderate heart rate values predicted higher levels of cooperation whereas each subsequent utterance in the confederate sessions was associated with higher heart rate and lower cooperation. Interestingly, cooperation levels never became negative. This finding suggests that a recovery of cooperation took place soon after the confederate's provocation.

The hypotheses that facial expression predicts cooperative and uncooperative behaviour was tested via a logistic regression model which took into account cooperation weights as dependent variable and facial expressions, gaze, interaction role and gender as the predictors. We find out that cooperation was highly predicted by three lower face configuration - smile, open smile and lip biting- and one upper face configuration - eyebrows in the normal position.

Our results are very interesting with respect to computational models usually adopted to explain facial expression recognition. Basically, the two models applied to categorize facial expressions are Principal Component Analysis and Non negative Matrix Factorization. As cooperation is predicted by a group of mouth configurations and by only one eyebrow configuration, one can argue that the mouth is the component to look at to reliably predict cooperation. However, the upper face predictor of cooperation (eyebrows relaxed/in neutral position) can give us some information too. As smile and lower lip biting are associated with neutral eyebrows position, one could argue that the facial expressions displayed have more to do than with emotions. A number of studies pointed out that a facial expression that are seemingly emotional often do not indicate the expressive individual's emotional state but served as strictly communicative functions (Fridlund, 1994). Facial expressions frequently represent a way for communicating empathy (Bavelas, Black, Lemery & Mullett, 1987) or other communicative functions transcending simple indications of one' s current feelings. In this view, a smile could effectively communicate empathy and lip biting could be related to expressing difficulty pursuing the task. As regards the uncooperative behaviour, none of the facial patterns could predict it. This is consistent with the notion that emotional expression is differentially driven by the results of sequential appraisal checks, as postulated by the componential appraisal theory (Sherer & Heiner, 2007). In this view, facial expressions are not “readout” of motor programs but indicators of mental states and evaluation processes. Moreover, a recent study (Van Mechelen & Hennes, 2009) demonstrated that personal differences had very different response to externally induced disadvantage. In the confederate's Map Task a specific disadvantage was elicited to induce frustration and anger. Yet, for some of the participants, to be effective the thwarting should be characterized by norm violation. Moreover, the unease must be appraised as unfair and deliberate in order to experience anger. Combined, our findings demonstrated that an emotion with high arousal and unpleasant evaluation can occur in combination with different patterns of appraisals. Appraisal varied as a function of the communicative situation and the person's characteristics.

As regards gaze, our data confirmed that seeing the other speaker's face during a dialogue is a predictor of cooperation. Moreover, the model confirmed that Map Task dialogues are an equal enterprise. Based on the Kappa statistics data, the Follower was more cooperative than the Giver. As both Giver and Follower were not predictors of cooperation there was a joint responsibility for a good task result, instead of it being the main responsibility of the Giver's. Finally, evolutionary scientists argued that human cooperation is the product of a long history of competition among rival groups. There are various reasons to believe that this logic applies particularly to men so that they are in general less cooperative than females are. Our model confirmed that in a dialogue setting males were more likely to exhibit uncooperative behaviour compared to females.

Chapter 6

Conclusion

In this research, uncooperative communication and cooperation were investigated by recording cooperative and uncooperative dialogues between two speakers. Cooperative behavior and its relationship with emotions is a topic of great interest in the field of dialogue studies. In our study we considered a range of features that had never been deeply analyzed before, but that are necessary to the progress in the domain of dialogue studies. Our initial hypothesis was that a negative emotion elicitation would lead to a reduced level of cooperation in the other participant. To test this, the level of cooperation following each negative emotion elicitation was measured. In line with the appraisal theory of emotion, we collected the psychophysiological recordings, the ratings of pleasantness/unpleasantness of the situation, and the corresponding facial expressions of each participant. A further condition was added: in half of the interactions a screen divided the two speakers, so that they could not see each others' faces. The corpus we collected, named RECC, is a unique resource in both the way it was collected and the phenomena that it challenged. RECC is the first multimodal corpus with audiovisual recordings aligned with psychophysiological data.

As regards emotion annotation, our coding scheme was partially derived from cognitive and neuroscience corroborations. The reliability test we run confirmed the usefulness of the chosen features. RECC coding scheme constitutes an important step towards the creation of an annotated multimodal resource to investigate several aspects of human communication.

One of the main goals of this thesis was to detect the predictors of cooperative and uncooperative behavior in a task-oriented dialogue setting. Using data from RECC, we addressed the following five research questions:

- Research question 1: *are psychophysiological measures, specifically heart rate, predictors of cooperation?*

A negative linear regression between cooperation and heart rate was found. This result provided support for the hypothesis that during an interaction negative emotion elicitation and uncooperative utterances would reduce the level of cooperation in the other participant.

- Research question 2: *is facial expression a predictor of cooperation?*

Upper and lower facial expressions were investigated as predictors of cooperation. Eyebrows in relaxed position, smile, open smile and lip biting were found to be strong predictors of cooperation. Surprisingly, uncooperative behavior had no facial predictors. Though the arousal value and the post rating test showed that emotions appraised by the participants could be attributed to the area of anger, it seemed that the emotion displaying is attenuated or masked. Consequently, none of the facial expressions predicted uncooperative behavior. These results are very interesting in relation to the computational models usually adopted to recognize facial expression of emotions. Most of the algorithms have been tested on static pictures of basic emotions. According to our data, cooperation is predicted by a group of mouth configurations and only one eyebrow configuration. Therefore, one can argue that mouth configuration is the principal component to be taken into account when predicting cooperation. Nevertheless, as smile and lower lip biting are associated with a neutral eyebrow position, the facial expressions displayed might be related with communicative functions or empathy rather than with emotions. Moreover, our results are consistent with the notion of emotional expression postulated by componential appraisal theory. In this view, facial expressions are not functions of motor programs but are evidence of mental states and evaluation processes.

- Research question 3: *is eye contact a predictor of cooperation?*

In order to start and maintain an interaction it seems to be an advantage that the speaker's eyes are easily visible to the other speaker. In our corpus, half of the interactions had a high screen that separated the Giver and the Follower. In the other half, a short barrier allowed the speakers to see each others' faces and at the same time prevented the participants from seeing each others' maps. In the logistic regression model short screen condition, enabling mutual gaze, was found to be a predictor of cooperation.

- Research question 4: *Does dominance affect cooperation?*

The logistic regression model confirmed that Map Task is an equal enterprise. Nor the Giver neither the Follower was a predictor of cooperation. Therefore, our data confirmed that the task result was a speakers' joint responsibility rather than the Giver's main job.

- Research question 5*:* *is gender a predictor of cooperation?*

In many studies on both western and non western cultures women were considered more cooperative and men more confrontational/competitive. Our model confirmed that in a Map Task dialogue setting males are more likely to be more uncooperative than females.

Given the premises that RECC coding scheme started and the results we achieved, we consider promising the implementation in computational coding schemes of cognitive and neuroscience inspired features. Our results shed light on a crucial aspect of communication, and the methods we adopted can be used to investigate and model other aspects of human interaction. Corpora annotated according to the RECC coding scheme will represent a useful resource for model feedback, turn management and facial expressions of multimodal agents. Furthermore, our findings are utilizable for the design of Human Computer Interfaces.

Bibliography

Abrilian, S., Devillers, L., Buisine, S., & Martin, J.-C., (2005). EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. In *Proceedings of Human-Computer Interaction International*, Las Vegas, USA.

Allen, J. F. & Perrault, C. R., (1980). Analysing intention in utterances. *Artificial Intelligence*, 15, 143-178.

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, *Language Resources and Evaluation*, 41, 273-287.

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, J.-C., Kühnlein, P., Paggio, P., Stiefelhagen, R., & Pianesi, F. (eds.) *Multimodal Corpora: From Multimodal Behavior Theories to Usable Models*, 38-42.

Anderson, J. C., Linden, W., & Habra, M. E., (2005). The importance of examining blood pressure reactivity and recovery in anger provocation research. *International Journal of Psychophysiology*, 57,159-163.

Anderson, A.H., & Boyle, E.A., (1994). Forms of introduction in dialogues: Their discourse contexts and communicative consequences. *Language and Cognitive Processes*, 9, 101-122.

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., (1990). The Biological Basis of Rapport. *Psychological Inquiry*, 1, 296 - 300.

Argyle, M., & Cook, M., (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Athanaselis, T., Bakamidis, S., Dologlou, I., Cowie, R., Douglas-Cowie, E., Cox, C., (2005). ASR for emotional speech: clarifying the issues and enhancing performance, *Neural Networks*, 18, 437- 444.

Axelrod, R., Hamilton, W. D., (1981).The Evolution of Cooperation. *Science*, 211, 1390- 1396.

Bard, K.A. & Vauclair, J., (1984). The communicative context of object manipulation in ape and human adult-infant pairs, *Journal of Human Evolution,* 13, 181–190.

Bargiela-Chiappini, F., & Harris, S., (1997). *Managing Language: The discourse of corporate meetings*. Amsterdam: John Benjamins.

Bartko, J. J., & Carpenter, W. T. Jr. (1976). On the methods and theory of reliability. *Journal of Nervous and Mental Disease*, 163, 307–317.

Bavelas, J. B., Black, A., Lemery, C. R., & Mullett, J., (1987). Motor mimicry as primitive empathy In Eisenberg N. & Strayer J., (eds.), *Empathy and its development* (pp. 317-338). Cambridge, UK: Cambridge University Press.

Berger, C. R., & Bradac*,* J. J., (1982*). Language and social knowledge: Uncertainty in interpersonal relations*. London: Edward Arnold Publishers.

Boyd, R., & Richerson, P. J., (2005). Solving the Puzzle of Human Cooperation, In: S. Levinson (ed.) *Evolution and Culture* (pp. 105–132). Cambridge MA: MIT Press.

Bradley, M. M. & Lang, P. J., (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry*, 25, 49-59.

Brennan, S.E., & Clark, H.H., (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition,* 22, 1482-1493.

Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. A., & Zelinsky, J. C. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, 106, 1465-1477.

Brown, P., (1980). How and why are women more polite: Some evidence from a Mayan commmunity. In McConnell-Ginet, S., Borker, R., & Furman, N. (eds.) *Women and Language in Literature and Society*. (pp. 111-136). New York: Praeger.

Brunner, L.J., (1979). Smiles can be backchannel. *Journal of Experimental Social Psychology*, 37, 728-734.

Bruyer, R., Laterre, C., Seron, X., Feyereisen, P., Strypstein, E., Pierrard, E. & Rectem, D., (1983). A case of prosopagnosia with some preserved covert remembrance of familiar faces. *Brain and Cognition*, 2, 257–284.

Buciu, I., & Pita, I., (2006). NMF, LNMF, and DNMF modeling of neural receptive fields involved in human facial expression perception. *Journal of Visual Communication and Image Representation*, 17, 958-969.

Calder, A. J., Burton, M., Miller, P., Young, A. W., & Akamatsu, S., (2001). A principal component analysis of facial expressions. *Vision Research*, 41,1179-1208.

Callejas, Z. & López-Cózar, R., (2008). Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Communication*, 50, 416-433.

Cameron, D., McAlinden, F., & O'Leary, K., (1988). Lakoff in context: The social and linguistic functions of tag questions. In Coates, J. & Cameron, D. (eds.) *Women in their Speech Communities* (pp. 74-93). London: Longman.

Carletta, J., (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41, 181-190.

Carletta, J., (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22, 249-254.

Carletta, J., (1992). *Risk-taking and Recovery in Task-oriented Dialogue*. Unpublished PhD thesis, University of Edinburgh.

Carletta, J., & Mellish, C., (1996). Risk-taking and recovery in task-oriented dialogue. *Journal of Pragmatics*, 26, 71-107.

Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J.C. & Anderson, A. H., (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23, 13-31.

Cavicchio, F., & Poesio, M., (2008). Annotation of Emotion in Dialogue: The Emotion in Cooperation Project. In *Multimodal Dialogue Systems Perception*. *Lecture Notes in Computer Science* (pp. 233-239). Heidelberg: Springer Berlin.

Cerrato, L., (2004). A coding scheme for the annotation of feedback phenomena in conversational speech. In Martin, J.-C., Os, E.D., Kühnlein, P., Boves, L., Paggio, P., & Catizone, R. (eds.) *Proceedings of Workshop Multimodal Corpora: Models of Human Behavior for the Specification and Evaluation of Multimodal Input and Output Interfaces*, (pp. 25-28). Heidelberg: Springer Berlin.

Clark, H. H., (1996). *Using language*. Cambridge: Cambridge University Press.

Clark, H.H., & Krych, M.A., (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50, 62-81.

Clark, H.H, & Brennan, S.E., (1991). Grounding in communication. In Resnick, L., Levine, J., & Teasley, S. (eds.) *Perspectives on Socially Shared Cognition*. (pp. 127-149). Washington, DC: American Psychological Association.

Clark, H.H., & Schaefer, E.F., (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2, 19-41.

Clark, H.H. & Wilkes-Gibbs, D., (1986). Referring as a collaborative process. *Cognition*, 22. 1-39.

Coates, J., (1988). Gossip revisited: Language in all-female groups. In Coates, J., & Cameron, D. (eds.) *Women in their Speech Communities*. (pp. 110-131). London: Longman.

Cohen, J., (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

Cohen, P.R., & Levesque, H.J., (1990). Intention is choice with commitment. *Artificial Intelligence*, 42, 213-261

Cohen, P.R., Morgan, J., & Pollack, M.E., (Eds.), (1990). *Intentions in Communication*. Cambridge MA: MIT Press.

Colletta, J.-M., Venouil, A., Kunene, R., Kaufmann V., & Simon, J.-P., (2008). Multitrack Annotation of Child Language and Gestures. In Martin, J.-C., Patrizia, P., Kipp, M., & Heylen, D., (eds.) *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*, (pp. 36-45). Heidelberg: Springer Berlin.

Connor, U., & Upton, T.A., (eds.) (2004). *Discourse in the Professions: Perspectives from corpus linguistics*. Amsterdam: Benjamins.

Coupland, N., Giles, H. & Wiemann, J.M., (eds.) (1991). *'Miscommunication' and Problematic Talk*. London: Sage.

Craggs, R., & Wood, M., (2004). A Categorical Annotation Scheme for Emotion in the Linguistic Content of Dialogue. In *Affective Dialogue Systems*, (pp. 89-100) Elsevier.

Davies, B. L., (2006). *Leeds Working Papers in Linguistics and Phonetics 11*, http://www.leeds.ac.uk/linguistics/WPL/WP2006/2.pdf (2006).

Davies, B.L., (1998). *An Empirical Examination of Cooperation, Effort and Risk in Task-oriented Dialogue*. Unpublished PhD thesis, University of Edinburgh.

Devillers, L, Vidrascu, L, & Lamel, L., (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18, 407-422.

Di Eugenio, B., & Glass, M., (2004). The kappa statistic: A second look. *Computational Linguistics*, 30, 95–101.

Donato, G., Bartlett, M.S., Hager, J.C., Ekman, P. & Sejnowski, T.J., (1999).

Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21, 974–989.

Dreber, D. G., Rand, D., Fudenberg, M. A., & Nowak, A. (2008). Winners don't Punish. *Science*, 348-352.

Drew, P. & Heritage, J., (eds.) (1992). *Talk at Work. Interaction in Institutional Settings*. Cambridge: Cambridge University Press.

Douglas-Cowie, E., Devillers, L., Martin, J.-C., Cowi, R., Savvidou, S., Abrilian, S., & Cox, C., (2005). Multimodal Databases of Everyday Emotion: Facing up to Complexity. In *9th European Conference on Speech Communication and Technology* (Interspeech'2005) Lisbon, Portugal, (pp. 813-816).

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 2, 161-180.

Duncan, S., (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology*, 23, 283-292.

Duncan, S. & Fiske, D., (1977). *Face-to-Face Interaction*. Hillsdale, NJ: Erlbaum,

Duncan, S. D., & Niederehe, G., (1974). On signaling that it's your turn to speak. *Journal of Experimental Social Psychology*, 10, 234-247.

Ekman, P., (1992). An Argument for Basic Emotions. *Cognition and Emotion*, 6, 169-200.

Ekman, P., (1984). Expression and the nature of emotion. In K. Scherer and P. Ekman (Eds.), *Approaches to emotion* (pp. 319-343). Hillsdale, N.J.: Lawrence Erlbaum

Ekman, P., Friesen, W. V. (1986). A new pan-cultural facial expression of emotion. *Motivation and Emotion*, 10, 159-168

Ekman, P. & Friesen, W. V., (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, California: Consulting Psychologists Press.

Ekman*, P. &* Friesen*, W. V., (1975*). *Unmasking the face. A guide to recognizing emotions from facial clues.* Englewood Cliffs, New Jersey: Prentice-Hall.

Enfield, N.J., (2006). Social Consequences of Common Ground In Enfield, N. J. & Levinson, S. C. (eds.), *Roots of human sociality: Culture, cognition and interaction* (pp. 399-430). Oxford: Berg.

Etcoff, N.L., & Magee, J.J., (1992). Categorical perception of facial expressions. *Cognition*, 44, 227–240.

Farah, M.J., O'Reilly, R.C., & Vecera, S.P., (1993). Dissociated overt and covert recognition as an emergent property of a lesioned network. *Psychological Review*, 100, 571–588.

Feldman Barrett, L., Lindquist, K. A., & Gendron, M., (2007). Language as Context for the Perception of Emotion. *Trends in Cognitive Sciences*, 11, 327-332.

Fishman, P., (1983). Interaction: The work women do. In Thorne, B., Kramarae, C., & Henley, N. (eds.) *Language, Gender and Society* (pp. 89-101). Rowley, Massachusetts: Newbury House.

Fleiss, J. L., (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P., (2007). The world of emotion is not two-dimensional. *Psychological Science*, 18, 1050–1057.

Freed, A. & Greenwood, A., (1996). Women, men and type of talk: What makes the difference? *Language in Society*, 25, 1-26.

Fridlund*, A., (*1994). *Human Facial Expression. An Evolutionary View*. San Diego: Academic Press.

Frijda N. H., (2009). Emotions, individual differences and time course: Reflections. *Cognition and Emotion*, 23, 1444 – 1461

Frijda, N. H., (1986). *The Emotions*. New York: Cambridge University Press.

Goeleven, E., De Raedt, R., Leyman, L., & Verschuere, B., (2008). The Karolinska Directed Emotional Faces: A validation study. *Cognition and Emotion*, 22, 1094 -1118.

Grice, H. P., (1989). *Studies in the Way of Words*. Cambridge, Mass: Harvard University Press.

Grice, H.P., (1975). Logic and conversation. In Cole, P. & Morgan, J.L. (eds.) *Syntax and Semantics, Vol. 3: Speech Acts* (pp. 41-58). New York: Academic Press.

Grosz, B. & Sidner, C., (1986). Attention, intention, and the structure of discourse. *Computational Linguistics*, 12, 175-206.

Guerini, M., Stock, O., & Zancanaro, M., (2007). A Taxonomy of Strategies for Multimodal Persuasive Messagge Generation. *Applied Artificial Intelligence*, 21, pp. 99-136.

Goodwin, C., (1981). *Conversational Organization: Interaction between Speakers and Hearers*. New York: NY Academic Press.

Gut, U., Looks, K., Thies, A., & Gibbon, D., (2003). CoGesT–Conversational Gesture Transcription System. Version 1.0. *Technical report*. Bielefeld University.

Hall, J. A., Levin, S. (1980). Affect and verbal-nonverbal discrepancy in schizophrenic and non-schizophrenic family communication. *British Journal of Psychiatry*, 137, 78-92.

Hancock, P. J., Bruce, V., & Burton, A. M., (1998). A comparison of two computer-based face identification systems with human perception of faces. *Vision Research*, 38, 2277–2288.

Hay, D. C., Young, A. W., & Ellis, A. W., (1991). Routes through the face processing system. *Quarterly Journal of Experimental Psychology*, 45, 123-156.

Hietanen, J. K., Leppänen, J. M., & Lehtonen, U., (2004). Perception of Emotions in the Hand Movement Quality of Finnish Sign Language. *Journal of Nonverbal Behavior*, 28, 53-64.

Holmes, J., (1986). Functions of you know in women's and men's speech. *Language in Society*, 15, 1-22.

Holmes, J., (1984). Hedging your bets and sitting on the fence: Some evidence for hedges as support structures. *Te Reo, Journal of the Linguistic Society of New Zealand*, 27, 47-62.

Holmes, J., (1983). The function of tag questions. *English Language Research Journal*, 3, 40-65.

Houghton, G., & Isard, S.D., (1987). Why to speak, what to say and how to say it: Modelling language production in discourse. In Morris, P. (ed.). *Modelling Cognition* (pp. 249-267). Chichester: Wiley.

Ide, S., (1982). Japanese sociolinguistics: Politeness and women's language. *Lingua*, 57, 357-385.

Isard, A. & Carletta, J., (1995).Transaction and action coding in the Map Task corpus. *Tech. Rep. HCRC/RP-65*. Edinburgh, Scotland: Human Communication Research Centre, University of Edinburgh.

Izard, C., (1993). Four systems for emotion activation: Cognitive and non-cognitive processes. *Psychological Review*, 100, 60–69.

Izard, C., (1977). *Human Emotions*. New York: Plenum Press.

Kendon, A., (1967). Some Functions of Gaze Directions in Social Interaction. *Acta Psychologica*, 26, 1-47.

Kipp, M., (2001). ANVIL - A Generic Annotation Tool for Multimodal Dialogue. In Eurospeech 2001 Scandinavia *7th European Conference on Speech Communication and Technology*.

Kipp, M., Neff, M., & Albrecht, I., (2006). An Annotation Scheme for Conversational Gestures: How to economically capture timing and form. In Martin, J.-C., Kühnlein, P., Paggio, P., Stiefelhagen, R., & Pianesi, F. (eds.) In *Multimodal Corpora: From Multimodal Behavior Theories to Usable Models*. (pp. 24-28). Berlin: Springer Verlag.

Kowtko, J. C., Isard, S. D., & Doherty, G. M., (1992). Conversational games within dialogue*. Technical Report HCRC/RP-31*, Human Communication Research Centre, University of Edinburgh

Krauss, R.M., & Morsella, E., (2000). Conflict and communication. In, Deutsch, M., & Coleman, P. (eds.), *The handbook of constructive conflict resolution: Theory and practice* (pp. 131-143). San Francisco: Jossey-Bass.

Krauss, R. M., Fussell, S. R., & Chen, Y., (1995). Coordination of perspective in dialogue: Intrapersonal and interpersonal processes. In Markova, I., Graumann, C. G., & Foppa, K. (eds.), *Mutualities in dialogue* (pp. 124-145). Cambridge, Eng: Cambridge University Press.

Krippendorff, K., (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30, 411-433.

Krippendorff, K., (1980). *Content Analysis: an introduction to its Methodology*. Sage Publications.

Kuppens, P., Van Mechelen, I., Smits, D.J.M., De Boeck, P. & Ceulemans, E., (2007). Individual differences in patterns of appraisal and anger experience, *Cognition and Emotion*, 21, 689–713.

Kurucz, J., Feldmar, G., & Werner, W., (1979). Prosopo-affective agnosia associated with chronic organic brain syndrome. *Journal of the American Geriatrics Society*, 27, 91–95.

Labov, W., (1972). *Language in the Inner City*. Philadelphia: University of Pennsylvania Press.

Lee, Ch. M., & Narayanan, S., (2003). Emotion recognition using a data-driven fuzzy inference system. In EUROSPEECH-2003, 157-160.

Lee, D. D., & Seung, H. S, (1999). Learning the parts of objects with nonnegative matrix factorization, *Nature*, 401, 788-801.

Levinson, S. C., (2006). On the human "interaction engine". In Enfield, N. J. & Levinson, S. C. (eds.), *Roots of human sociality: Culture, cognition and interaction* (pp. 39-69). Oxford: Berg.

Levinson, S. C., (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge: MIT press.

Levinson, S. C., (1995). Interactional biases in human thinking. In Goody, E. N. (ed.), *Social intelligence and interaction* (pp. 221-260). Cambridge: Cambridge University Press.

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Litman, D., & Hirschberg, J., (1990). Disambiguating cue phrases in text and speech. In *Proceedings of the Thirteenth International Conference on Computational Linguistics*, 251-256.

Maat, M., & Heylen, D., (2009). Turn Management or Impression Management? Amsterdam *IVA 2009*, 467-473.

Magno Caldognetto E. & Poggi, I. (2001). Dall'analisi della multimodalità quotidiana alla costruzione di agenti animati con facce parlanti ed espressive. In Atti delle XI Giornate di Studio del Gruppo di Fonetica Sperimentale. *Multimodalità e Multimedialità nella Comunicazione*. Padova, 29 novembre-1 dicembre 2000. Padova, Unipress, 2001, pp.47-55.

Magno Caldognetto E. & Poggi, I. (2002). Una proposta per la segmentazione e l'etichettatura di segnali multimodali. *Atti dell'Ottavo Convegno dell'Associazione Italiana per l'Intelligenza Artificiale*. Siena, 10 – 13 settembre 2002, pp. 195-203.

Magno Caldognetto, E., Poggi, I., Cosi, P., Cavicchio, F., & Merola, G., (2004). Multimodal Score: an Anvil Based Annotation Scheme for Multimodal Audio-Video Analysis. In Martin, J.-C., Os, E.D., Kühnlein, P., Boves, L., Paggio, P., & Catizone, R. (eds.) *Proceedings of Workshop Multimodal Corpora: Models of Human Behavior for the Specification and Evaluation of Multimodal Input and Output Interfaces*, (pp. 29-33).

Markova, I., & Foppa, K., (eds.) (1991). *Asymmetries in dialogue*. Prentice-Hall.

Martell, C., & Kroll, J., (2006). Using FORM Gesture Data to Predict Phase Labels. In Martin, J.-C., Kühnlein, P., Paggio, P., Stiefelhagen, R., & Pianesi, F. (eds.) *Multimodal Corpora: From Multimodal Behavior Theories to Usable Models,* (pp. 29-32).

Martel, C., Osborn, C., Friedman, J., & Howard, P., (2002). The FORM Gesture Annotation System. In Maybury, M., & Martin, J.-C. (eds.) *Proceedings of Multimodal Resources and Multimodal Systems Evaluation Workshop*, (pp. 10-15).

Martin, J.-C., Caridakis, G., Devillers, L., Karpouzis, K. & Abrilian, S., (2006). Manual Annotation and Automatic Image Processing of Multimodal Emotional Behaviors: Validating the Annotation of TV Interviews. In *Fifth international conference on Language Resources and Evaluation* (LREC 2006), Genoa, Italy.

Maynard, S. K., (1986). On back-channel behavior in Japanese and English casual conversation. *Linguistics*, 24, 1079-1108.

Mehrabian, A., (1971). *Silent messages*. Belmont, California: Wadsworth.

Milroy, L., (1987). *Language and Social Networks*. Oxford and New York: Basil Blackwell.

Nordenstam, K., (1992). Tag questions and gender. In *Swedish conversations. Working Papers on Language, Gender and Sexism*, 2, 75-86.

O'Barr, W., & Bowman, A., (1980). Women's language' or `powerless language. In McConnell-Ginet, S., Borker, R., & Furman, N. (eds.) *Women and Language in Literature and Society* (pp. 111-136). New York: Praeger.

Ortony, A., & Turner, T. J., (1990). What's basic about basic emotions? *Psychological Review*, 97, 315-331.

Padgett, C., & Cottrell, G., (1995). Identifying emotion in static face images. In *Proceedings of the 2nd Joint Symposium on Neural Computation*, San Diego, CA: University of California

Passonneau, R. J., & Litman, D., (1993). Intention-based segmentation: human reliability and correlation with linguistic cues. In *Proceedings of the 31st Annual Meeting of the ACL*, (pp. 148-155).

Pianesi, F., Leonardi, C., & Zancanaro, M., (2006). Multimodal Annotated Corpora of Consensus Decision Making Meetings. In Martin, J.-C., Kühnlein, P., Paggio, P., Stiefelhagen, R., & Pianesi, F. (eds.) *Multimodal Corpora: From Multimodal Behavior Theories to Usable Models*, (pp. 6-19).

Pillutla, M.M., & Murnighan, J.K, (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes,* 68, 208–224.

Poggi, I., (2007). *Mind, hands, face and body. A goal and belief view of multimodal communication*. Berlin: Weidler Buchverlag.

Poggi, I., & Vincze, L., (2008). The Persuasive Impact of Gesture and Gaze. In Martin, J.-C., Patrizia, P., Kipp, M., & Heylen, D., (eds*.) Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*, (pp. 46-51). Berlin: Springer Verlag

Poggi, I., Cavicchio, F., & Magno Caldognetto, E., (2007). Irony in a judicial debate: analyzing the subtleties of irony while testing the subtleties of an annotation scheme. *Language Resources and Evaluation*, 41, 215-232.

Reidsma, D., & Carletta, J., (2008). Reliability Measurement without Limits. *Computational Linguistics*, 34, 319-326.

.Reidsma, D. Heylen, D., & Op den Akker, R., (2008). On the Contextual Analysis of Agreement Scores. In Martin, J.-C., Patrizia, P., Kipp, M., & Heylen, D., (eds.) *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*, (pp. 52-55). Berlin: Springer Verlag.

Rodríguez, K., Stefan, K. J., Dipper, S., Götze, M., Poesio, M., Riccardi, G., Raymond, C., & Wisniewska, J., (2007). Standoff Coordination for Multi-Tool Annotation in a Dialogue Corpus. In *Proceedings of the Linguistic Annotation Workshop at the ACL'07* (LAW-07), Prague, Czech Republic.

Russell, J.A., & Barrett, L.F., (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76, 805-819.

Sander, D., Grandjean, D., Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks*, 18, 317–352.

Sanfey, AG, Rilling, JK, Aronson, JA, Nystrom, LE, & Cohen, JD., (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science,* 300, 1755-1758.

Sacks, H., Schegloff, E., & Jefferson, G., (1974). A simple systematics for the organization of turn-taking for conversation. *Language,* 50, 696-735.

Scherer, K. R., (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, 23, 307-351.

Scherer, K. R., (2001). *The nature and study of appraisal: A review of the issues*. New York and Oxford: Oxford University Press

Scherer, K. R., (2000). *Psychological models of emotion*. Oxford/New York: Oxford University Press.

Scherer, K. R., (1992). *What does facial expression express?* Chichester: Wiley.

Scherer, K. R., (1993). Studying the Emotion-Antecedent Appraisal Process: An Expert System Approach*. Cognition and Emotion*, 7, 325-355.

Scherer, K. R., (1987). Toward a dynamic theory of emotion: The component process model of affective states. *Geneva Studies in Emotion and Communication*, 1, 1–98.

Scherer, K. R., (1984). *Emotion as a multicomponent process: A model and some cross-cultural data*. Beverly Hills: CA: Sage.

Scherer, K. R., & Ellgring, H., (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*, 7, 113–130.

Scherer, K. R., & Heiner E., (2007). Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion*, 7,158-171.

Scherer, K. R., & Ceschi, G., (2000). Criteria for emotion recognition from verbal and nonverbal expression: Studying baggage loss in the airport. *Personality and Social Psychology Bulletin*, 26, 327–339.

Schober, M.F. & Clark, H.H., (1989). Understanding by addressees and over hearers. *Cognitive Psychology*, 21, 211-232.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19, 321–325.

Shadbolt, R.N., (1984). *Constituting Reference in Natural Language Dialogue: The problem of referential opacity*. Unpublished PhD thesis, University of Edinburgh.

Searle, J.R., (1979). A Taxonomy of Illocutionary Acts. In P. Cole, & J. Morgan (eds.). *Expression and Meaning: Studies in the Theory of Speech Acts*, (pp. 1–19). Cambridge: Cambridge University Press.

Siegel, S. & Castellan, N.J., (1988). *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill.

Sinclair, J., & Coulthard, M., (1975). *Towards an Analysis of Discourse: The English used by teachers and pupils*. London: Oxford University Press.

Smith, J., (1992). Women in charge: Politeness and directiveness in the speech of Japanese women. *Language in Society*, 21, 59-82.

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G., (2005). Transmitting and Decoding Facial Expressions. *Psychological Science*, 16, 184-189.

Sosnovsky, S., Brusilovsky, P., Lee, D.H., Zadorozhny, V., & Zhou, X., (2008). Re-assessing the Value of Adaptive Navigation Support. In *E-Learning Context in Adaptive Hypermedia and Adaptive Web-Based Systems 5th International Conference* (pp. 193-203). Heidelberg: Springer Berlin.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, 106 , 10587-10592.

Tassinary, L. G., & Cacioppo, J. T., (2000). The skeletomotor system: Surface electromyography. In Tassinary, L.G., Berntson, G.G., Cacioppo, J.T. (eds) *Handbook of psychophysiology* (pp. 263-299). New York: Cambridge University Press.

Taylor, T.J., & Cameron D., (1987). *Analysing Conversation: Rules and Units in the Structure.* Oxford: Pergamon.

Tickle-Degnen, L., & Rosenthal, R., (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, 1, 285–293.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H., (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behavior Brain Science*, 28**,** 675–735.

Traum, D.R., (1994). *A Computational Theory of Grounding in Natural Language Conversation*. Unpublished PhD thesis, University of Rochester.

Traum, David R., & Allen, J. F., (1994). Discourse obligations in dialogue processing. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics,* (p.1-8), June 27-30, Las Cruces, New Mexico [doi>10.3115/981732.981733]

Truong, K., (2009). *How Does Real Affect Affect Affect Recognition in Speech?* PhD Thesis Dissertation, Alpeloong, The Netherlands.

Valentin, D., Abdi, H., & O'Toole, A.J., (1994). Categorisation and identification of human face images by neural networks: A review of the linear autoassociative and principal component approaches. *Journal of Biological Systems*, 2, 412–429.

Van Mechelen, I., & Hennes, K., (2009). The Appraisal Basis of Anger Occurrence and Intensity Revisited*. Cognition and Emotion***,** 23,1451-1487.

Xiao, E., & Houser, D., (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, 102, 7398-7401.

Xue, Y., Tong, C. S., & Zhang, W., (2007). Survey of Distance Measures for NMF-based Face Recognition. In *Lecture Notes in Artificial Intelligence* (LNAI) 4456, (pp. 1039-1049). Berlin: Springer-Verlag.

Yngve, V. H., (1970). On getting a word in edgewise. Sixth Regional Meeting of the Chicago Linguistics Society, 567-577.

Yuan, J., Liberman, M., & Cieri, C., (2007). Towards an Integrated Understanding of Speech Overlaps in Conversation. In *International Conference of Phonetics,* Saarbrucken August 2007. http://www.icphs2007.de/conference/Papers/1617/1617.pdf

Wagner, H. L., (1993). On measuring performance in category judgment studies on nonverbal behavior. *Journal of Non-verbal Behavior*, 17, 3-28.

Wagner, H. L., & Smith, J. (1991). Facial Expressions in the Presence of Friends or Strangers. *Journal of Nonverbal Behavior*, 15, 201-214.

West, C. & Zimmerman, D., (1983). Small insults: A study of interruptions in cross-sex conversations between unacquainted persons. In Thorne, B., Kramarae, C., & Henley, N. (eds.) *Language, Gender and Society*. (pp. 102-117), Rowley, Massachusetts: Newbury House.

Woodworth, R. S., *Experimental Psychology*. New York: Henry Holt (First Edition 1938).

Zimmerman, D., & West, C., (1975). Sex roles, interruptions and silences in conversation. In Thorne, B., & Henley, N. (eds.) *Language and Sex: Difference and Dominance* (pp. 105-129). Rowley, Massachusetts: Newbury House.

APPENDIX

Coding Guidelines for the **Annotation of the RECC Corpus**

December 3rd 2008

Version 1.0

Rovereto Emotive Copperation Corpus (RECC) is built up of 240 minutes of Map Task interactions.

In this task, two persons, the Giver and the Follower, are facing each other having slightly different maps. The Giver has the aim of driving the Follower from a starting point to an end point.

You will see and hear some videos regarding the dialogues between the two. Your aim is to annotate their interaction from a multimodal point of view. Thus you will not only rely on words transcription but also to facial display and trunks movements, if any.

You will find on the screen display the chunks to label in a box like this

Right-clicking on it you will have a sliding menu opening.

If the chunks are too long or too shorts with respect to the event to code (> 200 ms), you can modify the event. Click on start and move the event to the new start line and do the same with the end, if applicable.

If there is a tag error, you can modify it as well.

If you have too many events, you can delete the unnecessary ones. If there is a missing event, place yourself at the start of the event, then move forward the duration of the offset and introduce the new event.

To code the events, choose edit from the menu. A new window will open on the upper left corner of the screen with a sliding menu. Read carefully the given labels and attribute it to the chunk just clicking on the label you choose. If you want to see again the chunk in motion click on the play button, which is located in the bottom of the new window you have just opened

In the following, we list and describe the modalities and the annotation features of our multimodal annotation scheme:

Modality	**Expression type**
Facial displays	Eyebrows
	Head
	Gaze
	Mouth
Speech	Speech
	Collaboration
	Turn management

1. Cooperation Annotation:

Instructions (Cooperation Typology)

No answer to question: no answer given when required

Inappropriate reply : failure to introduce useful information when required

No Spontaneous Add/Repetition of Instruction: information is not spontaneously added or repeated after a check

Giving Instructions: task baseline, when the Giver introduces a new feature **IT DOES NOT APPLY TO FOLLOWER ANNOTATION**

Acknowledgment: a verbal response which minimally shows that the speaker has heard the move to which it responds

Question answering (Y/N): Yes-No reply to a check

Check: questions (function or form) which solicit other understanding of information already offered

Repeating Instructions: repetition of an instruction following a check

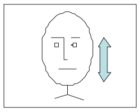
Question answering + adding information: Yes-No reply + new information introduction

Spontaneous info/description adding: introduces new information relevant to the task

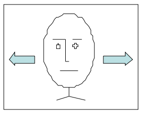
2. Turn Management:

The turn management system regulates the interaction flow and minimizes overlapping speech and pauses. It is coded by the three general features *Turn gain*, *Turn end* and *Turn hold*. In addition, a turn gain is either a *Turn take* if the speaker takes a turn that wasn't offered, possibly by interrupting, or a *Turn accept* if the speaker accepts a turn that is being offered. Similarly, turn end can be achieved in different ways: the speaker can release the turn under pressure (*Turn yield*), offer the turn to the interlocutor (*Turn offer*).

- **Turn giving/offer**: the speaker give or offer the conversational turn to the other
- **Turn accept:** the speaker accepts a turn that is being offered/given
- **Turn yielding**: the speaker can release the turn under pressure of the other speaker
- **Turn holding**: hold conversation turn even with *ehhhm, uhmmm* speech sounds or word repetition
- **Turn taking**: the speaker take a turn that wasn't offered, possibly by interrupting the other speaker
- **Backchannel:** one of the speaker answers with *yes/no, ok* or head movements (*shake* or *nod*) alone



head nod



head shake

3. FACIAL CUES ANNOTATION

In the following we describe the conformation analysis of emotive labial movements as implemented in our annotation system. It is based on a little amount of signs similar to emoticons. We sign two levels of activation using the plus and minus signs.

3.1 Mouth:

- **Closed lips/lips in relaxed position:** when the mouth is closed**, -**
- **Corners up:** e.g. when smiling**,); +)** wide smile/ laugh
- **1 corner up**: asymmetric smile
- **Corners down:** e.g. in a sad expression**, (**
- **Lower lip biting:** the subject is biting her/his lower lip
- **Protruded:** when the lips are rounded, O.

3.2 Eyebrows:

- **frown:** when the eyebrow are frowning
- **+frown:** when eyebrow are very frowned
- **Up**: both eyebrows up
- **+Up:** eyebrow very up
- **--:** eyebrow in normal position

3.3 Gaze:

- **Up**: when the person looks up
- **Down**: when the person looks down
- **Sideways**: when the person looks on the side
- **Unfocused/no gaze**: when the speaker/listener is looking at the space, without focusing on anything or anybody in particular, this is not the same as "neutral" since it shows the interlocutor is "lost in his/her thoughts"
- **To the interlocutor**: refers to a situation in which the two interlocutors are looking at each other, usually in the region of the face, this can include eye contact.

RECC CORPUS SPECIFICATION FILE

```
  <?xml version="1.0" encoding="ISO-8859-1" ?>
- <annotation-spec>
  <doc>WORKINPROGRESS for anvil 4.0-FEDERICA</doc>
- <head>
- <valuetype-def>
- <valueset name="cooperationtipo">
  <value-el>repeating given instruction</value-el>
  <value-el>checking instruction</value-el>
  <value-el>spontaneous info/description adding</value-el>
  <value-el>yes-answer</value-el>
  <value-el>no-answer</value-el>
  <value-el>yes-answer + adding</value-el>
  <value-el>no-answer + adding</value-el>
  <value-el>no answer to question</value-el>
  <value-el>no information add when required</value-el>
  <value-el>inappropriate reply (no giving info)</value-el>
 <value-el>backchannel</value-el>
<value-el>acknowledgment</value-el>
    </valueset>
- <valueset name="turnmanagementipo">
  <value-el>turn holding</value-el>
  <value-el>turn giving/ offer</value-el>
  <value-el>turn taking</value-el>
  <value-el>turn accept</value-el>
  <value-el>backchannel</value-el>
  <value-el>backchannel head nod</value-el>
<value-el>backchannel head shake</value-el>
  <value-el>turn yielding</value-el>
  <value-el>turn concluding</value-el>
    </valueset>
- <valueset name="eyebrowmovimentotipo">
  <value-el>--</value-el>
  <value-el>Up</value-el>
  <value-el>+Up</value-el>
  <value-el>frown</value-el>
  <value-el>+frown</value-el>
    </valueset>
- <valueset name="gazemovimentotipo">
  <value-el>to the interlocutor</value-el>
  <value-el>down/to the map</value-el>
  <value-el>up</value-el>
  <value-el>Waggle</value-el>
  <value-el>side-turn</value-el>
  <value-el>unfocused</value-el>
    </valueset>
- <valueset name="mouthopeningtipo">
  <value-el>O</value-el>
  <value-el>closed</value-el>
  <value-el>)</value-el>
  <value-el>+ )</value-el>
  <value-el>(</value-el>
  <value-el>1 corner up</value-el>
  <value-el>lip biting</value-el>
    </valueset>
    </valuetype-def>
    </head>
- <body>
- <group name="Speech Analysis">
  <track-spec name="Speech" type="waveform" height="2" />
- <track-spec name="Giver Transcription" type="primary" height="0.5">
  <attribute name="parole" valuetype="String" />
    </track-spec>
- <track-spec name="Follower Transcription" type="primary" height="0.25">
  <attribute name="valore" valuetype="String" />
    </track-spec>
- <track-spec name="Cooperation Type" type="primary" height="0.25">
  <attribute name="tipo" valuetype="cooperationtipo" />
    </track-spec>
- <track-spec name="Turn Management" type="primary" height="0.25">
  <attribute name="tipo" valuetype="turnmanagementipo" />
    </track-spec>
    </group>
- <group name="Mouth Movements Analysis">
- <track-spec name="Mouth movements analysis" type="primary"
    height="0.25">
  <attribute name="tipo" valuetype="mouthopeningtipo" />
    </track-spec>
    </group>
- <group name="Eyebrows Movements Analysis">
- <track-spec name="eyebrows movements type and intensity"
    type="primary" height="0.25">
  <attribute name="tipo" valuetype="eyebrowmovimentotipo" />
    </track-spec>
    </group>
- <group name="Eyes and Gaze Analysis">
- <track-spec name="Gaze Movements" type="primary" height="0.25">
  <attribute name="tipo" valuetype="gazemovimentotipo" />
    </track-spec>
    </group>
    </body>
    </annotation-spec>
```