Eyebrow raising in dialogue: discourse structure, utterance function, and pitch accents

María L. Flecha-García, M.Sc.

A thesis submitted in fulfilment of requirements for the degree of
Doctor of Philosophy

to

Theoretical and Applied Linguistics
School of Philosophy, Psychology and Language Sciences
University of Edinburgh

July 2004

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

María L. Flecha-García

Abstract

Some studies have suggested a relationship between eyebrow raising and different aspects of the verbal message, but our knowledge about this link is still very limited. If we could establish and characterise a relation between eyebrow raises and the linguistic signal we could better understand human multimodal communication behaviour. We could also improve the credibility and efficiency of computer animated conversational agents in multimodal communication systems.

This thesis investigated **eyebrow raising** in a corpus of task-oriented English dialogues. Applying a standard dialogue coding scheme (Conversational Game Analysis, Carletta et al., 1997), eyebrow raises were studied in connection with **discourse structure** and **utterance function**. Supporting the prediction, more frequent and longer eyebrow raising occurred in the **initial utterance of high-level discourse segments** than anywhere else in the dialogue (where *'high-level discourse segment'* = *transaction*, and *'utterance'* = *move*, following Carletta et al.). Additionally, eyebrow raises were more frequent in **instructions** than in requests for or acknowledgements of information. Instructions also had longer eyebrow raising than any other type of utterance. Contrary to the prediction, the start of a lower-level discourse segment (*conversational game*) did not have more eyebrow raising than any other position in the dialogue, and queries did not have more eyebrow raising than any other type of utterance.

Eyebrow raises were also studied in relation to intonational events, namely **pitch accents**. Results showed evidence of **alignment** between the brow raise start and the start of a pitch accent. Most pitch accents were not associated with brow raising, but when brow raises occurred they tended to immediately precede a pitch accent on the speech signal. To investigate what could explain the alignment

between the two events, pitch accents aligned with eyebrow raises were compared to all other pitch accents in terms of: phonological characteristics (*primary* vs. *secondary* pitch accents, and *downstep-initial* vs. *non-initial* pitch accents), information structure (*given* vs. *new information* in referring expressions, and the *last quarter* vs. *earlier parts* of the utterance length) and type of utterance in which they occurred (*instruction* vs. *non-instruction*). Those comparisons suggested that brow raises may be aligned more frequently with **pitch accents in *downstep-initial* position and in *instructions***. No differences were found in terms of information structure or between *primary/secondary* accents.

The results provide evidence of a link between eyebrow raising and spoken language. Eyebrow raises may signal the start of linguistic units such as discourse segments and some prosodic phenomena, they may be related to utterance function, and they are aligned with pitch accents. Possible linguistic functions are proposed, such as structuring and emphasising information in the verbal message.

Acknowledgements

I would like to thank, first of all, my supervisors Dr. Ellen G. Bard and Prof. D. Robert Ladd, for their valuable advice and support in this long project. Thanks also to the computing support team at Linguistics (specially Cedric and Eddie). I am also grateful to The Engineering and Physical Sciences Research Council (EPSRC) for providing funding for this research.

There is a long list of friends I would like to thank as well. Here I will only mention (in alphabetical order) some of those who helped me in some way in the final year: Susana Cortes, Francesca Filiaci, Markus Guhe, Ruth Hanson, Christine Haunz, Cassie Mayo, Scott McDonald, Terri McKeigan, Keith and Laure Mitchell, Hannele Nicholson, and Ivan Yuen.

Finally, my deepest thanks go to my husband (Jaime), and to my parents (Honorino and Socorro) and sister (Susana). In spite of the long distance, they were always with me. Gracias por vuestro enorme cariño y apoyo. A pesar de la larga distancia os tuve presentes en todo momento. Siempre os llevo conmigo. Jaime, esta tesis va dedicada a ti.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

1.1 What this thesis is about

Like many of the subtleties of human communication, the use of the face is something we believe we understand but cannot yet describe. Human facial movements have attracted a great deal of research, but the information we have about the use of the face in multimodal communication is still very limited. Research on facial movements has been largely dominated by the study of the expression of emotion (see 2.1). By contrast, and leaving aside movements that are necessary for the articulation of speech, studies on facial movements in connection with spoken language have been scarce. The difficulty in carrying out this type of research may explain to some extent the apparent neglect in the literature. Observations have, however, been made which suggest possible conversational functions of eyebrow raises in particular, and in recent years a few studies have taken an empirical approach to this issue (see 2.5.2 and 2.5.3). Many results do not appear robust and more research is needed, especially in the temporal coordination of facial movements with the linguistic signal.

The sketchy nature of the evidence for coordination is particularly surprising because other body movements have been shown to have a non-random relation with the speech they accompany. For instance, hand gestures, head movements, body shifts, and gaze seem to integrate with language to deliver a message (see 2.4). If this is the case, then the intuition that eyebrow movements are also related to the speaker's message is probably well founded.

In this thesis, I investigate the relationship between the linguistic signal and eyebrow raises in dialogue. In particular, I am interested in *when* these movements

appear in relation to speech, and ultimately in *why* they happen at all. If we could describe the use of facial movements in dialogue, embodied conversational agents in multimodal communication systems would not be hampered by strikingly poor coordination between the verbal and the visual channel. One of the aims of this research is to provide some information that may improve this coordination in the design of such systems. The point is not to make these computer-animated agents look simply more 'human' in appearance, but to make them more efficient at communicating with us by showing the right movements at the right time. A cartoon face could in fact convey a message more efficiently than a more human-like animated face if the former raised its eyebrows at the same points as a real speaker would. Some studies have already shown how the coordination of synthetic speech and eyebrow raises in computer-animated talking heads can affect perception (e.g. House et al., 2001; Krahmer et al., 2002a,b, see 2.5.3). However, a lot more research on the production of natural conversation is necessary for an accurate reproduction or generation of eyebrow raising during speech.

Using the Map Task (Anderson et al., 1991) and a dialogue structure coding scheme (Carletta et al., 1997), I collected and annotated a set of task-oriented English dialogues to observe eyebrow raising in natural interaction. The Map Task allowed the study of spontaneous, yet controlled, behaviour of the speakers in the conversation. The aim was to investigate if brow raises could signal the kind of linguistic phenomena that some studies have associated with body movement in general (see 2.4) and eyebrow raises in particular (see 2.5). As will be explained in Chapter 2, many of these studies have based their conclusions on non-empirical evidence, or have followed an inductive approach (Chovil, 1989). The few exceptions that have used a hypothetico-deductive approach have studied languages other than English and have been inconclusive (Cavé et al., 1996, 2002) or have used synthethic data in perception experiments (Krahmer et al., 2002a; House et al., 2001). The main linguistic phenomena previously investigated in relation to body movement were related to: discourse structure and utterance function, information structure, prosodic phenomena, and the alignment between events in the linguistic signal and events in body movement. The current thesis addresses all of these issues in a single set of collected data.

Discourse structure will be described briefly in section 2.3 and explained in more detail within the framework of Conversational Game Analysis in 3.3.1. Although the term discourse is not synonymous with dialogue, in this thesis discourse

structure generally refers to the structure of dialogue. Thus it refers to how a conversation can be segmented into sections, each with a coherent communicative purpose. A short example would be the following extract, slightly modified for the purpose of exposition, from one of the dialogues investigated. In this sample Rita (R) is giving navigation instructions from her map to Claire (C) so that Claire can draw a route on a slightly different copy of that map. Utterances are numbered u1 to u11:

u1 R: *Now, have you got an almond tree?*
u2 C: *No*
u3 R: *Ok*
u4 R: *Have you got an anemone?*
u5 C: *Yeah*
u6 R: *Well, about an inch below the anemone, in a sort of straight line down, I've got an almond tree*
u7 C: *Yeah*
u8 R: *And you want to go sort of above the anemone but below the almond tree*
u9 C: *Ok*
u10 R: *Mmm ... and then ... have you got a vineyard?*
u11 C: *Yes, I have*
... (conversation continues)

Utterances u1 to u3 are clearly connected, and so are u4 and u5. In each case there is a request for information, which is then provided by the interlocutor and so there is a sense of completion of the initial goal. If we had to divide the conversation into segments like that, utterances u6 and u7, and u8 and u9 would also form segments: in the first one, an explanation is provided and then acknowledged, and in the second one, an instruction is given and then also acknowledged. At the same time, these four segments seem to be connected. In u1, Rita has referred to a landmark, an almond tree, that Claire does not have in her map. The task they are doing requires Claire to draw a route line around her landmarks, while avoiding also Rita's landmarks. In order to tell Claire how to avoid the almond tree, Rita asks about a different landmark, the anemone, and then after an explanation about the distance relation between the two landmarks, she is able to instruct Claire how to go below where the almond tree is. This higher goal links the four segments which then form another segment at a higher level. Then, at this level, the next utterance (u10) starts another segment that will deal with a different portion of the map route, around the vineyard. And in this way the structure of the dialogue is developed. Some body movements, such as head movements (McClave, 2000) and body shifts (Cassell et al., 2001), seem to occur more frequently at the start of discourse segments than somewhere else in the

discourse structure (see 2.4.3). The same has been reported for eyebrow movements (Chovil, 1989; Cavé et al., 2002; see 2.5.2). Eyebrow raises, then, could signal a shift into a new discourse segment, for instance in u10 in the example above. This will be investigated in Chapter 4.

But there is more to discourse structure than just segments: **utterance function** is strongly linked to the structure of the dialogue. As we have seen in the example above, the utterances within these segments are not all of the same type. They can be distinguished by their purpose or function. For instance, u1 requests information, u2 provides the elicited information, u3 acknowledges that the information has been received and understood. Some studies have claimed that brow raises can have a questioning function (e.g. Ekman, 1979; Bavelas and Chovil, 1997), but there is no strong empirical evidence to support this claim. This issue is addressed in Chapter 4 in the current thesis. Additionally, due to the nature of the dialogues under investigation, the possible marking of another kind of utterance function is also investigated, namely instructing. Instructions are the most important type of utterance in the task performed by the participants in this study. For this reason, instructions may need some kind of marking or reinforcing device to make them distinct from other utterances, and this device could be in the visual channel in the form of eyebrow movements.

Another linguistic area that has been related to some body movements is **information structure** (McNeill 1992, for hand gestures; Cassell et al. 1999, for gaze; Krahmer et al. 2002b, for synthetic eyebrow raises). Section 2.3.3 presents a brief description of what information structure refers to, and Chapter 5 investigates brow raises in relation to the contrast between *given/new* information. In the example above the references to the 'almond tree' in utterances u1 and u6 are different in terms of information structure. In u1 this landmark is mentioned for the first time in the conversation and thus the 'almond tree' is considered *new* information there, as opposed to u6, where it is mentioned for the second time and is *given* information. Chapter 5 investigates whether eyebrow raises could signal this kind of contrast.

Another important issue addressed in this thesis is the alignment of eyebrow raises and some prosodic events, namely pitch accents. Body movement has been strongly associated in the literature to **prosodic structure** (see 2.4.2). Short

brow raises, in particular, have been linked to pitch accents in natural conversations in French (Cavé et al., 1996, 2002) (see 2.5.2) and in the perception of synthetic stimuli in Dutch (Krahmer et al., 2002a,b) and Swedish (House et al., 2001) (see 2.5.3). Prosodic phenomena are more difficult to illustrate in the example above, because they are not related to what speakers say, but to how they say it (I will return to this in 2.3.2). For instance, it is likely that Rita emphasised 'almond tree' and 'anemone' when she first mentioned them in the conversation. These were most likely phonologically marked with a pitch accent. And for extra reinforcement, perhaps they were accompanied with eyebrow raising as well. Chapter 5 addresses this kind of question in English.

The aim of this thesis in addressing the issues presented above is to determine when eyebrow raises occur in relation to the verbal message, and ultimately why. This is not only interesting from a psycholinguistic point of view, but also, as we will see in section 2.7, from an engineering perspective: if we could predict when real speakers will raise their eyebrows as they talk, we could presumably generate this behaviour in animated agents in order to make more efficient multimodal dialogue systems. Thus, one of the goals of the thesis is to provide some useful information for the design of embodied conversational agents. Another goal is to present and evaluate a methodology that could be used in the study of facial movements accompanying speech.

1.2 Clarifications on terminology

After describing the main topics of this thesis, two things need to be explained about the terminology used, especially in Chapter 2. First, when referring to muscle facial activity in general I will mainly use the phrase 'facial movement', except when discussing some fields of research in which other terms have been preferred. Psychologists investigating the link between the face and the expression of emotion, have almost unanimously used the term 'facial expression'. But there is some controversy about whether facial movements that are commonly called expressions, are indeed the 'expression' of inner states. Ekman (1997), one of the representatives of this line of research, admitted that he was not comfortable with this term in his early research "because it admits that some inner state is being manifested or shown externally" (Ekman et al., 1972, p. 3). However, he

then adopted the term with the belief that the facial movements it refers to are "outward manifestations of changes that have occurred and are occurring internally in the brain. (...) Expression is part of those changes and a sign that those changes are happening" (Ekman, 1997, p. 322). Izard (1997), another representative researcher of the face-emotion link, prefers the term "facial pattern". Finally, Fridlund (1997) and Chovil (1991a), who do not see emotion as the main cause for facial activity, have used the term "facial display". I prefer the general term "facial movement" because it is not associated to any particular theory about what these movements mean.

Second, when referring to facial movements I will not include those necessary for the articulation of speech[1] or the ones used in sign language. The former, such as lip shapes, jaw and tongue movements, have been studied carefully by phoneticians and others. They play a central role in the articulation of speech and therefore in communication, but it is a different role to the one I am interested in. These movements are necessary for the articulation of the different sounds in spoken language, whereas the movements I refer to in this thesis are clearly not necessary, even though they may add useful information to a linguistic message. I should therefore qualify them as "non-articulatory". In the interest of simplicity, however, and once this has been clarified, I will not use that term. Similarly, I will not deal with the facial movements of signers. It has been shown that movements of the upper face are used in sign language to fulfil prosodic functions (e.g. Wilbur, 1994). Corina (1989, cited in Chovil (1997)) found that facial displays helped mark the introduction of topics, clauses, questions and other syntactic constructions. For instance, *yes/no* questions are conveyed by raised eyebrows and a forward head tilt. This is a very interesting use of the face which provides evidence of how the upper face can fulfil linguistic functions that could have been conveyed by some other behaviour. However, the use of the face in signed and spoken language cannot be directly compared. In the former, there is no auditory modality available and facial movements are used in a systematised way that is not necessary in spoken language.

[1]Some movements of the articulators will be briefly mentioned in section 2.5.3, when discussing ongoing research on their correlation with emphatic speech. Notice however that these represent a departure from neutral articulatory movements

1.3 Thesis chapter outline

In the following outline I use the abbreviations BR for eyebrow raise, and PA for pitch accent. These abreviations will also be used in the Method section of Chapters 4 and 5.

Chapter 1 As an introduction to the topic of this thesis, I have explained that we know very little about facial movements in the linguistic context, and that it is important for both research and industry to find out how movements such as BRs may be related to the verbal channel. The general questions of this thesis are briefly summarised as *when do BRs occur in relation to the linguistic signal?* and *why?* The specific issues addressed have been introduced here.

Chapter 2 In this chapter I first provide a background to research on facial movements by describing some of the literature on facial expressions of emotion (2.1). Then I present an alternative view that treats facial movements as social communicative signals (2.2). Moving on to linguistic communication, I provide some linguistic background (2.3) before presenting the kind of research that has related body movements to the verbal message (2.4). Special attention to eyebrow movements in a separate section (2.5) will show how the few studies available suggest that eyebrow raises may have an important linguistic role in communication but how we still have very limited knowledge about this behaviour in conversation. Next, I describe some of the methods that have been used in research to measure facial movements (2.6). Finally, I present some background on Embodied Conversational Agents (ECAs) as an area that can profit from research such as the one in this thesis (2.7).

Chapter 3 Here I describe the methodology used in the corpus collection and annotation for this thesis. First there is a description of the experimental setup used in the data collection (3.1). Then I present the method used to record the corpus (3.2). And finally, I describe the method employed to annotate dialogue structure, PAs, information structure, and BRs, in the dialogues selected from the corpus (3.3). This last section includes some images to provide examples of eyebrow raises from the participants in this study.

Chapter 4 In this first experimental chapter I present an investigation into possible linguistic functions of BRs associated to discourse structure and to utterance function. In particular, I report multiple regression analyses carried out to find out whether speakers raise their eyebrows more frequently at the start of a new segment in the dialogue structure and when giving instructions or asking a question, versus at other positions in the structure and when producing utterances with different purposes, respectively.

Chapter 5 The second experimental chapter investigates where within the utterance BRs occur and which linguistic roles might be inferred from this. Motivated by previous observations that BRs are aligned with certain prosodic phenomena, I look at the alignment between BRs and PAs in the dialogues under investigation.

Chapter 6 In answer to the question '*when* does brow raising occur in dialogue', this final chapter summarises the findings in the thesis (6.2.1) and their relation to previous studies (6.2.2). It then examines reasons why eyebrow raising occurs (6.3). Issues related to the methodology employed in the thesis are discussed next (6.4). Also, practical applications in the development of Embodied Conversational Agents are explained for this kind of research (6.5). Finally, future research directions are suggested before concluding the chapter.

CHAPTER 2

Literature review

In this chapter I will provide some background to the current study by looking at previous research on body movement in general and eyebrow movements in particular. Facial movements have been investigated from different points of view. I will first summarise some research on facial movements as expressions of emotion and then I will present an alternative view that treats these movements as communicative social signals. Body movement, in general, has been related to linguistic phenomena. I will provide some linguistic background before presenting previous studies that have associated body movements, such as hand gestures, with certain linguistic functions. Next, I will describe the research that has been done on possible linguistic functions of eyebrow raises in particular. Unfortunately, as we will see, this is an area where empirical research has been scarce. Contributing to this is the fact that measuring spontaneous facial movements is not an easy task. I will describe different methods that have been used for this. Finally, as one of the motivations for this thesis, I will briefly present an area of current technology where the need for more informative studies about facial behaviour is pressing: Embodied Conversational Agents within multimodal dialogue systems.

2.1 Facial movements and the expression of emotion

Psychological studies of facial movements have traditionally been directed towards the expression of emotion. A link between emotional states and facial

expressions was already discussed by Darwin. Previous to his publication *The Expression of the Emotions in Man and Animals* (Darwin, 1872), this link had already been suggested by Bell (1844) and Duchenne de Boulogne (1862). Though Darwin cited both of them in his book, he did not support their belief that certain muscles had been given by God so that man could express emotion. Darwin claimed that facial expressions of emotion were innate, universal, and a product of evolution, and were not exclusive to the human race. He described some expressions of "states of the mind", such as "surprise", and also tried to explain why we produce them (pages refer to Ekman's edition, 1998):

> Attention is shown by the eyebrows being slightly raised; and as this state increases into surprise, they are raised to a much greater extent, with the eyes and mouth widely open. The raising of the eyebrows is necessary in order that the eyes should be opened quickly and widely; and this movement produces transverse wrinkles across the forehead. (p. 278)

> As surprise is excited by something unexpected or unknown, we naturally desire, when startled, to perceive the cause as quickly as possible; and we consequently open our eyes fully, so that the field of vision may be increased, and the eyeballs moved easily in any direction. But this hardly accounts for the eyebrows being so greatly raised as is the case, and for the wild staring of the open eyes. The explanation lies, I believe, in the impossibility of opening the eyes with great rapidity by merely raising the upper lids. To effect this the eyebrows must be lifted energetically. (p. 280)

As Ekman explains in the introduction to the third edition, Darwin's book was initially a best-seller, but his ideas about expressions and emotion were soon criticised and then simply ignored for decades. One of the reasons, according to Ekman, was that Darwin's notion of expressions as innate, a product of evolution and therefore part of our biology, was contrary to the behaviourist views that dominated psychology around the first half of the 20th century. It was also incompatible with cultural relativism of anthropologists such as Bateson and Mead (1942), who claimed that facial expressions differ from culture to culture and that they are communicative signals tied to the flow of conversation rather than to internal states.

From the 1960s onwards the study of facial expression returned to the arena of psychological research and was legitimized by several events as explained by

Rosenberg (1997). In 1962 Tomkins published a theory of affect in which the face played a central role as a site of emotion. His ideas were consistent with Darwin's, and together with McCarter he reported a study (Tomkins and McCarter, 1964) in which observers consistently identified facial expressions as indicative of certain emotions. Ekman and Izard's crosscultural work on facial expressions of emotion in literate and preliterate cultures (e.g. Ekman and Friesen, 1971; Izard, 1971) presented evidence for the universality of the recognition of facial expressions of emotion. Subsequent work by them and their followers strengthened the face-emotion link, which has dominated the study of facial expression in psychology until thus far.

The bias towards emotion on the study of the face has been discussed by Russell and Fernández-Dols (1997) who argued for a broader approach. They explained how by the 1980s this "Facial Expression Program" dominated research on the face through a network of assumptions, theories and methods. This program is characterised by the establishment of a small number of basic emotions, universal and discrete, with characteristic facial expressions that result from evolutionary adaptation. Six basic emotions, are generally agreed on (Ekman and Friesen, 1971): happiness, surprise, fear, anger, disgust, and sadness, with contempt sometimes added to the list. Other emotions are subcategories or combinations of these. Facial expressions of emotion are considered involuntary (Ekman and Rosenberg, 1997) but can also be voluntarily managed by what has been called "display rules" (Ekman et al., 1972), that determine what type of emotion is appropriate to a particular situation in a particular culture. One of the main ideas defended by psychologists in this program, the universality of facial expressions of emotion (e.g. Ekman, 1980), was criticised by Russell and Fernández-Dols, who argued that there is insufficient evidence for identical expression-emotion pairs across cultures. They even questioned the link between emotion and the face, since the assumption that facial expressions are caused by emotions has not been tested.

In the analysis carried out in this thesis, expressions of emotions were not considered. I believe the relationship between some facial behaviour and internal emotional states accounts for only a small portion of our non-articulatory facial movements, and by focusing on it other important aspects of facial activity that may be crucial for communication have been neglected.

2.2 Facial movements as social communicative signals

Researchers have considered functions of facial movements beyond the expression of emotion. In this section I will introduce some research that has emphasised the social function of certain facial movements.

2.2.1 Ethological approach to facial movements

As we have seen in the previous section, the notion that facial expressions are involuntary indicators of underlying emotions has been challenged from different sides. One source of criticism is human ethology. For instance, after working within the tradition of the face-emotion link for many years, Fridlund later proposed an alternative explanation for facial expression: the behavioural ecology view (e.g. Fridlund, 1997). Within this view facial displays are not expressions of discrete emotions, involuntarily produced and sometimes modified by display rules. Instead they are social signals, "messages which influence others' behavior because vigilance for and comprehension of signals co-evolved with the signals themselves" (Fridlund, 1997, p. 104). These signals signify our intentions in a given social interaction and are only interpretable within that context. For example, as Fridlund explains, a display explained by the emotions view as an expression of "anger" is interpreted in the behavioural ecology view as showing "readiness to attack". Similarly, a "fear face" in the former, would be a sign of "readiness to submit or escape" in the latter, and so on. The behavioural ecology view therefore, emphasises the social function of facial displays. And this "sociality" applies even in situations when there is no "interactant" present, because "people are always implicitly social even when schematically alone" (p. 123). Some aspects of Fridlund's view may not seem very different from the emotions view. Indeed, Fridlund claims that the behavioural ecology view is not antagonistic to emotion, "it simply regards the term as unnecessary to understand how our facial expressions both evolved and operate in modern life" (p. 124).

Eibl-Eibesfeldt has also studied facial and other body movements as social signals and as part of human communicative behaviour not necessarily linked to emotion. He was interested in the social functions of human behaviour across cultures and pointed out that expressive behaviour has a communicative function, and can be interpreted by an onlooker even if a message is not intended.

When someone shivers, for instance, "he does not necessarily intend to communicate 'I am cold' or 'afraid', but the perceiver of the behavior may recognize the mood of the sender and either learn to attach significance to it or phylogenetically adapt to it" (Eibl-Eibesfeldt, 1979, p. 11).

Motivated by what he considered an inadequate documentation, particularly of social behaviour, in film libraries at the time, Eibl-Eibesfeldt started a programme with Hass on the crosscultural documentation of human expressive behaviour (Eibl-Eibesfeldt and Hass, 1967, cited in Eibl-Eibesfeldt, 1972). They used angle lenses to film people without their awareness by having the camera point in a different direction. Most of the events were filmed in slow motion at 48 frames per second, allowing them to observe movements that otherwise would not have been noticed. But sometimes they speeded up to 2–7 frames per second to record longer events as a whole, such as a ritual, a flirting couple, etc., and to observe sequences of patterns. Every shot was accompanied with a detailed commentary stating the context in which each pattern occurred and what the person did before and after that recording was made. In this way they could interpret specific movements by their recurrence in certain contexts, and they could compare data recorded in different cultures.

Eibl-Eibesfeldt paid particular attention to what he called the "eyebrow flash" (Eibl-Eibesfeldt, 1972; Grammer et al., 1988): a very quick raising of the eyebrows, which in his recordings were maximally raised for approximately $\frac{1}{6}$ of a second. He first observed this movement in different cultures in situations of greeting, especially over a distance, in which people would also smile and nod. Later he observed it in several other situations, such as flirting, approving, seeking confirmation, thanking, and emphasising a statement (calling for attention). He concluded that in these situations "the basic common denominator is a 'yes' to social contact and that the eyebrow flash is used either for requesting such a contact or for approving a request for contact" (1972, p. 300).

Looking at other contexts in which eyebrow raising occurred, Eibl-Eibesfeldt observed that people often raise and hold their eyebrows up for a while when they are surprised or, in conversation, when asking a question. He explained that in both cases people attend and open their eyes to perceive better and their eyebrows are raised in connection with the opening of the eyes. He hypothesised

that "the eyebrow lift of surprise - originally part of the opening of the eye - was the starting point for the ritualization of several 'attention' signals" (1972, p. 301). He grouped some of these into "friendly attention signals" represented by the eyebrow flash, which can be accompanied by nodding and smiling. As can be seen in Figure 2.1, emphasising a statement was included into the contact and approval seeking eyebrow flash, whereas eyebrow raising when asking questions was considered a different attention signal.

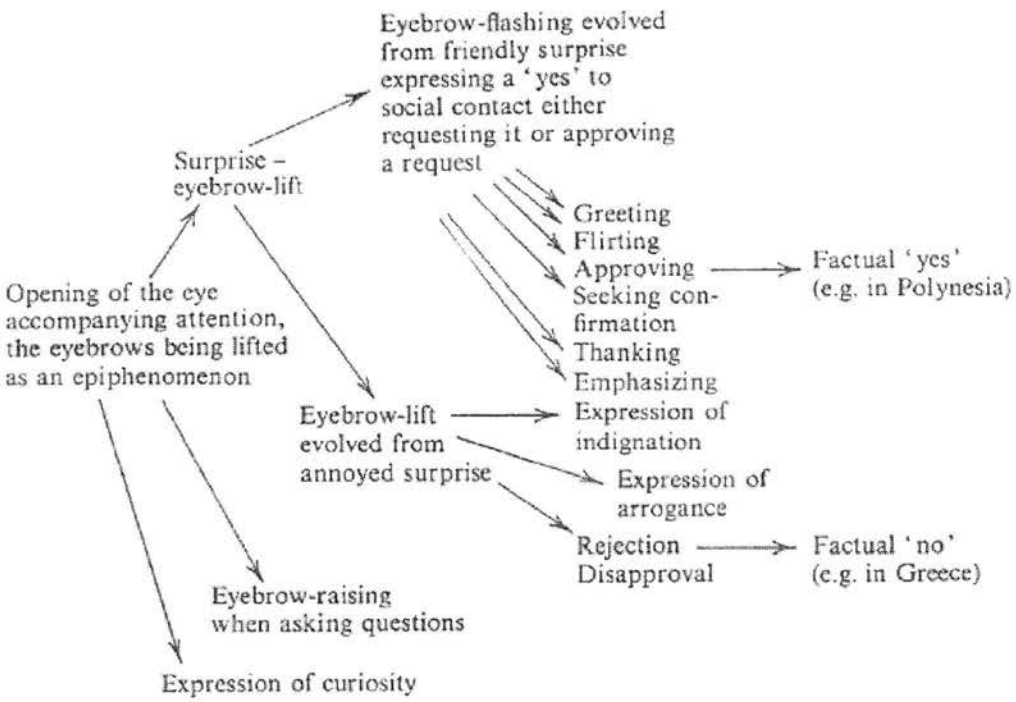

Figure 2.1: Hypothetical evolution of eyebrow movements into signals in man, after Eibl-Eibesfeldt (1972), p. 302.

Grammer et al. (1988) analysed 255 instances of eyebrow raising from recordings of three different cultures that Eibl-Eibesfeldt made between 1975 and 1983: Eipo (West New Guinea), Trobriand (Papua New Guinea), and Yanomami (Venezuela). In this analysis they aimed at describing the temporal and structural organisation of brow raising and how it is connected to other facial movements. They also tried to define the context in which brow raising occurred, in an effort to assess its basic meaning. To annotate the eyebrow raises in these recordings they used the Facial Action Coding System (Ekman and Friesen, 1978). This is an observational coding system that provides a way of identifying and classifying all visually discernible facial movements. The Facial Action Coding System will be

described in more detail in section 2.6. Grammer et al. (1988) coded all visible head and facial movements according to their onset, apex, and offset. Using the respective fieldnotes and cues from the film sequence, the context of the movement was described as age and sex of the receiver of the facial signal, and also as an opening situation (introduction of interaction) or as an interaction during a conversation. The number of brow raises studied for each cultural group was 80 in Eipo, 84 in Trobriand, and 91 in Yanomami.

To analyse the temporal properties of the brow raises they calculated the medians (due to considerable skew of the distributions) for total duration, onset, apex, and offset times. Their temporal structure was described as a fast onset (80 msec), followed by a variable apex time (the maximum contraction) and a slow offset (120 msec) where the brows returned to their starting point. The three cultural groups only differed in apex duration. In terms of total duration, the distribution suggested that there could be two different types of brow raises: a short one (up to approximately 800 msec), which was the most frequent type, and a long one (typically more than 1200 msec). When comparing contexts, they found that in the three cultures brow raises at the beginning of social interactions had a longer total duration than those occurring during interactions. Looking at the co-occurrence of other facial movements, brow raises were mostly accompanied by smiles.

Assessing the possible function of brow raising in its facial context Grammer et al. pointed out it could stress the meaning of other social signals, mostly positive signals like smiling. And it could also be combined with "single verbal utterances" to mark their meaning for the interlocutor. Thus they interpreted the eyebrow flash as a universal "social marking-tool" and concluded that it may have received this function through its prominent position in the face guaranteeing its visibility.

The possible social functions of eyebrow raising will not be addressed in the analysis of the current thesis. However, it is interesting to notice the reported temporal properties of the brow raises in the study above, such as the tendency for a fast onset. According to Grammer et al. this represents a marked change in the behavioural flow which is necessary for a particular behaviour to become

a stimulus interpretable by a perceiver. Another interesting and relevant finding is the fact that brow raising was longer at the start of interactions.

2.2.2 *Facial movements as social signals in dialogue*

The social communicative functions of facial movements have also been emphasised by psychologists Bavelas and Chovil (1997). They described facial displays, a term they prefer to facial expressions, as "active, symbolic, components of integrated messages (including words, intonations, and gestures)" (p. 334). They believe that "although they often depict emotional reactions by self or other, they are not emotional expressions; they signify rather than reveal" (p. 337). Since most facial displays occur in social interaction, these authors emphasise that they must be studied in dialogue, in order to explore their communicative functions and their meaning in context.

Chovil (1989, 1991a) carried out the first such systematic study of facial displays produced by pairs of subjects engaged in spontaneous conversation. Her study aimed at providing an alternative to the view that facial movements were related to underlying emotions that presumably caused the facial displays. She explained that the difference in her approach was that, as in ethological studies, hers was based on a model of communication, and instead of looking *within* the individual to interpret the facial display she looked *outside* to the social interaction for understanding. She believed facial displays were produced in social interactions to convey information that could be used by others. In one of the studies in her doctoral thesis (Chovil, 1989, 1991b) she tested the extent to which social factors regulated the occurrence of facial displays. She hypothesised that if facial displays were socially elicited, then their frequency of occurrence would be affected by changes in the social nature of the situation. To test this, participants (all female) were videotaped as they listened to another person relating a personal experience in four different conditions: (a) Tape-recording (alone listening to a tape recording), (b) Partition (listening to a participant co-present in the room but separated by a partition), (c) Telephone (listening to another participant over the telephone), or (d) Face-to-face (listening face-to-face to a co-present participant). The relative level of sociality for these conditions was determined by the ranking (1 to 4) provided by another group of 65 participants who considered

(a) to be the least social, and (b), (c), (d) as increasing in sociality, with average rankings of .14, 1.28, 1.81, and 2.81, respectively. Ten dyads were recorded in each interactive condition and ten participants alone in the non-interactive, tape-recording condition. The frequency of motor mimicry displays by the listeners was measured for each condition. Paralleling the sociality rankings, the mean number of displays increased from (a) to (d) (0.11, 0.32, 0.71, and 1.14, in each condition respectively) and a linear contrast between the two sets of scores (sociality and frequency of displays) proved to be significant. A contrast was also found between the number of displays in the three interactive conditions and the tape-recording condition, with the latter having significantly less mimicry displays than the other three conditions. Finally, visual availability appeared as a significant factor determining the likelihood of facial displays, with the face-to-face condition having more displays than the three non-visual conditions. In summary, the degree to which individuals could interact was found to affect the extent to which listeners exhibited motor mimicry displays, and these displays were affected by actual presence and visual availability of the story-teller, which increased their frequency. From this she concluded that facial displays are more frequent in social interactions and they have an important role in conveying messages to others in face-to-face communication.

In another study in her thesis, summarised in Bavelas and Chovil (1997) and Chovil (1991a), she used an inductive approach to investigate what kinds of communicative information could be provided by facial displays in conversation. She identified some linguistic functions in which eyebrow raises were the most frequent displays. This study is more relevant for the current thesis, and will be presented later, in the review of research of eyebrow movements in relation to linguistic phenomena (2.5).

In summary, in this section I have reviewed the work of some researchers who criticised the involuntary expression of emotional states as an explanation for facial behaviour. Instead, they proposed an alternative explanation that assigns social communicative functions to these movements. Within this approach a relation between facial activity and the verbal message has been suggested. This is the relation that will be investigated in the current thesis, for eyebrow raises, but only from a linguistic perspective, leaving aside possible social functions.

As we will see later, a relation with the verbal message has been explored more extensively for non-facial body movements, such as hand gestures. Some of this work will be presented in section 2.4 as an introduction to the more relevant studies on eyebrow raising and speech in section 2.5. But before that, I will provide below some linguistic background for the kind of linguistic functions that have been investigated in relation to body movement.

2.3 Linguistic background for functions associated with body movement

The purpose of this section is to introduce some linguistic background for the functions that have been associated with body movement and eyebrow raises in particular. I will deal with discourse structure and utterance function, intonational prominence, and information structure. These will play a central role in the current analysis. The presentation here will be from a purely linguistic point of view. Body movement will be discussed in section 2.4.

2.3.1 Discourse structure and utterance function

When we engage in a conversation we do not produce isolated utterances at random. Our utterances are linked to other utterances, and in this way they convey meaning and allow communication. Thus, some utterances form groups that combine into larger groups to make up the structure of the conversation. A simple example was presented in the first chapter of this thesis (section 1.1). Several schemes have been proposed in the research literature to segment the discourse into smaller units, normally as a tree structure. One criterion that has been used to analyse discourse structure is the utterance purpose. That is, utterances can be connected according to the speaker's purpose in producing them. In a computational approach to dialogue analysis, Power (1974, 1979) studied conversation in terms of the underlying goal of utterances. He proposed the notion of "conversational procedures" using a computer model of conversation. In his computer model a programme generated a conversation between two robots on either side of a series of doors. To move through the doors they had to cooperate with the other robot, since doors could be bolted on the opposite side. He proposed

several "conversational procedures" in which the robots announced their intention to move, they requested help, etc. Developing Power's model, Houghton (Houghton, 1986; Houghton and Isard, 1987) proposed four "interaction frames" used to accomplish simple goals in a similar robot simulation: getting attention, providing information, requesting information, and accomplishing an action. Based on Power's and Houghton's work, the Conversational Games Analysis (Carletta et al., 1997) was developed to describe the structure of real human dialogues. Although it can be applied to different types of discourse, it was originally developed for task-oriented dialogues. Conversational Games Analysis is the scheme used in this thesis to describe the structure of the dialogues under investigation (other schemes will be mentioned later in sections 2.4.3 and 3.3.1).

The basic unit of the Conversational Games Analysis is the conversational move, an utterance classified by its function. The next level up is the game, in which a sequence of moves pursue and finally achieve or abandon a goal. Games make up larger structures at the highest level, transactions, which correspond to a step in the extra-linguistic task which the dialogue aids. Conversational Games Analysis will be described in more detail in Chapter 3 and is applied in Chapter 4 to investigate the distribution and possible function of eyebrow raising across the discourse structure. An association between body movement and discourse structure has been suggested in previous studies, e.g. for hand gestures, body shifts, and eyebrow raises. This will be discussed in 2.4 and 2.5. Before that, I will introduce other linguistic functions that have been suggested for facial and other body movements: prominence and information status marking. To discuss these we move into another area of linguistics, namely prosody.

2.3.2 *Intonational prominence*

Utterances take their meanings from words, word order, and how words are said. As we will see later, body movement, particularly brow raising, is thought to be associated with the phonological prominence of words. In a linguistic message not all words have the same weight. Some words stick out and are more "prominent" than others. This does not happen at random, it is part of the linguistic meaning. For instance, imagine a conversation where someone has just mentioned a person named Susana who is unknown to the interlocutor. When asked *"who is Susana?"*, this someone answers *"she is a Spanish teacher"*. Now,

with the phrase "Spanish teacher" he can mean one thing if he makes "Spanish" prominent (Susana comes from Spain and is a teacher) and something else if he makes "teacher" prominent (Susana teaches Spanish). This illustrates how different prominence patterns can convey different meanings on the same string of words. Ultimately, prominence is a perceptual phenomenon always determined in relative terms from a relation between weak and strong elements in an utterance, but how exactly it is realised phonetically is not a simple issue. It is associated to suprasegmental features of the linguistic signal, that is, fundamental frequency (F0), intensity, and duration, and therefore to intonation, which is described by Ladd (1996) as the use of these "suprasegmental phonetic features to convey 'postlexical' or sentence-level pragmatic meanings in a linguistically structured way" (p. 6). In English, major sentence-level prominence is associated with the occurrence of a pitch accent on the prominent word.

The definition of pitch accent is not always clear, but there is very good agreement (80.6%) on the identification of presence/absence of pitch accents by listeners (Pitrelli et al., 1994). A pitch accent can be defined as "a local feature of a pitch contour - usually but not invariably a pitch change, and often involving a local maximum or minimum - which signals that the syllable with which it is associated is prominent in the utterance" (Ladd, 1996, p. 46). Pitch accents are elements of intonational contours, which are often analysed in terms of two distinctive levels: High and Low. They often consist of an F0 peak (High tone), or valley (Low tone), or a combination of these two levels or tones (but High and Low tones need not imply peaks or valleys and can refer to prominent syllables in relatively smooth stretches of F0).

Stress is also related to prominence and can be interpreted as "acoustic salience" that can be cued by the presence of a pitch accent but also by increased intensity and duration. Stress is one of the most difficult concepts to define in intonation, because it is "a complex perceptual amalgam only indirectly relatable to psychophysical and physical dimensions" (Ladd, 1996, p. 6). In English, stressed syllables are often accompanied by pitch accents and some studies have used both terms interchangeably (different approaches to intonation have even used them in opposite ways). Ladd (1996), following the autosegmental-metrical theory, makes a distinction between them, and points out that pitch accents do not

represent the acoustic realisation of stress (for further discussion see Ladd, 1996, pp. 46–51).

The second study in this thesis deals with pitch accents without specific reference to stress and without specifying their High and Low constituent tones. It does, however, consider them in relation to the phenomenon of downstep. Downstep refers to a relation between two like tones (usually High tones) in a sequence where the second one is realised lower than the first one to an extent that cannot be accounted for by some background declination (the tendency for F0 to decline over the course of an utterance). The notion of downstep was first applied to English by Pierrehumbert (1980), inspired by many sub-Saharan African languages where in a sequence of High-Low-High tones the second High is lower than the first, and subsequent High tones in the same intonation unit are gradually realised lower. The function or meaning of downstep has not been much discussed. Ladd has pointed out that "downstepping adds a nuance like finality or completeness, but does not make the accent 'less prominent' in the way it affects the focus of the phrase" (1996, p. 76).

Prominence has been associated with another linguistic function briefly described below: the marking of information structure.

2.3.3 *Information structure*

Information structure refers to the structure of a linguistic message into units of information with different relationships to previously presented information. A common distinction traditionally made in information structure is that between *new* and *given* information. This distinction was adopted and developed by Halliday (1967b), following the Prague School that worked within the "functional sentence perspective". New/given information can be defined respectively as "information that the addressor believes is not known to the addressee" versus information "which the addressor believes is known to the addressee (either because it is physically present in the context or because it has already been mentioned in the discourse)" (Brown and Yule, 1983, p. 154). In the imaginary conversation above, an interlocutor asked "who is Susana?" after Susana had been mentioned by the main speaker. The first time "Susana" was mentioned would be an example of new information, whereas in the question "Who is Susana?",

it would be given information. As Brown and Yule explained, Halliday believed that in English speakers marked the new/given distinction by means of intonation[1]. Others studied information structure in terms of its syntactic form. In their presentation of different approaches, Brown and Yule concluded that although there are no rules for the specification of new/given information, there are some "regularities". For instance, new entities are usually introduced by indefinite referring expressions and intonational prominence. However, as they pointed out, information structure is very difficult to pin down formally, and linguistic form alone cannot determine information status: "if we have to rely on linguistic forms alone to determine information status, it seems that the relevant status will not always be clearly marked and, indeed, if syntactic and intonational forms are both regarded as criterial for 'givenness', that these forms may supply contradictory information to the hearer." (Brown and Yule, 1983, p. 188).

If the linguistic form of a message does not always provide conclusive cues for its information structure, perhaps in face-to-face communication it is facial cues that we should attend to in order to understand how information status may be expressed. This thesis investigates a potential role of eyebrow raises as information structure markers in face-to-face dialogues.

2.4 Observations on linguistic functions of body movements

In section 2.2 we saw how some researchers have proposed the study of facial movements in terms of their communicative functions, rather than the expression of emotion. Their emphasis was on social functions, but in this context facial behaviour has also been linked to the verbal message and researchers have emphasised the need to study it in interaction and in relation to language (e.g. Bavelas and Chovil, 1997). Other body movements have also been associated to the verbal message. This section presents studies on body motion in relation to the verbal channel. The purpose of this presentation is to show what kind of linguistic phenomena have been related to body movement in previous research and how this encourages and lends support to the study of eyebrow raises within the linguistic context.

[1]But see Bard and Aylett (1999), who found that decreased intelligibility and length of second mentions of entities in their task-oriented dialogues was not due to deaccenting of the referring expression, which only occurred 15% of the times

Three aspects of the linguistic signal are important here (and were introduced in the previous section): intonation and prominence, discourse structure, and information structure. Another important issue that will be relevant for this thesis is the temporal alignment between body motion and linguistic events. Research on eyebrow raises is postponed until section 2.5, where special attention will be given to those same aspects of the linguistic message.

2.4.1 Body movement structure and its alignment with the linguistic signal

Birdwhistell, like the researchers discussed in section 2.2.2, emphasised the importance of studying non-verbal behaviour in its social context. Within this approach, he looked carefully at the structure of body movement. He microanalysed filmed material and observed body motion as an important part of the communication process. According to Birdwhistell, communication is a continuous process in which one or more channels of all sensory modalities are always in operation (Birdwhistell, 1970). He explained that communicative body motion occurs at the kinesthetic-visual channel and is studied by kinesics, of which he was a pioneer. He defined kinesics as "the systematic study of the communicational aspects of human body motion (Birdwhistell, 1952, p. 11, cited in Wiltshire 1999).

Birdwhistell's treatment of motion owes much to structural linguistics. Thus, by analogy with phones, allophones, phonemes, and morphemes, in the audio-acoustic channel, he proposed kines, allokines, kinemes, and kinemorphemes to describe the structure of body motion in the kinesthetic-visual channel. Kinemes are "building blocks with structural meaning" and as they are "combined into orderly structures of behavior in the interactive sequence they contribute to social meaning" (Birdwhistell, 1970, p. 99). He devised a detailed notation system to microanalyse body motion and in his study of "American movement" he isolated 32 kinemes in the face and head area, of which four are kinemes of brow behaviour: "lifted brow", "lowered brows", "knit brow", and "single brow movement" (p. 100).

Birdwhistell had a direct influence on Condon, another author who investigated body motion in relation to speech. Condon observed and described a phenomenon he termed "synchrony" (e.g. Condon, 1970; Condon and Ogston, 1971), by which

our body movements are synchronised with the speech segments we produce (self-synchrony) and even with the speech produced by another speaker when we are the listener (interactional synchrony). He described the flow of body movement as small waves (e.g. blinks and finger movements) within larger waves (e.g. arm movement). The point at which movement changed in the small waves, coincided with changes in the large waves. And he observed a synchronisation between this flow of movement and changes in speech. Thus he described a "rhythm hierarchy" in which movements were synchronised with different levels of the speech: the phone, syllable, word, phrase, half second, and second. For instance, changes in the fingers could be synchronised with phones, while wrist movements with the syllable, hand movements with words, and the whole arm movement with the whole phrase. The point at which the units of motion changed coincided with the boundaries of the units of speech.

Condon made his observations by very close inspection of recordings on 16mm sound film. With the use of a time-motion analyser he could advance the film frame by frame or across a series of frames, and compare the observed movements with the accompanying speech by means of an oscilloscopic display (frame-numbered) (Condon, 1979). Using modern computer equipment, Wiltshire (1999) found examples of self-synchrony at the onset of speech units in her data, and she attributed earlier failures at replicating the phenomenon to a lack of understanding of Condon's theory and methodology.

Condon's findings and methodology influenced Kendon, who, as we will see below, became a very important figure in the study of gesture[2]. Kendon, like other researchers introduced below, studied the alignment of body motion and speech, but he analysed linguistic units larger than Condon's and looked at suprasegmental aspects such as prosodic structure.

2.4.2 *Body movement and prosodic structure*

In his study of body motion, Birdwhistell (1970) associated some movements to linguistic stress, such as head nods, eye blinks, and thorax thrusts, among others,

[2]In the field of gesture studies, gesture is usually defined as "spontaneous bodily movements that accompany speech. The most common body parts used are the hands, fingers, arms, head. face, eyes, eyebrows, and trunk" (Loehr, 2004, p. 7). Most authors, however, have used gesture to refer only to hand and arm movements

as part of a kinesic stress system. He also agreed with Kenneth Pike that phonetic pitch "may contain some of the secrets of linguistic-kinesic interdependence" (p. xiv). But he did not go into that subject although he admitted that "it seems likely that some kind of systematic relationship exists between certain stretches of kinesic behavior and certain aspects of American English intonation behavior" (p. 128-129).

Kendon, as we mentioned above, studied the alignment of gestures and speech and paid particular attention to suprasegmental features. He concentrated mainly, but not exclusively, on speech-related movements of the hands and arms, which he referred to as "gesticulation". After detailed analysis of utterances from filmed conversations he established an intimate relationship between body movements and units of the prosodic structure of speech, as described below.

Kendon (1972, 1980) divided the prosodic structure into tone units as defined by Crystal and Davy (1969) (more or less a group of syllables with a complete intonation tune). Tone units combined into "locutions", that generally comprised complete sentences. Locutions made locution groups and these were organised into locution clusters, which were like paragraphs of the discourse marked by a pause or a change in voice quality, loudness or pitch range, and a shift in subject matter. Finally, locution clusters were combined into the highest level, the discourse, which in his studies corresponded to one speaker turn. The structure of gesticulation was described in terms of "gesticular units and phrases". A gesticular unit comprises the movement of the limb away from the body until it returns to its rest position. Within such unit the limb can make several distinct movements classified as gesticular phrases. These are distinguished for having several phases: the preparation (optional), the stroke (obligatory) or moment of most accented movement, and the recovery phase (optional), where the limb moves back to its rest position or is prepared for another stroke. Kendon found a close relationship between the prosodic structure and the kinesic structure. Table 2.1 presents the correlation between the different levels of both hierarchies. Each level in the prosodic structure was matched by a distinctive pattern of bodily movement. In his 1972 study, for the duration of the discourse the speaker maintained a body posture different from the one he sustained before and after it; for each locution cluster he used different arm movements; in each locution group he had consistent head movement patterns; and so on. In the 1980 study,

each locution was found to have its own gesticular unit, and tone units were matched with a distinct gesticular phrase (although the latter was a more complex relationship). From these findings Kendon claimed that "this bodily activity is so intimately connected with the activity of speaking that we cannot say that one is dependent upon the other. Speech and movement appear together, as manifestations of the same process of utterance" (1980, p. 208).

Kinesic Hierarchy	Phonological Hierarchy
Consistent arm use and body posture	Locution Cluster
Consistent head movement	Locution Group
One G-Unit	Locution
One G-Phrase	Tone Group
One Stroke	Most Prominent Syllable

Table 2.1: Correlation between kinesic and phonological hierarchies, from McNeill (1992), based on Kendon (1972, 1980)

Kendon (1972, 1980) looked at the temporal relationship between the units in both channels, acoustic and kinesic. He found that the phrases of gesticulation tended to precede their associated speech phrases. The stroke of the gesticular phrase was completed before the most prominent syllable in the tone unit or just at the onset of this accented syllable. According to Kendon this finding could suggest that the kinesic channel is easier and more readily called upon than the verbal channel, which would agree with the idea that language first appeared in the form of gesture (e.g. Hewes, 1973). McNeill (1992) also looked at synchronisation between hand gestures and speech and found the same temporal relationship that Kendon described. McNeill termed this the "phonological synchrony rule", by which the stroke phase of a gesture is completed before or at the accented syllable of the accompanying speech (with the optional preparation phase obviously preceding that syllable). This is a very interesting observation with relevance for the study in Chapter 5 in this thesis, where the alignment between eyebrow raises and pitch accents is investigated.

Intonation has been the object of much research on body movements. Bolinger (1983, 1986) claimed that intonation belongs with gesture, and that when the two occur together, they are not only synchronised but they also move up and down in parallel. That is, body movements, such as those of eyebrows, hands, head, or shoulders, go up and down in parallel with pitch rises and falls. He suggested

this up and down movement was due to emotional tension, and that we "read intonation the same way we read gesture ... We know how we feel when we are tense and we have already associated the high pitch of our own voice with that feeling; when we hear a high pitch from someone else, we infer tension. The fluctuations of pitch are to be counted among all those bodily movements which are more or less automatic concomitants of our states and feelings and from which we can deduce the states and feelings of others" (Bolinger, 1983, p. 157). But several authors who tested this parallel movement hypothesis did not find evidence for it. McClave (1998) found no significant correlation between pitch direction and the direction of manual gestures. But she did find an alignment between intonational phrases and gestural phrases, and that beats[3] coincided with the most prominent syllable in a tone unit, supporting Kendon's (1972; 1980) and McNeill's (1992) claims. Similarly, Loehr (2004), in the first study of intonation and gesture that used a full framework of intonational phonology (Pierrehumbert, 1980) measured acoustically, did not find evidence for Bolinger's hypothesis in the relation between gestures and pitch direction. However, he did find a strong relationship between the two modalities, manifested in the alignment of the apexes (the peak of the stroke) of hand movements with pitch accents, and of gestural phrases with intermediate phrases of intonation. Loehr (2004) included head movements in part of his analysis and found that, as for hand gestures, the direction of head movements in relation to pitch did not support Bolinger's hypothesis of parallel movement. What he found was a kind of common rhythm in which pitch accents, hand and head movements were all aligned at some points, though not very frequently, much like the instruments in a jazz music piece (in his own analogy).

2.4.3 Body movement and discourse structure

When analysing the speech signal, Kendon (1972, 1980) paid most attention to prosodic structure, but some aspects of his description are also related to discourse structure: locution clusters are described as "paragraphs of discourse" and they correspond to the speaker's main discourse themes. In a later paper, Kendon (1997) claimed gestures "can provide a visible indication of different "levels" of discourse structure" (p. 112).

[3]Beats have been termed "batons" by other authors such as Efron (1941) and Ekman (1979). They are typically simple flicks of the hands or fingers

McNeill agreed with Kendon that language is more than words: "gestures are an integral part of language as much as words, phrases, and sentences– gesture and language are one system" (McNeill, 1992, p. 2). These movements are part of the discourse and often "we can see the overarching discourse structure more clearly in the gesture than in the words and sentences" (p. 2). For instance, he noted that beats have a discourse function. Beats can vary in size but are typically simple flicks of the hands or fingers (up and down, or back and forth) occurring on stressed syllables. Although they may look "insignificant", "beats reveal the speaker's conception of the narrative discourse as a whole". A beat "indexes the word or phrase it accompanies as being significant, not for its own semantic content, but for its discourse-pragmatic content" (1992, p. 15). An example of its use is to mark the introduction of new entities or new themes into the discourse (McNeill, 1992; Levy and McNeill, 1992). This type of gesture is therefore related to both information status and discourse structure.

Other types of gesture, as well as beats, have also been related to discourse structure. Following Kendon's work, McNeill et al. (2001) studied possible cues for discourse structure from manual gestures (regardless of type) in videotaped conversations describing living spaces. The analysis described in their publication comes from a 32-second section in which a female conversant describes her living quarters to an interlocutor. Using a systematic procedure by Nakatani et al. (1995b)[4], the discourse structure of the transcribed text was recovered by a set of questions that revealed the speaker's goal in producing the utterances. As for gestures, hand movements were automatically traced with motion analysis techniques (Quek et al., 1999) and coded in terms of recurring form features. According to McNeill et al. (2001) and as suggested earlier by Kendon (1972, 1980), common discourse themes will produce gestures with recurring features. Therefore gesture will give clues for discourse structure. Indeed, by comparing their annotations of the different channels, McNeill et al. observed that recurring gesture features revealed a discourse organisation that correlated (100%) with the hierarchical structure derived from Nakatani et al's discourse annotation system.

[4]Nakatani et al. (1995b) presented a set of instructions to do discourse segmentation from text. These instructions are based on the theory of discourse structure by Grosz and Sidner (1986) and were prepared for naive segmenters who have not studied discourse theory or discourse processing methods

Head movements have also been investigated in the context of speech (e.g. Loehr, 2004) though not as much as hand gestures. McClave (2000), reviewing previous research on head movements, pointed out that Kendon (1972) was the first to observe that some head movements are connected to the discourse structure of an utterance. McClave (2000) carried out a microanalysis of two dyadic conversations between native speakers of American English (male-male, female-female). The subjects were asked to talk about topics of their choice for approximately an hour. Two cameras recorded their upper body while a third camera captured their full bodies. A timecode was generated on each tape to allow the coordination of the movements of one participant with those of the other. The analysis was carried out using a VCR machine with the muting device off to hear the sound as the film was advanced frame by frame. This allowed them to match the observed movements to the simultaneous speech. McClave found that the speakers moved their heads with great frequency. She described the patterns of movement observed and she associated several functions to these movements including semantic, discourse, and interactive functions. Among the discourse functions, she associated changes in head position with a switch from indirect to direct quotes and with listing or presenting alternatives. She described how "a speaker's head often will assume a new orientation slightly preceding or coinciding with the beginning of a quote in a striking parallel to American Sign Language (ASL)" (p. 863). And that in lists or alternatives "characteristically, the head moves with each succeeding item – often to a contrasting position" (p. 867).

Postural shifts, not restricted to the head, were investigated in relation to discourse structure by Cassell et al. (2001). They recorded subjects in "pseudo-monologue" and in dialogue. In the "pseudo-monologues" subjects first had to describe their homes and then they gave directions between four pairs of locations they knew well. The experimenter acted as a listener providing only backchannel feedback (thus the name "pseudo-monologue"). In the dialogues, two subjects had to generate an idea for a class project that they would like to work on. They were told to perform their task in 5–10 minutes. Cassell et al. analysed seven "pseudo-monologues" (29.2 mins.) and five dialogues (42.5 mins.). The data was transcribed and coded for: discourse segment boundaries, turn boundaries, and posture shifts. Following Grosz and Sidner (1986), a discourse

segment was taken as "an aggregation of utterances and sub-segments that convey the discourse segment purpose, which is an intention that leads to the segment initiation" (p. 108). In their analysis they decided to look at high-level discourse segmentation phenomena, using as segmentation points the time at which speakers started the assigned task topics. Turn boundaries were marked in the dialogues at the point in which the start/end of an utterance cooccurred with a change in speaker, excluding backchannel feedback. Finally, posture shifts were defined as "a motion or a position shift for a part of the human body, excluding hands and eyes", and they were labeled with their start and end time, body part involved, and an estimated energy level of the movement. Posture shifts were found to occur more frequently at discourse segment boundaries than within discourse segments in both monologues and dialogues. And they also tended to be more energetic at the boundaries. Turn structure also had an effect on the occurrence of posture shifts, with subjects five times more likely to make a posture shift at a boundary than within a turn. So both turn and discourse structure had an influence: speakers tended to generate a posture shift when initiating a new discourse segment, which was often at the boundary between turns. From this they concluded that posture shifts can signal *boundaries* of units. Their empirical findings were used to derive an algorithm for generating posture shifts in an animated embodied conversational agent (Cassell et al., 2000a)[5] with the aim of improving the naturalness of this dialogue system.

2.4.4 *Body movement and information structure*

As mentioned above, beats have been related to the information structure of a message by marking the introduction of new entities or themes into the discourse (McNeill, 1992; Levy and McNeill, 1992). Another behaviour that has been studied in connection with information structure in speech is gaze. Cassell et al. (1999) explained that in previous research gaze behaviour was associated to turn-taking behaviour in conversation: the speaker *looks away* from the interlocutor to keep the floor and *looks at* the interlocutor to give up the floor (e.g. Duncan, 1972, 1974; Goodwin, 1981). However, Cassell et al. found that gaze behaviour is better explained in terms of both turn-taking and information structure.

[5]For more details on embodied conversational agents see section 2.7

They collected data from three pairs of subjects engaged in conversation. All subjects were native speakers of North American English and were unfamiliar with each other. They were asked to talk about a topic of their choice for at least 20 minutes while their upper bodies were recorded by two cameras. Cassell et al. selected 100 turns from these conversations and transcribed the verbal and nonverbal behaviour of the two interactants. The annotated verbal behaviour consisted mainly of words and pauses, and the nonverbal behaviour was mainly the speaker's gaze (beginning of a look away from the listener and of a look toward the hearer) and listener's head nods. Three units of analysis were employed: turns, themes, and rhemes. Following Halliday (1967a) they defined the theme as "the part of the utterance that links it to the previous discourse and specifies what the utterance is about", and the rheme as the part that "specifies what is contributed to the discourse with respect to the theme ... [it] specifies what is new or interesting about the theme" (Cassell et al., 1999, pp. 146–147)[6]. An example of theme and rheme provided by Cassell et al. is in the following two turns:

Q: *What do you do?*

A: *I work with Mike B.*

where, in the answer to the question, *I work with* would be the theme and *Mike B.* would be the rheme.

By looking at the frequency of two gaze patterns, look-away and look-towards, Cassell et al. found support for previous claims in the research literature: of all the turn beginnings in their data, 44% were accompanied by look-aways. But as Cassell et al. hypothesised, a stronger pattern was found when looking at gaze in information structure behaviour: in 70% of the parts of utterances that were labeled as *theme*, the speaker initially looked away from the hearer. And more interestingly, in the beginning of a theme that coincided with the beginning of a turn, the speaker *always* looked away. As for gaze directed towards the listener, previous researchers claimed the speaker showed this behaviour at the end (or near the end) of turns. Cassell et al. found this occurred in 16% of all their ends of turns, whereas at the beginning of rhemes, the speaker looked towards the listener in 73% of the cases. And again, in the beginning of a rheme that coincided

[6]Halliday (1967b) also talked about another distinction in information structure discussed in section 2.3.3 above: given/new information. While theme/rheme is a contrast at sentence level, given/new information applies to words.

with the end of a turn (specifically, within one word of the end) the speaker *always* looked at the listener. From this, Cassell et al. concluded that "turn-taking only partially accounts for the gaze behavior in discourse" and that "a better explanation for gaze behavior integrates turn-taking with the information structure of the propositional content of an utterance". As an explanation they suggested that speakers looked toward hearers when new information or the key point of their contribution is being conveyed (at the beginning of the rheme) and this may focus the attention of speaker and hearer on this key part of the utterance. And also that this is not entirely independent from turn behaviour, because speakers may be more likely to give up the turn once they have conveyed this important material of their contribution.

To summarise, in section 2.4 we have seen body movement studied in relation to linguistic phenomena. Studies have shown that some body movements are aligned with certain linguistic events indicating a connection between them. In this alignment the movement usually starts before the word or phrase with which it is associated. The link between the two modalities has also been studied in terms of prosodic structure, discourse structure, and information structure. Although not always based on a sound empirical approach, the conclusions encourage further research, particularly into whether facial movements can be related to these linguistic phenomena. The current thesis investigates these issues in relation to eyebrow raises. In the following section I will review previous studies of eyebrow raises in connection with speech.

2.5 Eyebrow raises and speech

In this section I will present what we know from previous research about eyebrow raises in relation to speech. First, I will summarise some observations that were presented in the past without much supporting evidence (2.5.1). Then I will describe studies that have more recently investigated empirically a possible relation between eyebrow raising and linguistic phenomena: first some production studies (2.5.2) and then perception studies that have used synthetic stimuli (2.5.3). Of special interest is the connection between eyebrow raises, on the one hand, and discourse structure, utterance function, information structure, and intonation, on the other.

2.5.1 *Observations in descriptive studies*

Several of the authors presented in section 2.4 mentioned eyebrow movements in their observations about body movements. Birdwhistell (1970) identified four kinemes of brow behaviour, one of which was the "lifted brow". Condon (e.g. Condon, 1979) also included eyebrow movements, which would be like the small waves in his description of the flow of motion, synchronised with small units of speech. Kendon has also pointed out the importance of studying facial movements, including eyebrow movements, in interaction (Kendon, 1975). Bolinger (1983) claimed that head and facial movements are regularly coupled with intonation because of their proximity to the vocal organs, and he specifically mentioned eyebrow movements as one of the gestures that go up and down in parallel with pitch fluctuations. Loehr (2004) did not investigate facial movements but he mentioned that these are certainly related to intonation and he added that eyebrows, in particular, are worth studying. All these comments about eyebrow raising, however, were not based on quantitative systematic studies. In fact, empirical investigation of eyebrow movements during speech is very scarce even today.

Section 2.1 above dealt with research that saw facial movements as expressions of emotion. Ekman, one of the representatives of this view, published an influential paper about brow movements not only as emotional signals but also conversational (Ekman, 1979). He described conversational functions of eyebrow movements of speakers and listeners, and of brow movements with no accompanying words. But, as with the authors above, he made these observations without presenting supporting evidence. As he himself warned, then, they should be considered "preliminary, tentative, and only a suggestion about what may be found" (p. 183).

Like Birdwhistell (1970) and Eibl-Eibesfeldt (1972), Ekman observed that brow movements could sometimes be used for emphasis, and he referred to these movements as *batons* and *underliners*. Baton is a term from Efron's (1941) classification of gesture that corresponds to McNeill's "beat". Batons have to do with the tempo of speech and are used to emphasise words. Underliners, also provide emphasis but stretched over more than a single word. Ekman noted underliners coincided with speech changes such as sustained loudness, increased

pauses between words, or stretching out of words, all of which are used for emphasis. For the annotation of facial movements in his studies, Ekman used the Facial Action Coding System (Ekman and Friesen, 1978). This observational system allows the annotation of all visually distinguishable muscle movements in the face and provides a classification into different "action units". The Facial Action Coding System will be described in more detail in section 2.6. The different action units (AUs) of the brow/forehead are presented in Figure 2.2 (taken from Ekman's publication). From these, Ekman observed AU 1+2 as the most frequent baton, followed by AU 4, and much more rarely AU 1+4. These were also the most common brow actions used as underliners.

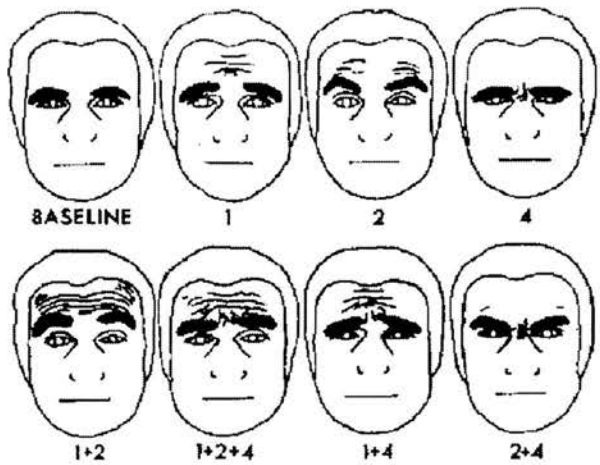

Figure 2.2: Action units for the brow/forehead, after Ekman (1979, p 174)

Those brow actions could also have other conversational functions, according to Ekman. Apart from emphasising, they could function as "punctuation marks": as commas, for instance when inserted in a pause between each of a series of events being described, or as a period or exclamation mark when in a juncture pause at the end of a phonemic clause[7]. In addition, like Birdwhistell (1970) and Eibl-Eibesfeldt (1972), he pointed out the use of brow movements as question markers. And another use by speakers was to hold the floor in conversation during word search. Listeners' eyebrow movements could function as agreement responses or requests for information, if for instance what the speaker said was not understood. Finally, both speaker and listener could use eyebrow actions as

[7]Phonemic clause is a term coined by Trager and Smith (1951) that corresponds to *prosodic phrase* or *intonation phrase*

emblems, signals that have specific semantic meaning that can be understood without words.

Although Ekman's publication has been influential in later research, his observations were presented without supporting evidence, as was mentioned above, and he did not pursue this line of research further. The lack of empirical studies remained very much the same for years. More recently, though, a few researchers have studied possible linguistic functions of eyebrow raising using a quantitative, empirical approach. These will be presented below.

2.5.2 Empirical production studies

As mentioned in section 2.2.2, Chovil (1989, 1991a) carried out the first systematic study of faces in spontaneous dialogue. She analysed facial displays (a term she prefers to "expressions") in terms of the information they provided to investigate their contribution to the production of messages in conversation. Twenty-four subjects were videotaped conversing in pairs: four female, four male, and four mixed pairs. The average length of their conversations was 11 minutes and 25 seconds. They were given three topics to discuss: planning a meal together, retelling a conversation involving a minor conflict or argument between themselves and another person, and describing a close-call situation they had experienced or heard about. The videotapes were then analysed using a large monitor and an industrial-quality VCR. Any noticeable movement or change in one or more areas of the face was marked as a facial display, excluding movements that were byproducts of other actions such as speech articulation, blinking, swallowing, inhaling, and laughing. Adaptors, however, such as biting a lip or wiping the lips with the tongue, were included as facial displays, whereas smiles occurring with no other facial action were not because of their high frequency, which according to Chovil would have overwhelmed the other types of displays. For every facial display the following information was taken: time of occurrence, who made it (speaker/listener), general description of its most obvious actions (e.g. eyebrow raising, eye squinting, etc.), a transcript of the verbal content surrounding the display, including the words with which it started and ended, and finally, whether the information conveyed by the display was also conveyed by the accompanying words. With an inductive approach Chovil questioned what the display was doing at that point in the conversation and how it conveyed

meaning in that context. In this way she classified facial displays into different categories. Reliability scores on 20% of the data from a second scorer showed 82% to 97% agreement. Table 2.2 lists Chovil's general categories as published, with their percentage and raw frequencies of occurrence.

Linguistic Categories		
Syntactic	27%	(315)
Semantic Speaker (Redundant)	21%	(243)
Semantic Speaker (Nonredundant)	14%	(162)
Listener comment	14%	(160)
Nonlinguistic Categories		
Adaptor	25%	(301)
Not assigned a Category	< 1%	(3)
Total		(1184)

Table 2.2: Distribution of facial displays across general categories, after Chovil, 1991a, p. 175

The most frequent displays, and the most relevant for this thesis, were the ones she classified as *syntactic* (1991a, p. 175):

> These were facial displays that (a) appeared to mark stress on particular words or clauses, (b) were associated with syntactic aspects of an utterance or (c) were associated with the organizational structure of the talk (e.g initiation of topics).

The most common actions within this category were eyebrow movements (raising or lowering). Table 2.3 lists the frequencies of the specific kinds of syntactic displays found in her data.

In a total of 315 syntactic displays, the most common was the *emphasizer*, also observed by Birdwhistell (1970) and Ekman (1979), which "occurred with a stressed (prosodically marked) word in an utterance" (1991a, p.177). Next in frequency were the *underliner*, as observed by Ekman (1979), followed by the *question marker*. Another interesting group of displays, but with a much lower frequency, were

Specific Category	% of Total Syntactic Displays	
Grammatical Markers		
Emphasizer	50%	(156)
Underliner	18%	(57)
Question Marker	14%	(45)
Offer	4%	(13)
Sentence Change	3%	(9)
End of Utterance	2%	(5)
Comma	< 1%	(1)
Organization of Story		
Story Announcement	2%	(5)
Story Continuation	6%	(18)
End of Story/Topic	1%	(3)
Topic Change	< 1%	(1)
Speech Corrections		
Pronunciation Correction	< 1%	(1)
Self Correction	< 1%	(1)

Table 2.3: Distribution of facial displays across the specific syntactic categories, reproduced after Chovil, 1991a, p. 176

those that seemed to help structure the conversation. These marked the beginning, continuation, and end of a story or topic. The most frequent function within this subgroup was marking the continuation of a story/topic after detracting from the main point. Two other types of displays that were not very frequent were displays that marked the end of an utterance and those marking "sentence changes", in which a speaker began to say something but then decided to express it differently and made a facial display at the point of change. There are five other subtypes described by Chovil within the syntactic facial displays (see Table 2.3), but these are either too infrequent or not relevant here.

Chovil's study is very important because it emphasised the need to study faces in dialogue and in relation to the verbal channel. It also provided evidence for the importance of eyebrow movements among the different facial actions. Some of their functions that are relevant for the current thesis were to mark emphasis on words and on longer utterances, to signal questions, and to mark the structure

of the conversation. In her methodology, Chovil presented frequencies of displays and the context in which they occurred, and it was by looking at these that she associated certain displays with certain conversational functions. However, she did not compare those occurrences to the frequencies of similar contexts that were not accompanied by a facial display. For instance, she did not provide the percentage of emphasised words, questions, or beginnings of topic that did not occur with a facial movement. Also, some of the types that she described had a very small frequency of occurrence. In addition, although Chovil made a good point by annotating the timing of the displays from the video recording, a more refined annotation of both the auditory and visual channels is necessary in order to minimise possible perceptual mistakes in their temporal relation. Modern technology now facilitates a more detailed annotation of this kind of data.

In conclusion, Chovil's work was ambitious in looking at different types of facial displays and with a number of subjects considerably large for this type of study. She made very interesting observations and encouraged the study of facial movements in dialogue. Eyebrow raising appeared as one of the most relevant movements in relation to the verbal message. But her methodology could lead to subjective conclusions and so further investigation should be done with a different method to study further some of her claims. In the current thesis, eyebrow raises in dialogue were studied using the hypothetico-deductive method, and using precise temporal measures for both auditory and visual stimuli.

A deductive approach has been used by a research group investigating the relation between facial expressions and voice variations. One of their studies (Cavé et al., 1996) looked at the relation between eyebrow raises and fundamental frequency (F0) variations in French. They made recordings of subjects talking to an interviewer. Their eyebrow movements were automatically recorded with a system (Elite) that captured the trajectories of small infrared markers attached to the subject's skin (Ferrigno and Pedotti, 1985). The eyebrow movement curve was displayed with the simultaneous F0 curve (if accompanied by speech). Cavé et al. (1996) analysed "rapid rising-falling movements" (Cavé et al., 1993) of at least 3mm displacement or more for at least one eyebrow. These were annotated in terms of duration and magnitude of the movement. Results for three subjects with a total of 78 movements showed a large degree of speaker variability in magnitude but not in mean duration (376 msec, s.d. not reported). But the latter

could well be an artefact of selecting rapid movements to begin with. Thirty-eight percent of the movements were made during pauses and from this fact Cavé et al. suggested these could have been used as back-channel or turn-taking signals. The 48 movements that occurred during speech (62% of the cases) were accompanied most frequently (71%) by intonation patterns containing a rise in F0: rising, rising-falling (the most common), and falling-rising. But because this was not always the case, that is, because some eyebrow raises occurred during "flat or slightly rising pitch patterns", they concluded that eyebrow movements and F0 changes are not the result of muscular synergy but result from linguistic and communicational choices. Given the small set of data used and the lack of specificity in their report of the analysis, their conclusions should be treated with caution until further evidence can be provided. For instance, they only present raw frequencies of pitch contours accompanying eyebrow raises, and they do not specify the frequency of those contours in the dialogues in general. However, if reliable, their findings would suggest that eyebrow raises play some role in the production of linguistic messages. Also, their empirical approach is important, since previous studies were mostly limited to non-empirical observations.

In a later paper Cavé et al. (2002) used the same analysis tool to investigate whether the same kind of rapid eyebrow raises could have a role in turn-taking. They measured the interval between the eyebrow movements and the beginning and end of speaking turns and found that eyebrow raises occurred significantly closer to the beginning than to the end (of the turn in which they occurred or of the next one when the brow movement was produced during silence). From this they concluded that eyebrow raises cue a new speaking turn. But although they reported a significant statistical result, they did not present the statistical test and they did not describe their sample size. In the same paper they also investigated the relationship between eyebrow movements and fundamental frequency. They compared the number of eyebrow movements in F0 contours with accentuating and non-accentuating values. Accentuating F0 contours were described as "those that contained points where there was a change in direction, called target points". As in their previous paper, they explained that although 93.75% of the movements occurred with accentuating contours (mostly rising-falling contours), the fact that 6.25% occurred with non-accentuating ones suggested that eyebrow movements and F0 variations were not automatically linked but instead

were controlled by the speaker for communicational purposes. Again, they presented percentages but they did not describe their sample size. Also it is not clear how they did the annotation of fundamental frequency contours and how they determined co-occurrence of brow raises and F0 pattern variations.

Cavé et al. (1996, 2002) used an automatic motion analysis system (Ferrigno and Pedotti, 1985) that reconstructed the trajectories of infrared markers glued onto the subject's skin. While this seems an accurate and very objective way of measurement, the markers could have the disadvantage of interfering with the natural movements, and making the subject aware of the purpose of the experiment. I will return to this issue in section 2.6.

The studies by Cavé et al. (1996, 2002) have the merit of being the first and practically the only empirico-deductive investigation to date on the production of natural eyebrow movements in relation to the speech signal. Their results are based on French. The purpose of the current thesis was to investigate eyebrow raises in connection with the verbal message in English.

2.5.3 *Perception studies with synthetic stimuli*

A number of recent studies have investigated the perception of eyebrow raises during speech to reveal their communicative functions. These studies have mostly used synthetic stimuli implemented in computer-animated talking heads [8]. Their method consists of implementing in the talking head observations from the literature of the kind discussed in the previous section. These are then tested in perception studies, making it possible to investigate further the claims in the literature and also to assess the naturalness of the talking head. Their conclusions are important for the current thesis and will be summarised below.

Granström et al. (1999, cited in House et al. 2001) found that words and syllables with accompanying eyebrow raising in a talking head in Swedish were perceived as more prominent than syllables without it. Following up on this, House et al. (2001) investigated both eyebrow raises and head movements as potential cues to prominence in Swedish, and the effect of varying the timing of such movements in relation to an accented syllable. Their synthesised test sentence was

[8] Computer-animated talking heads can be used in multimodal communication systems as "embodied conversational agents". For more details, see section 2.7 below

Jag vill bara flyga om vädret är perfekt, which in English would be "I only want to fly if the weather is perfect". The words *flyga* ("fly") and *vädret* ("weather") were acoustically accented. They used two sets of synthetic visual stimuli: one where the talking head showed synchronous eyebrow and head movements, and another one with asynchronous movements. Within each set they made six different stimuli by placing the movements in various positions between the two accented words in the test sentence, while keeping the acoustic signal and the articulatory visual movements constant. Thirty-three subjects were asked to listen to each stimulus while looking carefully at the talking head displayed on a computer screen. They were requested to choose which of the two words, *flyga* or *vädret*, was most prominently accented. They were also asked to indicate, on a scale from 1 to 5, how confident they were about their choice. House et al. found that both eyebrow and head movements cued prominence when synchronised with the accented syllable. When the movements were placed on different positions, both could act as independent cues to prominence, but head movements had a slight advantage. They suggested this could be explained by the larger surface involved in head movements compared to eyebrow raises. Also they found that the perceptual sensitivity to the timing of the visual stimuli in relation to the accented syllable was around 100 msec, showing that complete synchronisation with the syllable is not necessary for visual and auditory stimuli to be integrated.

House et al. (2001) described the synthetic eyebrow raises they created as being subtle movements, distinctive but not too obvious, with a duration of 300 msec, divided into: a 100 msec onset, a 100 msec static portion, and a 100 msec offset. Using a movement that is not too obvious seems a better choice than the exaggerated movements sometimes found in computerised talking heads. But we do not know if the temporal features they used are the most appropriate (for rise, static portion, and lowering of the eyebrows) since they are not reported as resulting from empirical observation. According to findings by Grammer et al. (1988) (section 2.2.1 above), a slightly faster onset and a longer offset might be more appropriate. Also, House et al's findings are for Swedish, a language where pitch can be used differently from other languages including English. Below I will describe some studies done on Dutch, which is prosodically very close to English, the language under investigation in this thesis.

Krahmer et al. (2002b) explored the signalling of important bits of information via pitch accents and visual cues in Dutch. In particular, they investigated eyebrow movements and pitch accents in relation to the perception of focus. Focus is related to information status. A piece of information is said to be in focus when it is new or contrastive with some other information in the surrounding context. In Dutch and other Germanic languages such as English, information in focus is made prominent usually by means of a pitch accent. Krahmer et al. used a computer-animated talking head with both synthetic and natural voices (two synthetic and four human). The natural voices were obtained from an earlier production experiment, in Dutch, in which a participant described geometrical figures to another participant (Krahmer and Swerts, 2001). From these recordings, Krahmer et al. collected instances of the phrase *blauw vierkant* ("blue square") with focus on one or both words. The words were in focus when they contrasted with the previous description. That is, if the figure that had been described previously was, for instance, a red square, then when describing the blue square *blue* would be in focus; if the previous figure had been a blue triangle, then *square* would be in focus; and if it had been a black triangle, i.e. not blue and not a square, then both words would be in focus. Acoustically, focus was marked by a pitch accent, both in the natural and synthetic voices. Visually, it was marked by raised eyebrows on the talking head. When both words were in focus, only one was marked with eyebrow raising. But some of the stimuli were created to have conflicting information, by having raised eyebrows on words that had no pitch accent, and vice versa.

Twenty-five native speakers of Dutch watched and listened to the talking head uttering the phrase "blue square" in the different conditions. Their task was to choose, out of three possibilities, what the preceding utterance would have described: a red square, a blue triangle, or a red triangle. This meant that they had to determine where the focus was on the phrase stimulus. The stimuli were displayed on a high-resolution computer screen and the participants could watch and listen to them as many times as they wanted. They were not told what kind of cues they should use for their task and they were not given any feedback about the "correctness" of their choices. There were a total of 36 stimuli (3 pitch accent distributions $\times$ 2 eyebrow versions $\times$ 6 voices). Krahmer et al. found that both auditory and visual information had an effect on the perception of focus,

but pitch accents were much more influential than eyebrow raises. The visual effect was stronger when the auditory cues were inconclusive for the required task because both words were accented. They suggested that because speakers normally do more with their pitch than with their eyebrows, listeners have learnt to pay more attention to the former. Indeed, all participants reported paying most attention to information in the auditory channel.

In a follow-up study, Krahmer et al. (2002a) investigated whether the greater impact of pitch accents on the perception of focus found above could be attributed to unnaturalness of their synthetic eyebrow raises, or to the possibility that these were simply not functional and were therefore ignored. Their materials were similar to the above, but this time they presented minimal pairs of the phrase "blue square", in Dutch, that were distinguished by the placement of an eyebrow raise either on the first or on the second word. That is, the only difference between the members of the pairs was that one member had an eyebrow raise on the first word, while the other one had a movement on the second word. Thus, some stimuli had a pitch accent and eyebrow raise on the same word, others on different words, and others had two pitch accents and only one eyebrow movement. Twenty-five participants were presented with twelve pairs each, and had to choose which animation they preferred, in terms of synchronisation between sound and image. Results showed a preference for pitch accents and eyebrow raises to be aligned on the same word, and for eyebrow raises to occur on the first word when both words were acoustically accented.

From the above, Krahmer et al. suggested that eyebrow movements may have the same function as pitch accents in making a word prominent. They tested this further by asking the same participants to rate the prominence of words in minimal pairs that differed on the presence or absence of an eyebrow raise on the accented word. Since unaccented words did not have brow raises, there were no inconsistent stimuli. Each subject was presented with eight pairs, four of which were distractors. Here they found that, interestingly, eyebrow raises not only boosted the perceived prominence of accented words, but they also scaled down the prominence of unaccented words next to them.

Eyebrow raises then seemed to be relevant for prominence perception, supporting earlier claims in the literature. However, as pointed out by the authors, they

were only minimally used in the previous study (Krahmer et al., 2002b) where participants relied mostly on auditory information. Krahmer et al. (2002a) concluded that it might be that eyebrow raising is used more consistently as a cue to different kinds of discourse information, or that listeners are biased to using auditory information, rather than the visual one, for the perception of focus. In relation to this, and discussing the synthetic stimuli used in their studies, Krahmer et al. (2002a) made a very important point by noting that "while the manipulations were inspired by claims in the literature, it would be nice to supplement the current results with findings of observations on real speakers to see whether they indeed use eyebrow movements for the determination of focus as suggested here, or whether these mainly signal other types of information, if any" (p. 1936). In the current thesis I addressed this point by examining eyebrow raises spontaneously produced by real speakers. Furthermore, these were observed as they occurred in natural dialogues, thus using a larger linguistic context than the short phrases above. Some of the issues investigated were whether brow raises are indeed aligned with pitch accents, and whether they are used to boost the prominence of new information in contrast with given information.

The temporal properties of eyebrow raises used in Krahmer et al. (2002a,b) were similar to those in House et al. (2001). Krahmer et al. stated that their choice of overall duration was based on the average duration (376 msec) of rapid eyebrow raises naturally produced by speakers in the study by Cavé et al. (1996). As they explained, they opted for slightly shorter duration (300 msec) because of the short length of their phrase stimuli, but there is no explanation for the division into equal rise, hold, and lowering portions of their brow movements. As I mentioned above, it could be that this is not the most appropriate structure. One of the disadvantages of using synthetic stimuli is that results may be biased by inaccurate representation of the natural behaviour.

Krahmer and Swerts (2004) discussed the advantages and disadvantages of the method employed in Krahmer et al. (2002a,b), which they called "analysis-by-synthesis". With this method, claims made in the literature are implemented in a talking head or embodied conversational agent, which is then tested to verify those claims. This is a powerful method that, as Krahmer and Swerts explained, offers direct control over relevant parameters, and allows the implementation and evaluation of different theories. However, they warn that it should be used

with caution. One possible disadvantage is that results may be incomplete when a more important parameter than the one manipulated is left out. Thus, they suggested a combination of the analysis-by-synthesis method with another one referred to as "analysis-by-observation". An example of the latter is given below in the description of a test they conducted to gain insight into which audio-visual cues to prominence human speakers actually use in Dutch.

In their analysis-by-observation test, Krahmer and Swerts (2004) asked twenty participants to pronounce nonsense words provided on printed cards. Each word consisted of three equal syllables with one syllable printed on capital letters: e.g. *"ga GA ga"*, *"MA ma ma"*. The participants were asked to emphasise that syllable when pronouncing the words, but they were not told what kind of cues they could use to do that. Looking at a camera, each participant pronounced twelve words in two different versions: neutral and exaggerated. Krahmer and Swerts reported that almost all the participants used verbal cues for emphasis, and many also used visual cues. Nine out of twenty raised their eyebrows, while four used head movements. Also, in the exaggerated condition the most obvious audiovisual cue for prominence was clearer articulation on the stressed syllable (used by 18 participants). Krahmer and Swerts related these findings to a production study by Keating et al. (2003) which showed a correlation between phrasal stress[9] and both head and eyebrow movements. In this study three native speakers of American English were recorded while reading 24 sentences in which the location of phrasal stress varied (e.g. *"So TOMMY gave Debby a song from Timmy"* versus *"So Tommy gave DEBBY a song from Timmy"*). To measure their facial movements, retroreflectors glued to the talker's face were tracked by a motion analysis system (see section 2.6 for more details on automatic measurement of facial movements). Keating et al. reported that talkers used eyebrow raising on almost all stressed words, and they also moved their heads more on stressed words. They also found differences in articulatory movements such as lip and chin displacement. Specifically, there were increased opening movements for the stressed syllables, which relates to the clearer articulation produced by speakers in the study by Krahmer and Swerts (2004)[10].

[9]The kind of prominence related to focus, as pointed out by Krahmer and Swerts (2004)

[10]Another study cited by Krahmer and Swerts (2004) in relation to clearer articulation on stressed syllables is the one by Erickson et al. (1998), who found an increase in jaw opening on emphasised words. See also Erickson (1998) and Erickson (2002)

Next, Krahmer and Swerts (2004) conducted a perception test to assess the relative contribution of the visual and auditory cues for prominence used by speakers in their "analysis-by-observation" study. The recorded utterances from five of the above participants were presented to three groups of fifteen subjects in three different conditions: audio only, visual only, and audio-visual. Subjects had to determine which was the emphasised syllable. In the audiovisual and audio only conditions they performed very well, as expected: 97.1% and 97.3% correct, respectively. With visual cues only they scored significantly less good, confirming their previous findings about the greater impact of auditory cues for the perception of prominence, as compared to visual cues. Nevertheless, in this visual only condition subjects were still surprisingly good at determining which syllable was stressed (overall 92.89% correct answers). These results suggested that there are clear visual cues for prominence, but it remains uncertain what these cues may be. As Krahmer and Swerts explained, a combination of this method, i.e. analysis-by-observation, with analysis-by-synthesis could provide more insight into possible cues used by speakers and how they are interpreted by perceivers.

Apart from the possible disadvantages of the analysis-by-synthesis method used by Krahmer et al. (2002a,b), it must be pointed out that their findings on Dutch may not generalise to the use of cues for prominence in English. Krahmer and Swerts (2004) reported different results for Italian, although they argued that these differences can be reduced to prosodic differences between Italian and Dutch. As a Germanic language, English is similar to Dutch in the use of prosodic features such as pitch accents to provide prominence. However, it may still be the case that speakers of the two languages differ in their use of visual cues. The current thesis investigates the use of eyebrow raises by speakers of English. The two studies described below also reported on English.

Massaro (2002) studied audiovisual cues to the perception of stress in English. Using an animated talking head and synthetic speech, he manipulated eyebrow raising, eye widening, amplitude, and F0 to investigate their relative contribution to perceived emphasis on the first or last word of noun-verb-noun sentences. Although all parameters influenced participants' judgements, amplitude was the most influential factor on the perception of stress. The relative contribution of eyebrow raising, compared to other factors, is not reported.

As we have seen, eyebrow movements have been mainly associated with the marking of prominence. But other roles have also been suggested in relation to utterance function. In particular, it has been suggested that they can signal a questioning function in the utterance they accompany, but there is not strong evidence for this. This is an important issue with relevance for the current thesis. On this topic, Srinivasan and Massaro (2003) investigated which auditory and visual (facial) characteristics could potentially differentiate statements from questions. They carried out three experiments, described below, using statements and echoic questions. An echoic question has the same word order as an equivalent statement, but it is interrogative in nature and differs from the statement in prosodic features such as fundamental frequency contour. In their first experiment, Srinivasan and Massaro presented twenty-two participants with four natural English utterances (from a recorded corpus) in statement and echoic question form and with stress placed on different words. The utterances were presented in three different modalities: audio only, visual (face) only, and audiovisual (white noise was added to avoid a ceiling effect and a lack of distinction observed in a pretest between the audio and audiovisual modalities). Participants had to identify the sentence as a statement or a question, paying attention to both the face and the voice of the speaker. Statement/question forms were distinguished across all modalities, except for one utterance where subjects did not distinguish them in the visual modality. Srinivasan and Massaro selected the most discriminable utterance pair (utterance No. 3: "We will weigh you") and examined its acoustic and visual characteristics. Acoustic differences were found in F0 contour and in duration and amplitude of the final syllable. As for visual cues, questions differed from statements in a significant eyebrow raise and head tilt across the length of the utterance. Details of these acoustic and visual measurements were used in the construction of synthetic versions of statement/question pairs for subsequent experiments, as explained below.

Using synthetic speech and a computer-animated talking head, versions of the four utterances above were created with the visual and auditory characteristics found in utterance pair No. 3. These were used as stimuli presented to sixteen participants with a similar procedure to the previous experiment. Results showed that first, statements and questions were significantly discriminated. Second, comparing the effect of the visual characteristics, the synthetic

characteristics were equal or better than the natural ones in terms of their effect on the discrimination of questions versus statements. This indicated that the visual cues of utterance No. 3 were effective and could be generalised to the other synthetic utterances.

In the third experiment, the synthetic version of utterance pair No. 3 was employed to construct a continuum from ideal statement to ideal question. Five levels were made changing pitch contour, amplitude, and duration, in the auditory channel, and eyebrow raise and head tilt, in the visual channel. For the bimodal modality, consistent and inconsistent stimuli were created by combining the same or different levels, respectively, of the auditory and visual dimension. The aim was to test how cues from both dimensions could be integrated in the perception of prosody. Forty-three participants were asked to identify the stimuli as either statement or question. Srinivasan and Massaro found that although participants made use of both auditory and visual information, visual cues had a weaker effect on their judgements. Two subsequent experiments failed at increasing the weak visual effect relative to the strong auditory effect. In experiment four, twenty-one participants were presented with enhanced (doubled in magnitude) visual characteristics in the synthetic stimuli. In experiment five, the enhanced visual cues were presented, to seventeen participants, with a more ambiguous auditory continuum that had been changed by attenuating the differences between the levels. But as mentioned, even with these manipulations the auditory cues remained more informative.

From Srinivasan and Massaro's results, it seems that eyebrow raising is related to questioning. However, a strong conclusion cannot be made for several reasons: the auditory information was much stronger than the visual one at conveying this function. Also, their results may be biased by the use of synthetic stimuli presented in isolation (not embedded in a larger linguistic context). Thus, it may be the case that eyebrow raising is not so clearly related to questioning as it was claimed in earlier observations. The current thesis addressed this issue by examining the use of eyebrow raises in different types of utterances, including queries, in real conversations (see Chapter 4).

To summarise, several findings in the literature suggest eyebrow raises may have

conversational functions but the results are inconclusive. Eyebrow raises seem to indicate aspects of dialogue structure by marking the start, continuation, and end of a topic (Chovil, 1989, 1991a), but there is no strong evidence for this function. They also seem to be aligned with some pitch accents (e.g. Cavé et al., 1996, 2002), but the details of this alignment are unknown. In relation to this, brow raises have been associated to prominence and specifically to the marking of information status (to mark contrastive information). However, this relationship is still unclear, in particular for English, and the supporting evidence is limited to short synthetic utterances (Krahmer et al., 2002a,b; House et al., 2001). This thesis investigates eyebrow raises as they occur spontaneously in a corpus of dialogues in English. Their relation to the dialogue structure and to pitch accents is analysed with a hypothetico-deductive approach in order to investigate possible discourse and prosodic functions.

2.6 Methods of measuring facial movements

One of the problems in studying spontaneous facial movements is how to obtain an accurate and objective measurement of these movements. This is probably one of the reasons why there is such a lack of substantial empirical findings in the study of conversational facial behaviour. This section describes some of the methods that have been used in previous research to study facial movements [11]. Details of the methodology employed in the current thesis will be provided in Chapter 3.

The early studies mentioned in 2.1 above (Bell, 1844; Duchenne de Boulogne, 1862; Darwin, 1872) did not have access to modern technology and relied largely on their own direct observations, or on the reports of others. Duchenne's description of facial muscle activity was the most systematic at the time. He applied electric current stimulation to the facial muscles of a man who had lost feeling on the face due to nerve damage. In this way he studied the effects on facial appearance of the individual facial muscles stimulated. Advances in technology now allow better, faster collection and analysis of data, as we will see below.

[11] For a more detailed description and comparison of different approaches, see Cohn and Ekman (to appear)

Wagner (1997) described two different types of method that are suited to different research questions on the study of facial behaviour: the judgment method and the measurement method. The judgment methodology, which has been widely used, relies on observers' judgments about facial movements in order to address questions about the information available in facial expression. There are two types of procedure: the category method and the rating method. In the first, judges are shown a series of stimuli and are usually asked to assign each to a category from a short list of responses provided by the experimenter. Less frequently, they are allowed to choose a response label freely, although they may be told what type of label is expected, for example "emotional" (e.g. Izard, 1971). In the second type, the rating procedure, judges rate the extent to which the stimulus faces show each of a number of properties, which are usually based on theoretical notions.

In contrast to the judgment method, the measurement method, as described by Wagner, addresses questions about the *structure* of facial expressions, not their interpretation by others. It involves describing or measuring facial movements, which can be done by automatic procedures such as electromyography (EMG), or by observational coding systems. Facial EMG, on the one hand, is the most objective: small electrodes on the skin detect muscle action and produce a signal from which muscular contraction is inferred. This can be done, for instance, while subjects view films that are considered to evoke emotions, for the study of expressions of emotion. EMG has the advantage that it can record activity that may not be observable, but it also has some serious disadvantages. First, it is invasive because it requires attaching electrodes to the skin. Second, it is intrusive: subjects can feel the electrodes on their skin and this not only draws attention to the topic of investigation but it can also change their normal facial behaviour. And third, it is relatively expensive, requiring specialised equipment and trained staff.

Observational coding systems, on the other hand, do not have such problems. Rosenberg (1997) described and compared several examples. The Facial Action Coding System (FACS) (Ekman and Friesen, 1978) has been the most widely used. FACS was developed to measure all visually discernible facial movements. It describes facial activity in terms of 46 "action units" (AUs) and some categories of head and eye positions. An AU does not correspond directly to an individual

muscle, since not all muscles produce different facial appearances, but to "an individually producible facial movement" (Wagner, 1997). Each AU has a numeric code and it can be coded in terms of intensity, on a 5-point scale, and timing. FACS is designed for human coders to score dynamic facial patterns from recordings by observing facial movements and assigning the relevant numeric codes to classify them into AUs. It appears to be both comprehensive and objective. And although it has mainly been used in the study of facial expressions of emotion, it allows the scoring of all kinds of facial movements with different goals (for a compilation of studies using FACS see Ekman and Rosenberg, 1997). FACS has, however, the great disadvantage of being very labour intensive, both because of the training it requires and because of the time actually spent on scoring. This will discourage the use of large corpora where an audiovisual analysis already requires a considerable amount of measurement of auditory data.

There are other scoring systems that are less time-consuming, such as the Facial Affect Scoring Technique (FAST) (Ekman et al., 1971), and MAX, the maximally discriminative facial movement coding system (Izard, 1979). These have some disadvantages. They are based on theoretical assumptions by which scorers look for certain facial configurations believed to be associated with certain emotions. Thus, they are limited to the actions that were considered relevant when constructing the system. Also, they can be subjective. Finally, they do not measure intensity of movement.

In recent years, using computer vision techniques, there have been advances towards the automatic analysis of facial movements. That is, computer systems have been developed that attempt to automatically measure facial movements from recorded images and to recognise patterns in those movements. Some of these, especially in early research in this area, made use of markers on the face to facilitate the extraction of facial features. Reflective markers can be attached to the skin and these can be tracked with computer motion techniques on the recorded images of that face. This is what was done for instance by Cavé et al. (1996, 2002) and Keating et al. (2003) above. But this technique has some of the disadvantages that facial EMG has, since the placing of markers may modify the subjects' normal behaviour and it is also rather expensive.

As Cohn and Ekman (to appear) explain, most current research in automatic facial image analysis requires no markers or other enhancement of facial features. For comprehensive reviews see Fasel and Luettin (2003), Pantic and Rothkrantz (2000), and Tian et al. (2003). Automatic facial analysis requires measuring the facial movements and recognising meaningful patterns. Tian et al. described three necessary steps in this automatic analysis: first, the face needs to be located in the input images. Then, information about facial changes need to be extracted and represented. Finally, those changes need to be recognised as a particular facial action. For this automatic recognition, most systems have used FACS to classify the facial activity into action units.

Examples of progress in the automatic recognition of spontaneous facial movements are described in, for instance, Cohn et al. (2003), where an initial test on images from ten subjects (one minute each) was successful at recognising Action Unit 45 (blink) with 98% accuracy (measured as agreement with manual FACS coding). With the same database and in a study more related to this thesis, Cohn et al. (2004) achieved 76% agreement in the automatic recognition of brow action units (brow raise, brow lower, and no brow action).

But there are still many challenges in the automatic analysis of facial movements, especially for *spontaneous* movements as opposed to posed ones. The analysis is hampered by individual differences between faces, such as those due to age, gender, or ethnicity differences. Other problems are changes in head orientation, distractors such as beards and glasses, and partial occlusion of the face, for instance, by a hand placed on part of it. Another serious challenge is the temporal segmentation of the facial actions, that is, the identification of start and end, which is particularly difficult since the transition from one pattern to another can be made without a neutral state. Finally, for the training of a fully working system there is still a serious need for larger image databases of spontaneous facial movements.

Automatic recognition would make research on facial movements a great deal faster by reducing almost completely the time spent on human coding. It would enable a detailed analysis, with an objective, non-intrusive, standardised measurement. However, although progress in the last few years is encouraging, this

area of research is still at an early stage and cannot yet be used reliably in the study of facial movements in spontaneous real conversations.

2.7 Embodied Conversational Agents

In this section I will briefly describe some aspects of Embodied Conversational Agents (ECAs) as an area that can benefit from research on human facial behaviour, such as eyebrow raising. First I will explain what ECAs are and I will give examples of some potential applications. Then I will discuss the challenge that these systems face in trying to reproduce human behaviour, a point to which I will return in the final chapter of this thesis.

2.7.1 What are ECAs?

ECAs have been defined as "more or less autonomous and intelligent software entities with an embodiment used to communicate with the user" (Ruttkay and Pelachaud, 2004, p. xv). Several other terms have been used, such as *talking head, avatar, virtual human*, or *humanoid*. They can be embodied in just an animated head, like the talking heads used in the methodology of the studies in section 2.5.3 above, or they can be represented in a full torso or body. They can be 2D or 3D. And they can have extremely human-like physical features, or look more like a cartoon, not necessarily depicting a human.

These "embodiments" have become an important part of modern multimodal communication systems which aim at providing a computer interface in which an agent communicates with the user as another human interlocutor would. This is of course a very ambitious goal which these systems are far from achieving yet. But the field of ECAs is attracting more and more research from different disciplines such as Computer Science, Psychology, and Linguistics, and some progress has been made.

ECAs can have a wide range of applications. They can be used as educational software to promote learning. For instance, they have been used in research as pedagogical agents tutoring subjects such as Newtonian physics and computer literacy (Graesser et al., 2004), marine biology (Darves and Oviatt, 2004), botanical anatomy and physiology (Lester et al., 1999a), internet packet routing (Lester

et al., 1999b) and naval operating procedures (Rickel and Johnson, 2000). They can also help in language training for the hearing impaired (Massaro and Light, 2004). ECAs can also have sales applications. For instance as a real estate agent who shows virtual properties to users (Cassell et al., 2000a) or provides general information about apartments for sale (Gustafson et al., 2000), or as a salesperson helping clients redesign their rooms (Foster, to appear). They can also be used as assistants in eBanking systems (Morton et al., 2004). In other public services, such as information kiosks, they can assist users by giving them directions to a certain location (Kopp et al., 2004, to appear).

With all the potential applications that ECAs have it is not surprising that they have rapidly become an important part of human-computer interaction research. But the requirements of a fully developed ECA system are still beyond current implementation capabilities as we will briefly illustrate below.

2.7.2 *The challenge of designing efficient ECAs*

It is not necessary to go into technical engineering details of ECA design to realise the many challenges that designers face. An obvious one is that in order to create an animated character that shows and interprets real human conversational behaviour we must have a very good understanding and model of this behaviour. Cassell et al. (2000b) edited a good collection of studies demonstrating the breadth of models and behaviours that are necessary to natural conversation. As Cassell et al. explain in the introduction to their book, those studies addressed four models in particular: emotion, personality, performatives, and conversational function. These are proposed by different researchers as explanations for the range of verbal and nonverbal behaviour in face-to-face conversation, and therefore as a way to realize conversational surface behaviours in a principled way in the design of ECA. But the models we have are still very rudimentary to allow us to generate complete human-like behaviour.

Apart from verbal behaviour, which poses a challenge in itself, there is a wide range of non-verbal behaviour that is part of face to face conversation. Different body movements normally accompany the speech: hand gestures of various kinds, body posture shifts, head movements, eyebrow movements, ... A multimodal dialogue system needs to decide whether the ECA should show body

movement at a certain time in the conversation, then it needs to select which particular movement or combination of movements is necessary, and it must integrate this temporally with the speech in an appropriate way. And the task becomes even more complicated when the system also needs to interpret the multimodal input that the user provides in the form of speech, body movements, writing, etc.

The previous sections in this chapter have demonstrated that, although facial movements such as eyebrow raises seem to be connected to linguistic phenomena in some way, we still do not know when and how exactly these are used in conversation. This knowledge is necessary for both the input and output of a fully developed ECA. The following chapters describe an empirical study investigating whether eyebrow raises are indeed connected to the speech they accompany, and if so, in which ways. One of the goals is to inform the design of multimodal dialogue systems where an ECA can make use of facial movements as part of their input/output.

CHAPTER 3

Corpus collection and annotation

In this chapter I will describe the methodology involved in the data collection and annotation of the current study. There is no standard method for the study of facial movements in relation to speech and, in fact, methodology seems to have been a weak point in a lot of the research on facial movements in communication. One of the contributions of this thesis is the presentation of a methodology for the collection and annotation of audiovisual data that can successfully produce a rich corpus for the study of eyebrow movements in dialogue. This method can be expanded to include the study of other facial movements and head movements as well.

The structure of this chapter is as follows: first there will be a description of the experimental setup used in the corpus collection. Then, in section 3.2, I will present the method used to record the data. And in 3.3 I will describe the method employed to annotate the data in the auditory channel (discourse structure, information structure, pitch accents) and in the visual channel (eyebrow raises). The final section includes some images illustrating examples of eyebrow raises produced by the participants in this study.

3.1 The Map Task

The data in this thesis was recorded using a variant of the Map Task (Brown et al., 1983; Anderson et al., 1991). In the Map Task, two participants, the *Instruction Giver* (*IG*) and the *Instruction Follower* (*IF*), sit opposite each other with slightly

different versions of a simple map. The *IG*'s map has a route navigating a set of labeled landmarks, whereas the *IF*'s has only landmarks. Their task is to draw the *IG*'s route on the *IF*'s map. But their sets of landmarks are not quite identical and they cannot see each other's maps, so both participants must collaborate to perform the task. Thus, they engage in conversation so that the *IG* can describe the route to the *IF*, who in turn can ask any questions or clarifications needed in order to draw that route. The fact that there are discrepancies between the two maps in terms of the landmarks that appear on them means that both participants must engage in efficient communication in order to achieve their goal. At the same time it concentrates their attention on fulfilling the goal, and does not make them self-conscious about what they say or how they should say it, or about the presence of recording equipment around them. The Map Task has already proved to be a good way of eliciting spontaneous dialogue while constraining the content and goal of the conversation. Knowing the task and goal makes it simpler for the analyst to discern the purpose of individual utterances in the dialogues. In addition, there is already a standard and reliable coding scheme that can be used to describe the structure of these dialogues and can also be applied to other domains. This scheme, Conversational Games Analysis (Carletta et al., 1997), will be described in section 3.3.1. The Map Task was therefore considered as the best choice for a means of collecting the kind of audiovisual data that was needed for the current study.

It may be argued that using this type of task is not a natural way of eliciting dialogue. An alternative way of collecting the data would be to leave participants to simply talk about an assigned topic or a topic of their choice. However, this will not necessarily produce more naturalistic conversation. In these situations, participants will be more aware of their conversation as an object of investigation. They are bound to feel observed and self-conscious about their communication behaviour, and this may change their natural patterns. By giving them a task, the participants' attention will shift to performing the task and dealing with the problems that come up. Also, when participants are free to talk about a topic, even if it is an assigned topic, we cannot predict what kind of interaction they will have and what they will say. In task-oriented dialogues, such as Map Task dialogues, the type of conversational exchange the participants will have is controlled. At the same time, the experimenter will know what the specific goals

are in the different segments of their interaction, and this will facilitate greatly the annotation of the dialogue in terms of communicative goals and structure. Also, task-oriented dialogues can be compared across interactions because they all belong to the same genre and they all share the same cognitive task.

While it is true that the Map Task can be thought of as a game more than a real task encountered in daily life, the type of communication exchange in which the participants engage is a common one. Like in the Map Task, people often engage in exchanges where there is asymmetrical knowledge between the two parties. That is, the information each of the interlocutors has is not the same and they must interact to exchange this information in order to perform a joint task. Also, the different purposes of individual utterances produced when doing the Map Task are also found in human-computer interactions in multimodal dialogue systems: requesting information, providing information, giving instructions, etc. For instance, in navigation systems interactions involve asking for directions, giving directions, expressing problems in understanding, requesting and providing clarifications, etc. These are all spontaneously produced in Map Task dialogues, providing a rich corpus of data for the study of audiovisual communication behaviour.

Typically, with the exception of studies involving gaze behaviour, research on dialogues collected with the Map Task has concentrated on the verbal channel. The current thesis is the first study to investigate eyebrow raises in Map Task dialogues. In the next two sections I will describe how the data was recorded and annotated. I will explain different choices that were considered and the reasons for following a particular approach, and I will discuss some problems that came up in the process.

3.2 Method: Data recording

3.2.1 Participants

Four female participants were recorded, aged between 22 and 26. They were two pairs of friends (*A1*, *A2*, and *B1*, *B2*), each pair unacquainted with the other. They were selected, without their awareness, to have clear visible eyebrows. One of the selected participants had to drop out before the recording. Her substitute had

much lighter and less visible eyebrows than the rest. But she had a clear forehead where wrinkles formed when she lifted her eyebrows, which aided identification of eyebrow raising. All participants came from England (Manchester, Liverpool, and South of England). At the time of the experiment they were students of the University of Edinburgh. They were told that the aim of the experiment was to study linguistic phenomena and that they would be filmed to obtain a complete record of the data. But the specific purpose of the study was not revealed to them before the recordings and there was no mention of facial movements as part of the interest in the investigation. Each participant was paid £15 at the end for their participation.

3.2.2 *Materials*

Four pairs of maps (composed of an *IG* and an *IF* map), plus a single *IG* map were created as materials for the task. Maps from the HCRC Map Task corpus (Anderson et al., 1991) were used as a blueprint to draw the map route and the distribution of landmarks around it. New landmark names and their drawings, however, were produced to contain as many sonorants as possible in the names. Although eventually not needed, this was done to facilitate F0 tracking in the speech signal when annotating pitch accents (sonorants, as opposed to obstruents, show a clear and stable display of F0).

All maps, reproduced on A3 paper, were intended to represent one of the following scenes: a zoo, a garden centre, a desert, a sea port, and a museum. In each map pair the *IG's* map had a line route, and the *IF's* did not. The starting point was drawn on both maps, but the end point appeared only on the *IG*'s map. There were about 13 landmarks on each map, distributed around the route (or where the route should be in the case of the *IF's* map). Some landmarks in the *IG*'s map did not appear on the *IF*'s, and vice versa, and some were mismatched. A rescaled copy of each map is included in Appendix A[1].

[1]The reader may notice that one of the landmark names in each map pair does not really fit the scene represented in the map. This was intentional, following the design of the original maps in the HCRC Map Task corpus

3.2.3 *Design and Procedure*

The design of the current experiment differed from the standard procedure used for the HCRC Map Task corpus collection (Anderson et al., 1991) in the addition of two conditions where each participant was recorded alone. These conditions are labeled *a)* and *b)* below. The motivation for including these here was that in the study there was originally a plan to compare speakers' facial behaviour in monologues and dialogues. However, due to time limitations the analysis of the monologue data was not done. Although these two monologue conditions were not included in the current analysis, the whole design is presented below to show the context in which the dialogues were recorded and to describe the available corpus that may be used for future studies.

a) Monologue rehearsal

Each participant, one at a time, was asked to rehearse giving instructions on a route of a single test map (the museum map). She was asked to describe the route aloud in front of a camera as if she was actually giving instructions to another participant. She was told that this would actually not be recorded but was necessary for the cameraman to adjust the settings on his camera and for herself to get used to the task. But in truth the whole session was recorded.

b) Monologue recording

In another session each participant was asked to give instructions as in the rehearsal, but this time for a recording. She had to imagine that she was describing the route for another person who would later have to draw this route watching the video of her instructions.

c) Dialogue recording

This session was similar to the standard Map Task design described above, and is the only one that was analysed in this thesis. Participants were recorded in pairs sitting opposite each other and collaborating to reproduce the *IG*'s route on the *IF*'s map. There were a total of eight dialogues in which four different map pairs were used. Each participant served as *IG* for the same map to two different *IF*s, and as *IF* for two different maps.

d) List reading

Following the standard Map Task design, the last recording of each participant was made while she read out a list that she was given with all the landmark names that appeared in the maps. It was suggested that they could look at the camera while reading each name, and indeed they chose to do this in most cases. This provided an audiovisual record of landmark names enunciation in list form.

There were 20 recordings in total. The order in which they were made is presented in Table 4.1. The first column shows the recording session number. Then there are four columns, one for each speaker. When a cell in a column is filled it means the speaker participated in that recording session, and the type of recording is specified with letters as *'a)'* (monologue rehearsal), *'b)'* (monologue recording), *'IG'* (*Instruction Giver* role in a dialogue), *'IF'* (*Instruction Follower* role in a dialogue), and *'d)'* (list reading). The map used for each session is also specified referring to the scene it represented.

The recordings were made and edited at the Media and Learning Technology Service (MALTS) of the University of Edinburgh. In the dialogue session, participants sat one in front of the other, approximately two metres apart, across two joined tables. On the table they each had a board where the A3 size map was placed. The board was slightly larger than A3, and was raised and angled so that they could not see the other participant's map but had a full view of their face. The angle also meant they did not need to lower their head completely when looking at their own map. The same tables and boards were used when participants sat alone in the rehearsal and monologue sessions described above.

A camera positioned high across one of the tables recorded the *IF*'s map all through the dialogue session. This was done to record her drawing activity through the dialogue, in case it was needed for future analyses of the data. Two cameramen, behind and slightly to the side of one participant each, recorded the opposite participant with a professional camera on a tripod. The cameramen stood still, in a shaded area at least one metre away from the participants, and monitored the cameras so that the participants did not move off the viewpoint. The participants were recorded from their mid torso up and from a front position. In the recording of the monologues and the list reading (in *a) b)* and

Session	*Speaker A1*	*Speaker A2*	*Speaker B1*	*Speaker B2*
1–4	*a)* museum	*a)* museum	*a)* museum	*a)* museum
5	*b)* desert			
6	*IG* desert		*IF* desert	
7				*b)* garden
8		*IF* garden		*IG* garden
9	*IF* zoo	*IG* zoo		
10			*IG* sea	*IF* sea
11		*IG* zoo		*IF* zoo
12	*IF* sea		*IG* sea	
13	*IG* desert	*IF* desert		
14		*b)* zoo		
15			*IF* garden	*IG* garden
16			*b)* sea	
17–20	*d)*	*d)*	*d)*	*d)*

Table 3.1: Order of Map Task recordings

d), above) one single cameraman recorded the participant, from the front, in the same way.

Having cameramen in the recording room could potentially make the participants nervous. However, this was necessary to obtain a good constant closeup of their faces. The participants' heads move considerably during the dialogue, and so if they had been recorded with fixed mounted cameras they would have been out of viewpoint in many occasions. The cameramen were able to monitor and adjust their cameras with small movements to maintain the closeup view of the participants' face. The fact that the participants' tables were illuminated with bright lights while the cameramen's area was dark, made the latter rather inconspicuous. The participants seemed to get used to this environment very

quickly, and because they were concentrating on doing their task they did not appear disturbed by the presence of cameras and cameramen.

In between recordings participants waited outside the recording studio in a room where refreshments were provided. The experimenter came into the recording studio before each session to instruct the participants, and then left and waited in the surveillance van with the staff that monitored the whole recording. Total recording time was around an hour.

The instructions to the participants were provided on paper and also briefly summarised by the experimenter, who clarified any questions they had. The written instructions are presented in Appendix B. Participants had no restriction on what they could say and when. But they were instructed not to gesture with their hands to show the shape of the route to their interlocutor. This was done for several reasons. If they could indicate the shape of the route by drawing it in the air with their hands, they would be able to perform the task faster and would make the dialogues shorter and less rich in terms of references to landmarks. They might also hide part of their face by waving their hands in front of them, which could make it harder for the experimenter to see their eyebrows. They were told they did not need to keep their arms still or restrict their movement. Simply, to make their task more challenging, they should not use gesture to show the map route to the other participant. When asked at the end if they had any comments about the task or about how they had felt, none of them reported feeling uncomfortable about this restriction. Indeed, inspection of the recordings shows them moving naturally. Only once one of them apologised for using a gesture in her description of a section of the route. Still, it could be argued that this restriction made the interaction abnormal in a way, since in face to face conversation we normally use our hands to describe spatial information.

All the recordings were edited by staff at the MALTS editing studio. The sessions with two participants were edited to show them in a split screen with an inset at the bottom showing the recording of the *IF*'s map. This format was recorded on VHS tape. Recordings were also edited to be burnt on CD-ROMs. Here the image was cropped to show only the head and shoulders of the participants, and no map inset was included. The recordings on the CD-ROMs were used for the

analyses described in this thesis and are available on request for research purposes[2]. Figures 3.1 and 3.2 show one video frame each from two of the recorded dialogues, with the *IG* on the left of the image and the *IF* on the right[3].

Figure 3.1: Example video frame from a dialogue recording with speakers *A1* and *B1*

Figure 3.2: Example video frame from a dialogue recording with speakers *B2* and *A2*

[2]Please contact the author at *marisa@ling.ed.ac.uk*, or through *Linguistics and English Language* at The University of Edinburgh

[3]In Figure 3.2 part of the face of the *IF* participant has been pixelated to protect her anonymity

3.3 Method: Data annotation

All the recordings were orthographically transcribed and then annotated by the author in terms of dialogue structure, pitch accents, information structure, and eyebrow raising. Below I will describe the annotation procedure and related issues for each of these.

3.3.1 Dialogue Structure: Conversational Games Analysis

As was already explained in the previous chapters, a dialogue is composed of a series of utterances which are not randomly produced, they are uttered with an intention and can be linked to other utterances and grouped into segments with a coherent communicative purpose. These segments, in turn, combine into larger groups, and in this way the structure of a dialogue develops. A short example was provided in Chapter 1 (section 1.1). Then, Chapter 2 (section 2.3.1) introduced the dialogue structure scheme known as Conversational Games Analysis (Carletta et al., 1997), which divides a dialogue according to the speaker's purpose in producing the utterances. The structure of the eight dialogues recorded here was annotated according to this scheme, which is described in detail below. But first, I will set out the reasons why this scheme was adopted.

The Conversational Games Analysis was considered as the most appropriate scheme to describe the structure of the dialogues in this investigation. Based on earlier work by Power (1974, 1979) and Houghton (Houghton, 1986; Houghton and Isard, 1987) (see section 2.3.1), Conversational Games Analysis was originally developed for use on the HCRC Map Task corpus (Anderson et al., 1991). So first of all, it was developed to represent the structure of dialogue, and more in particular, of task-oriented dialogues, which made it very suitable for the data collected here. The scheme was specially designed for Map Task dialogues, and so it defines the major communicative acts that occur in this type of task. Thus, it makes appropriate distinctions for the utterances in the dialogues collected in the current study, and for the object of this investigation. At the same time, these distinctions are not too specific and so they can be adapted to be used with other tasks in different domains. In fact, the scheme was intended to represent dialogue structure in general so that it could be used with codings of other dialogue phenomena.

The basic idea of this scheme that makes it appropriate for the current investigation is that it segments, classifies, and links utterances according to the speaker's intention and completion of goals during the discourse. There are other schemes that are developed around the same idea, and that also describe different levels of dialogue structure. For instance, as was mentioned in the previous chapter (section 2.4.3), Nakatani et al. (1995b) described a scheme based on the theory of discourse structure by Grosz and Sidner (1986) in which the segmentation of the discourse was also based on the utterance purpose (see also Hirschberg and Grosz, 1992; Grosz and Hirschberg, 1992). As in Conversational Games Analysis, in this scheme utterances are linked by their purpose and they form larger segments with a coherent communicative purpose. However, this scheme would not be appropriate for the dialogues collected for the current thesis, because it applies mainly to narrative discourse, and it has been used for the description of monologues. On the other hand, Conversational Games Analysis has been used as the standard scheme for Map Task dialogues, and it has been proved that it can be used reliably by different annotators (Carletta et al., 1997). I will refer again to the issue of reliability testing at the end of this section, when describing the annotation procedure for the current analysis.

The Conversational Games Analysis scheme distinguishes three levels of dialogue structure which are, from bottom to top, *conversational moves*, *conversational games*, and *transactions*. These are described below.

Conversational moves

The Conversational Games Analysis is based on the basic idea that human interactions are normally exchanges of initiating utterances, which signal some kind of dialogue purpose of the speaker, and response utterances to those initiations. Thus, a conversational move, the lowest level of the dialogue structure, is an utterance or part of an utterance that communicates an intention and can be classified according to its purpose in the communicative task and according to its form. A move does not necessarily correspond to a complete sentence or to a whole conversational turn. The concept of move can be better explained by describing its possible different functions. There are twelve types of move in this coding scheme: six initiation moves (*Instruct*, *Explain*, *Query-yes/no*, *Query-w*,

Check, and *Align*), five response moves (*Acknowledge*, *Clarify*, *Reply-yes*, *Reply-no*, and *Reply-w*), and one preparation move (*Ready*). The distinctions used to classify moves into these categories are summarised in Figure 3.3 (taken from Carletta et al., 1997) and further described below. As it will be explained later, these categories were reduced to a smaller set of broader categories for the analysis in the current study.

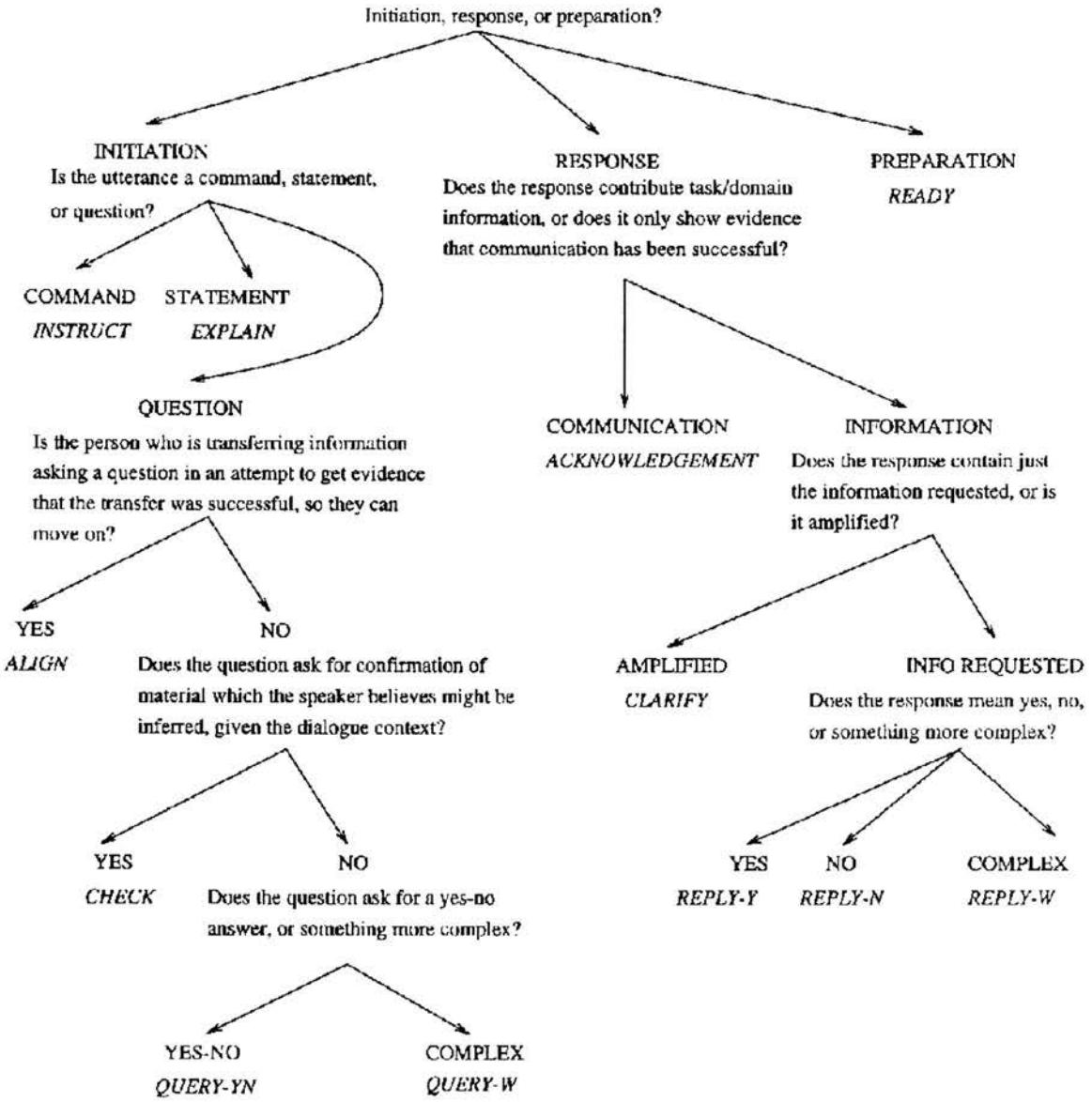

Figure 3.3: Conversational move types, after Carletta et al. (1997, p. 15)

***Instruct* move.** This is a command for the interlocutor to carry out an action. Very often it is an indirect command, but the implicit requested action is clear. Normally, *Instruct* moves are produced by the *IG* to tell the *IF* how to navigate part of the route, but they can also be some other kind of instruction, such as telling the interlocutor to repeat something they said. Occasionally, the *IF* can make an *Instruct* move, such as telling the *IG* to slow down, or to wait before giving the next instruction. Some examples would be:

- *IG: So then you go along just a few dashes*
- *IG: And when you are at the top of the camel man you want to go a direct ninety degrees to the right*
- *IF: Hold on*

Instruct moves carry the most important information for completing the task. A series of *Instruct* moves could in theory be enough for the participants to complete the task successfully. But in practice, of course, this is very rare, partly due to the difficulties encountered by the mismatches between the two maps.

***Explain* move.** This provides information that has not been elicited by the interlocutor (if it were, then it would be a response move). It may describe, for instance, some aspect of her map without intending that the interlocutor take any action on it. Examples:

- *IG: I don't have a picture of a weeping willow on my map*
- *IF: I'll just go straight down*
- *IF: I'm under the garden centre*

***Check* move.** A request to confirm information that the speaker believes but is not entirely sure about. It is usually about something that the interlocutor has said. When an *IG* is describing for the second time a map that she already described to a different *IF*, she can also use a *Check* to seek confirmation about something in the map that the current *IF* has not mentioned yet. Examples:

- *IG: Have you got, you've got the blooming lilac, haven't you?*
- *IG: You've got a flower alley, I think*
- *IF: so I'm going sort of straight down from there?*

***Align* move.** This checks for the interlocutor's attention, understanding, or readiness for the next move. Normally, the purpose of this move is for the *IG* to

confirm that the information has been transferred to the *IF* successfully, and that they can move on. *Align* moves are often short and are sometimes made even when the information transfer has already been acknowledged clearly by the interlocutor. Examples:

- *IG: Ok?*
- *IG: Yeah?*
- *IG: Yeah, have you got that?*

***Query-yes/no* move.** A question that requests a yes/no answer and is not a *Check* or an *Align* move. It often asks about what the partner has on her map. Examples:

- *IG: Now, have you got camellias?*
- *IG: Can you see the rare llamas at the top there?*
- *IF: so am I fairly close to the bottom of the page?*

***Query-w* move.** Any query not covered by the other move types, normally a wh-question but not necessarily, as in a question asking to choose an alternative in a set (except between 'yes' and 'no'). Examples:

- *IG: how many monoliths do you have?*
- *IF: so how far past the blooming lilac should I start heading left?*
- *IF: do I need to go to the left or to the right of the anemone?*

***Reply-yes* move.** Any reply to any query with a yes-no meaning, however it is expressed. Examples that could follow the *query-yes/no* moves above:

- *IF: Yes I have*
- *IF: Yup*
- *IG: Uh huh*

***Reply-no* move.** Similar to the *Reply-yes* above, but meaning 'no'. Example answers that could reply to the same *Query-yes/no* moves above:

- *IF: Er no*
- *IF: I don't have those*
- *IG: No, not really*

***Reply-w* move.** Any reply to any type of query which does not mean 'yes' or 'no', and is not a *clarify* (see below). Examples that followed the *Query-w* above:

- *IF: I've just got one*
- *IG: Just sort of ... like you're swinging round it*
- *IG: Oh, you're not you're not going quite as far as the anemone*

***Acknowledge* move.** A verbal indication that the speaker has heard, and normally has understood and accepted, the interlocutor's move to which it responds. It is typically a very short move that announces the speaker is ready for the next move. Examples, that could follow, for instance, the *Instruct* or *Explain* moves above:

- *IF: Ok, yeah*
- *IG: Right*
- *IF: Uh huh*

***Clarify* move.** This is described by Carletta et al. (1997) as a reply to any query in which the speaker tells the interlocutor something over and above what was strictly asked. When the information is substantial it is labeled instead as a reply followed by an *Explain* move. Carletta et al. explain that this is used, for instance, by the *IG* when the *IF* seems unsure of what to do. But that description seems to refer to a distinction which is too subtle and rather difficult to identify consistently. *Clarify* moves were in fact suggested as problematic in the reliability test of move segmentation and classification by Carletta et al. Therefore, this type was not included in the current study, in which moves with the purpose described above were classified as *Reply* types or *Explain*. An example *Clarify* move in the description by Carletta et al. is the *IG*'s utterance in the following consecutive moves:

[...instructions which keep them on land]

- *IF: So I'm going over the bay?*
- *IG: Mm, no, you're still on land*

***Ready* move.** This is another problematic type that was not included in the current data annotation. It is normally a short utterance such as *ok* or *right* that often occurs at the beginning of a new conversational game (see below). Carletta et al. (1997) pointed out that it is debatable whether this should be classed as a distinct move type or it should be treated as a discourse marker included in the following move. This was the source of some disagreement in the reliability test of move segmentation. Two example *Ready* moves (underlined) from Carletta et al. are:

- *IG: <u>Now</u> I have banana tree instead*
- *IG: <u>Ok</u>. Now go straight down*

An extra move type was coded in the current thesis for utterances that could not be classified into any of the types above:

***Unclassifiable* move.** An utterance (normally short) that did not fit into any of the categories above, for instance if a speaker suddenly talked about something unrelated to the task. An example (when one of the speakers apologised for gesturing with her hand) was:

- *IG: Sorry about the gesture there*

Moves cannot be embedded within other moves, but the *IG*'s and *IF*'s moves can overlap. Thus, sometimes moves are interrupted, and they can also be abandoned. In this thesis, interrupted or abandoned moves were labeled as belonging to one of the original types above when the purpose of the speaker was clear. When the purpose was not clear they were marked as *Unclassifiable*.

Utterances in the current data were segmented and classified into one of the categories described above, except *Clarify* and *Ready*, which were excluded for the stated reasons. Later, however, these categories were reduced to a smaller set of broader move types for the analysis. This was partly to produce categories with sizable representation and also to simplify the distinctions and make them more reliable. The resulting six categories were: ***Instruct***, ***Explain***, ***Query*** (grouping *Query-y/n*, *Query-w*, *Check*, and *Align*), ***Reply*** (grouping *Reply-y*, *Reply-n*, and *Reply-w*), ***Acknowledge***, and ***Unclassifiable***. The reduced set preserved the basic distinctions between moves' purposes: to make the listener follow an instruction, to acquire some information, to provide some information, and to acknowledge receipt of information.

Conversational games

The next level up from conversational moves consists of conversational games. These are sets of moves starting with an initiation move plus the subsequent moves that are produced until the purpose of that first move is fulfilled or abandoned. A full game could be just two moves, for instance a *Query-w* followed by a *Reply-w*, or it could include a longer series of moves starting with an *Instruct* which is followed by other moves pursuing the goal of the *Instruct* until the action required is carried out or the goal is abandoned. Games always begin with

an initiation move, but not all initiation moves start a game. All moves must be included in at least one game. Games inherit the type of their initiating move, and they can be embedded within other games when they serve the purpose of the larger game.

Transactions

At the top level of the dialogue structure, a transaction is a set of games that negotiate a section of the map route and so correspond to one step of the task which the dialogue furthers. These sections map onto a speaker's own division of the route into segments that are dealt with one by one. Carletta et al. (1997) describe a typical transaction as a subdialogue that gets the *IF* to draw one route segment on the map. All games are included in some transaction and the latter cannot be nested, thus a transaction starts where the previous one ended. There are four types of transactions in the coding scheme: *normal* (the default), *review* (when the speaker returns to a segment of the route already discussed), *overview* (when the speaker overviews an upcoming segment of the route), and *irrelevant* (when a speaker discusses something which is not relevant to the task). The vast majority of the transactions are *normal*, and so transaction type was not used in the current analysis. A transaction is almost always started by the *IG*, but very occasionally the *IF* can also start one.

The segmentation of dialogues into transactions by the coder can be subjective. It sometimes helps to rely on discourse markers and phrases indicating the completion of a section and the shift into a new one, e.g. *'Ok, and once you've done that now you have to ...'*. Other times the change into a new transaction is clear, when for instance the speaker asks about a new landmark not mentioned before and starts the discussion of the route around it.

A brief example of move, game, and transaction structure is provided here:

TRANSACTION 1

Start of *Instruct* game
***Instruct* move** *IG*: 'ok, so then you go along just a few dashes
***Instruct* move** *IG*: 'and then head upwards'
End of *Instruct* game

Start of *Query-y/n* game
***Query-y/n* move** *IG*: 'do you have a camel man?'
***Reply-y* move** *IF*: 'I do'
End of *Query-y/n* game

Start of *Instruct* game
***Instruct* move** *IG*: 'Head upwards towards the camel man'

> Start of embedded *Query-y/n* game
> ***Query-y/n* move** *IF*: 'on his left hand side?'
> ***Reply-y* move** *IG*: 'yeah, keep him on your right hand side'
> ***Acknowledge* move** *IF*: 'yeah'
> End of embedded *Query-y/n* game

End of *Instruct* game

TRANSACTION 2

Start of *Instruct* game
***Instruct* move** *IG*: 'And then when you are at the top of the camel man you want to go almost a direct ninety degrees to the right'

(...)

Annotation procedure

Dialogue structure was annotated by the author using the *xlabel* software on *Entropic/Xwaves* which segments the digitised speech signal into labeled units. Moves were segmented and labeled as one of the move types described earlier in this section (except for *Clarify* and *Ready*). These types were later reduced, as explained, to a smaller set of six broader categories: *Instruct, Explain, Query, Reply, Acknowledge* and *Unclassifiable*. Games were segmented aligning their start and end with the start of the first move and the end of the last move in the game, respectively. And they were labeled with the inherited type from their first move. Transactions were segmented also in alignment with the start of their first game (and move) and labeled as one of four types described above. In the

current thesis only *IG*'s utterances were analysed. But because game and transaction structure relies on both speakers, it was necessary to have a full record of dialogue structure that included the *IG*'s and the *IF*'s utterances.

Dialogue structure annotation was done by listening to the recordings. In principle, it could be done on text by reading the dialogue transcripts. However, for two main reasons access to the audio recording was considered essential in this procedure. First, it has been proved that there is information in the speech signal that is important for marking the discourse structure. For instance, Nakatani et al. (1995a) reported more consensus among labelers when they listened to the speech, as they segmented the discourse, than when they used text alone. From this they concluded that aspects of the speech signal can help disambiguate among alternate segmentations, and therefore, the availability of speech has a critical influence on the outcome of discourse structure analysis. Second, in the current thesis, by doing the segmentation and annotation on a digitised record of the audio signal, and using a tool like the one mentioned above, the timestamp of the segment boundaries was electronically recorded and labeled at the same time. This was important for an accurate analysis of the data. The annotation was done, however, without access to the video recordings to avoid a possible bias from facial movements on the segmentation and labeling of the dialogue structure. To ensure correct and consistent segmentation and labeling, several passes of annotation were done by the coder for each dialogue.

It would of course have been better to have a second labeler or more, at least for part of the data, to be able to do a reliability test on coders' agreement on the annotation of discourse structure. Due to limited time and resources in this project this was not an option. However, previous work has already shown that Conversational Games Analysis can be used reliably (Carletta et al., 1997) by different coders. Carletta et al. tested the level of agreement on four coders' annotation of four Map Task dialogues. Using the kappa coefficient[4], they showed that the coders had a very good agreement on their segmentation of the dialogues into moves, thus providing a solid foundation for move classification ($K = .92$,

[4]As explained by Carletta et al. (1997) "the kappa coefficient (K) (Siegel and Castellan, 1988) measures pairwise agreement among a set of coders making category judgments, correcting for chance expected agreement. $K = (P(A) - P(E))/(1 - P(E))$ where $P(A)$ is the proportion of times that the coders agree and $P(E)$ is the proportion of times that one would expect them to agree by chance"

$N = 4079$, $k = 4$). As for move classification, coders also had good agreement on the entire coding scheme ($K = .83$, $N = 563$, $k = 4$). The largest confusions were between (1) *Check* and *Query-yn*, (2) *Instruct* and *Clarify*, and (3) *Acknowledge*, *Ready* and *Reply-y*. Notice that these confusions would not have occurred to the same extent in the classification of moves in the current analysis, because as I mentioned above, the original set of move types was reduced to a smaller set of broader types. This smaller set eliminated the distinction between *Check* and *Query-yn*, which were included into a broader category *Query*. Also, it did not have the types *Clarify* and *Ready*. Next, the agreement on game coding was not as reassuring as move coding but still good. This was calculated as agreement on where games started and, for agreed starts, where they ended. On the start of games, coders showed a 70% agreement ($N = 203$). Finally, for transactions, the agreement reached by five coders as a group[5] on the identification of move boundaries as transaction boundaries was not as good as the previous results ($K = .59$, $N = 657$, $k = 5$). For this test, coders worked from the maps and transcripts, that is, they did not listen to the speech. This means they could not use intonational cues to identify transaction boundaries. Part of the problem was simplified by asking them to mark transactions only in particular sites of the transcripts marked with blank lines representing potential boundaries. These lines had been inserted between move boundaries except in between a *Ready* move and the following move. This was done because without prosodic cues a *Ready* move (e.g. *"Ok"*, *"Right"*) could have been interpreted in the text as a phrase closing the previous transaction instead of opening a new one. Nevertheless, it could be argued that not having access to the speech, coders probably missed many prosodic cues that could identify the start of transactions at the potential boundary sites they had available and this is perhaps the reason why the agreement was lower than for the tests on moves and games above. Thus, using the speech as it was done in the current study, would have yielded a better agreement between the different coders. One of the areas where coders disagreed, causing a lower agreement on transaction coding, was on introductory questions which started the description of a new route segment of the map. These are questions such as *'Have you got a pyramid?'*, when the *IG* is about to start describing a segment of the route using the pyramid. This could be corrected, as Carletta et al. explained, by clarifying the instructions. Being aware of this confusion, in

[5]This included four naive coders and the 'expert' developer of the coding instructions

the current study care was taken to identify such questions as transaction initial. And here again access to the speech in Carletta et al. would have caused less confusion and higher agreement.

In conclusion, results by Carletta et al. (1997) show evidence that the Conversational Games Analysis scheme can be used with good agreement by different coders. Furthermore, the fact that in the current study broader categories with less distinctions were used, as well as access to the speech when annotating, would ensure that the important aspects of the dialogue structure annotation in this study could very likely be reproduced reliably.

3.3.2 Pitch accents

Apart from their structure, another aspect of the collected dialogues that was investigated in relation to eyebrow raising was prosodic phenomena. This study will be reported in Chapter 5. In this section I will describe the scheme and procedure that was used for the annotation of pitch accents and related issues for that study.

Pitch accents were introduced earlier in this thesis in section 2.3.2 in relation to intonational prominence. Prominence is a perceptual phenomenon by which some words are perceived as more salient than others in an utterance, and in English it is associated to pitch accents. In this context, a pitch accent was defined as "a local feature of a pitch contour - usually but not invariably a pitch change, and often involving a local maximum or minimum - which signals that the syllable with which it is associated is prominent in the utterance" (Ladd, 1996, p. 46). Another related prosodic phenomenon that was introduced in that section was *downstep*. In a sequence of similar tones (usually High tones), downstep refers to a relation between them where the second one is realised with a lower F0 than the preceding tone to an extent that cannot be accounted for by background declination.

Pitch accent types

All pitch accents in the dialogues in this study were identified and then classified into one of five categories: *primary, secondary, downstep-initial, downstep-medial,* and *downstep-final.* Basically, they were labeled *primary/secondary,* except when

they were in a downstep group, in which case they were labeled according to their position (*initial, medial, final*) in the series of descending pitch peaks in the group. The five categories of accents in this system are described below:

***Primary* accent.** This was the default type.

***Secondary* accent.** This category was used when two (or occasionally more) words judged to be accented, and not in a downstep group, appeared closely linked intonationally as "weak-strong". An example would be the phrase "grey whale", where both words often sounded accented but "whale" was clearly more prominent. In conventional ToBI terms[6], these would be pairs of accents not separated by any break index greater than 1.

***Downstep-initial, downstep-medial,* and *downstep-final* accents.** These types were assigned to accents (two or more) that were clearly prosodically linked to each other and phonetically downstepping from one to the next. The first accent in the downstep group was classed as *downstep initial*, and the last one as *downstep final*. Any accents in between these two were *downstep medial*. In principle, *downstep-final* is a subtype of the *Primary* type, and *downstep initial* and *medial* are subtypes of the type *Secondary*. But in this study, downstepped accents were only classified in terms of their position in the downstep group.

This was a simple classification which did not take into account the levels of the constituent tones in the pitch accents. The purpose of the study was to investigate a possible relation between eyebrow raising and pitch accents as prominence cues. So the interest was in the salience of the accented syllable, rather than on whether this was realised as a pitch peak or valley, or a combination of both.

There are other prosodic phenomena that could have been labeled in the data, such as boundary tones. But due to known difficulties for annotators in distinguishing hesitation pauses from true intended boundaries, labeling boundary tones would have been problematic. This is more so because of the spontaneous nature of the conversations that contained frequent pauses.

[6]ToBI (Silverman et al., 1992) is a widely used system for transcribing the intonation patterns of spoken language

Annotation procedure

A phonetically trained expert[7] on intonation annotated the presence of pitch accents on the dialogue transcripts. The coder had access to the audio recordings but not to the videos. This was so that there was no bias from visual information on the identification of an acoustically accented syllable. If eyebrow raises or other facial movements could lend prominence without accompanying auditory cues, then by looking at the speaker's face while annotating pitch accents, the coder could be biased into marking some syllables as acoustically accented when in fact they were only accented visually. Though less likely, the reverse could also happen if coding with access to the images. That is, a syllable carrying a pitch accent could be wrongly labeled as acoustically unaccented due to bias from facial movements or the absence of movement on or around that syllable. In accordance with general transcription practices in the ToBI system (Silverman et al., 1992), a pitch accent was judged to have occurred if the coder perceived "prominence". Essentially this meant applying educated native speaker intuition about which syllables were prominent. All identified accents were classified by the coder into one of the five categories described above: *primary*, *secondary*, *downstep-initial*, *downstep-medial*, and *downstep-final*. And as explained, the High and Low constituent tones of the accents were not specified. However, as would be expected, the overwhelming majority of the accents were ToBI types H* or L+H*.

The annotation of pitch accents could have been done using the F0 display. In fact, as I mentioned in 3.2.2 when describing the materials for the dialogue recordings, the landmark names on the maps used by the participants were created to contain as many sonorants as possible to facilitate F0 tracking in the speech signal. However, this would not solve the problem for parts of the speech other than landmark names where obstruents may disturb the clear F0 display. In the interest of obtaining as much data as possible, it was decided that annotating the accents by carefully listening to the recordings would be a better option. The expertise of the coder in this area ensured a good judgment on identification of pitch accents.

[7] I am grateful to Prof. D. Robert Ladd for his assistance

The coder reported making two independent passes of the recorded materials. The second pass was done after refining the criteria for marking the accents and is the version that was used in the analysis. In particular, the second pass had some additional accents on monosyllabic words and utterance initial hesitations that were not coded on the first pass. It also had some secondary accents that in the first pass had been either primary or not marked. It may be argued that having only one coder poses problems on the reliability of the system used to annotate the pitch accents. However, as mentioned earlier in section 2.3.2, it has been proved that there is very good agreement (80.6%) on the identification of presence/absence of pitch accents by listeners (Pitrelli et al., 1994). As for the categories that the coder used to classify pitch accents, the reliability of their annotation has not been tested, but the high level of agreement between the coder's first and second pass is promising.

For the purpose of the study it was necessary to know the exact temporal location of pitch accents in the dialogues. Pitch accent codes from the dialogue transcripts of the *IG* speakers were recorded as codes for portions of the digitised speech signal. *Xlabel* from *Entropic/Xwaves* was used to assign a pitch accent onset label to the start and a pitch accent offset label to the end of the accented syllable. An alternative would have been to mark a single point in the syllable where the F0 reached its maximum excursion. However, when looking at pitch accents and eyebrow raises together in the current study, the interest was in the start of the events, and as shown by Ladd and Schepman (2003), the pitch excursion normally begins very close to the start of the syllable.

3.3.3 Information structure

As we saw in section 2.3.3, information structure refers to the way linguistic messages are structured into information units with different relationships to previously presented information. One distinction traditionally made in this context is that between *new* and *given* information. In conversation, speakers can refer to an entity that has not been previously mentioned and is considered new to their interlocutor, or they can refer to an entity which they believe is known to the addressee. The referring expressions used in each case have a different information status, one being *new* and the other being *given*. This distinction can be

marked in different ways. One way is to use prosodic features such as intonational prominence: accenting the referring expression introducing a new entity, and deaccenting the expression used to mention the same entity later on. However, speakers in Map Task dialogues have been shown not to use intonation in this way to mark the distinction *new/given* in referring expressions (Bard and Aylett, 1999). In this thesis I investigate whether speakers in the dialogues collected here make use of eyebrow raising as a cue to mark the contrast between *new/given* information when referring to landmarks on their maps. In order to do this, the following classification was made for referring expressions mentioning landmarks:

***First* mention.** A referring expression that referred to a landmark on the map that had not been previously mentioned in the dialogue by either of the speakers. This expression was normally the name that appeared on the label, but could be reduced in some way, for instance by using only part of the name: e.g. *gannets*, to refer to *hungry gannets*. In terms of information status, *first* mentions were considered to be *new* information.

***Second* mention.** A referring expression that referred to a map landmark that had been previously mentioned once by the same speaker earlier in the dialogue. This could be again the name on the landmark label or any reduced version of it, including pronouns. In terms of information status, *second* mentions were considered *given* information.

***Other* mention.** Any other referring expression that mentioned a landmark and did not fall into either of the two categories above.

In principle, the first mention of an entity is not necessarily always presented as *new* information. This is only the case if the speaker introducing that entity into the discourse believes that it is unknown to the interlocutor. But in the dialogues under investigation this was generally the case and so the classification into *new/given* information was based on whether the entity had already been mentioned or not.

Annotation procedure

The information status of referring expressions in the dialogues was annotated for the *IG* speakers. First, referring expressions mentioning map landmarks were

labeled on the transcripts as being a *first* mention, a *second* mention, or *other*, according to the definition above. Then, these labels were added to the electronic record of pitch accents occurring on those referring expressions.

3.3.4 Eyebrow raises

To investigate whether eyebrow raising was related to the linguistic phenomena described above, eyebrow raises in the corpus were annotated on the same timeline as the linguistic events. The scheme and annotation procedure devised for this purpose are described below. Examples of eyebrow raises will be provided in a series of images at the end of this section to conclude the chapter.

First, the movement under investigation was defined:

Eyebrow raise was any upward movement, from a baseline neutral position, of at least one eyebrow and observable by the author on the digital video recordings.

Notice that this definition would include different types of raising movements, including those in which only part of the eyebrow is lifted, such as the inner corners. Comparing it to the brow action units in the Facial Action Coding System (Ekman and Friesen, 1978), this definition would include, without distinction, units 1 and 2, and combinations of these, namely, 1+2, 1+2+4 and 1+4. See figure 2.2 in the previous chapter, and the following description summarised from Ekman et al. (2002):

AU 1 This action is the *inner brow raiser*, which pulls the inner portion of the eyebrows upwards. It normally causes horizontal wrinkles in the center of the forehead (except in infants and children) and may produce an oblique shape to the eyebrows.

AU 2 This is the *outer brow raiser*, which pulls the lateral (outer) portion of the eyebrows upwards. It stretches the lateral portion of the eye cover fold upwards, and it produces an arched shape to the eyebrows. For some people it can cause short horizontal wrinkles above the lateral portions of the eyebrows, and sometimes also in the center of the forehead but not as deep as the lateral ones.

AU Combination 1+2 This movement pulls the entire eyebrow upwards (both medial and lateral parts), producing an arched, curved shape. It bunches the skin in the forehead so that horizontal wrinkles may appear across the entire forehead. This movement was the most frequently observed in the corpus in the current investigation.

AU Combination 1+2+4 This is the combination of AU 1+2 and AU 4 (*brow lowerer*), but the appearance produced is not simply the addition of the changes observed in 1+2 and 4 individually. The eyebrows are pulled up and together but not as much as is done by 1+2 and 4 alone. This combination flattens the shape of the eyebrow between the inner corner and the middle portion. It bunches the skin in the central portion of the forehead producing horizontal wrinkles or wrinkles that show a little upward curve.

AU Combination 1+4 This pulls the medial portion of the eyebrows upwards and together (sometimes they may not appear to be drawn together). It also pulls up the mid to inner portions of the upper eyelid, and pulls the lateral portion of the brow down. The latter movement down is due to the partial action of AU 4 (the *brow lowerer*).

All these movements involve some lifting of the eyebrows and thus all were considered eyebrow raises in the scheme in this thesis. In principle they could have been classified into different categories of eyebrow raising, depending on their anatomical differences involving different muscle actions. However, these differences can be very subtle and doing this would have implied a considerable amount of training time and of time spent on annotation to allow consistent identification of the different categories. Also, since 1+2 was by far the most frequent movement, the other categories would have yielded rather small groups of data.

When saying *any upward movement*, in the definition of eyebrow raise above, movements of any intensity are included. The Facial Action Coding System (Ekman and Friesen, 1978; Ekman et al., 2002) provides a way of doing intensity scoring of its action units. However, this can lead to subjective scoring and was not considered appropriate for the current study. All upward movements observed by the coder were labeled as eyebrow raising regardless of the magnitude of the movement. The above definition also says it is a movement of *at least one eyebrow*. Cavé et al. (1996), in section 2.5.2 above, reported a link between fundamental frequency patterns and movements of the left eyebrow that

was not found for the movements of the right eyebrow. Asymmetries in facial movement between the right and left side of the face are common. In the current study some asymmetries between the right and left eyebrow were sometimes observed, but they were considered not relevant for the investigation and were therefore ignored. Thus, the upward movement of one eyebrow was qualified as an eyebrow raise regardless of whether the other eyebrow moved with the same magnitude or not.

Annotation procedure

The speakers' eyebrow raises in the collected dialogues were annotated by the author using an observational system. Automatic procedures of extracting information on facial movements were considered inappropriate for several reasons. As we saw in section 2.6, facial EMG techniques have the disadvantage of requiring the attachment of electrodes to the participant's skin. This is particularly undesirable when two participants are interacting face to face as in the current study. For the same reason, computer vision systems that required placing markers on participants' faces (such as the technique used by Cavé et al., 1996, 2002) were also rejected. As we saw in 2.6, other modern computer vision techniques do not require markers on the skin and are therefore not intrusive. However, these systems were not developed enough to the level necessary for this investigation at the time when the data annotation was done. The human visual system is still the best facial expression analyser, as pointed out by Pantic and Rothkrantz (2000), who take it as a reference point for the development of automated systems, since it can deal with obstacles such as head motion and partial occlusion of the face. For the current study a human observational coding system was considered the best option. This system had some disadvantages, such as the amount of time spent on annotation by human coders and also the fact that it can be unreliable. I will address this point further below.

Brow raises in the collected dialogues were annotated using the software *SignStream* (version 2.0) (Boston University, USA). This multimedia database tool allows the frame-stamped segmentation and labeling of digital video data. Thus, brow raises were annotated, for each speaker, as portions of the video signal with their start and end frame number. The digital video recording was observed in short sections, first at normal speed and then in slow motion several times until

the coder was confident of the presence or absence of brow raising in that particular section. If there was brow raising, the section was played again frame by frame until the start of the upward displacement of the eyebrows was observed and labeled in the immediately preceding frame. The preceding frame was labeled because due to the rate of video recording (25 frames per sec), a change observed in one frame could have actually started just before that image was captured. The end of the brow raise was marked on the frame where the last downward displacement of the eyebrows was observed. It is useful to point out that the start of the brow raise was easier to identify than its end.

The annotation was done without sound to avoid bias from the speech. At first, it was done with a full view of the face. But it was felt that other facial movements, such as lip or jaw movements, could be distracting at times, and then it was decided to hide the bottom part of the speaker's face during the annotation. This also seemed an appropriate measure to avoid a possible bias or interference from some articulatory movements which have been reported to correlate with stress (e.g. Keating et al., 2003; Erickson et al., 1998; Erickson, 1998, 2002, see section 2.5.3). Another type of movement that sometimes was felt to interfere with the annotation of eyebrow raising was head movement. The movement of the head changed the normal front view of the face, altering the relative visible distance of the different facial features. When the head was lowered, for instance, the visible vertical space between the eye and the eyebrow was narrowed and the upward displacement of the eyebrows was not so clear as when the face was in upright front position. This was not a problem when the eyebrow movement occurred while the head was relatively static. But if the head was in motion as the eyebrows were raised, more care was necessary to factor out the rotation and concentrate on the eyebrows to identify raising movements. In practical terms, this meant that the annotator had to spend more time on these segments, watching the images several times in slow motion. In these cases, the wrinkles that appeared on the forehead as the eyebrows were raised were a reliable cue of a raising movement.

All identified brow raises were annotated regardless of the amplitude of the rise. Sometimes, for very slight movements, it was difficult to decide whether there was brow raising or not. In some cases, changes in appearance on the forehead's skin aided identification of some very minor raises that would not have been

easily perceived at normal video speed. The fact that some movements were difficult to perceive by the coder brings up the issue of perceivability by the interlocutor in the dialogue. The current study investigates the *production* of eyebrow raising by a dialogue participant even when this behaviour might not have been perceived by the other participant. For this reason, all eyebrow raising was labeled without considering whether the interlocutor was looking at the speaker's face or not and whether the magnitude of the eyebrow raise was enough to have been perceived by her. The difficulty in identifying minor eyebrow raises could be a problem, however, in terms of the reliability of the annotation system. Because it means that different coders might not perceive and annotate an equal number of brow raises in the same data set. Reliability issues will be dealt with further below.

Another difficulty encountered, though not very frequently, involved what appeared as brow raising 'superimposed' on another brow raise. This happened when the eyebrows were raised, remained up for some duration, and from there they were raised further up, as if this second movement was embedded within the longer brow raise. The decision here was to annotate only one single brow raise from the first elevation of the eyebrows until their final lowering to the baseline position.

A more serious problem was related to the variability between speakers. Differences in facial physical appearance were not a problem. The four participants in the corpus had different eyebrow shapes and colour, but this did not hinder identification of eyebrow raising across them. However, differences in their eyebrow raising style was a problem and resulted in the exclusion of one speaker from the analysis. This speaker (*B1*) had some clear instances of eyebrow raising that did not differ much from the behaviour of the other speakers. She had a neutral baseline position from which she clearly lifted and then lowered her eyebrows, in the same style as the other participants (examples of eyebrow raising from all participants are illustrated at the end of this chapter). But she also had a tendency to arch her eyebrows when she was speaking, and to maintain them in a slightly raised position for very long stretches of time. In those long stretches, the raising and lowering of the eyebrows could be very gradual, making it difficult to identify the exact point at which that *raised state* started and ended. This

was a marked difference from the eyebrow raising style of the other three speakers. Also, from this raised position she could lift the eyebrows further up, as in a superimposed brow raise, or she could lower them to her original neutral position. It was almost as if she had two baseline positions, one slightly more raised at which the eyebrows could be kept during a series of utterances. However, it was difficult to find a consistent way of annotating her behaviour and this posed problems, especially for the reliability of the annotation scheme. Eventually, it was decided that speaker *B1* should be excluded from the analyses presented in the next two chapters. This decision was made primarily in the interest of time. It is very likely that with longer time available it would have been possible to observe *B1*'s behaviour more carefully to achieve a fairly consistent annotation of her eyebrow raises. However, this would have delayed the analysis and it also would have meant a different annotation procedure for this participant, which, among other things, would have posed problems for intercoder reliability.

The annotated record of eyebrow raising from the remaining three speakers was exported as text and, after conversion of frames into seconds, it was combined with the other annotated events described above to create a single timestamped record of visual and auditory events.

Reliability of the scheme for identification of number of eyebrow raises

As a human observational system, the scheme used to annotate eyebrow raises in the current corpus could lead to subjective and unreliable scoring. To test the reliability of the scheme, a very basic test was conducted. A second coder[8] was presented with a small set of the data and was asked to specify the number of eyebrow raises she observed. The coder was a research associate at *The University of Edinburgh* who had not been involved in any research on body movement. The set of data consisted of ten utterances (conversational moves) from each of the three speakers. In each set there were five utterances where the first coder had identified one eyebrow raise and five utterances that had been coded as having zero eyebrow raises. These were presented to the second coder in random order, and with the same procedure used in the first coding, e.g. using the same program, without sound, and hiding the lower part of the face of the recorded

[8]I am grateful to Francesca Filiaci for her assistance

participants. Before starting the task, the second coder was given a set of instructions on how to identify eyebrow raising. These instructions are attached in Appendix C.

Intercoder agreement between the first and second coder was measured on the number of eyebrow raises observed per utterance[9]. Results showed a 96.6% agreement: the two coders agreed on the number of eyebrow raises in 29 of the utterances, and they disagreed on one utterance (from speaker *A2*) in which the first coder identified one eyebrow raise, whereas the second coder reported two instances of eyebrow raising.

Although this was a very basic measure on a very small set of the data, the high level of agreement could be taken as tentative evidence that the scheme to identify eyebrow raising could be used reliably. A test of a much larger scale is obviously necessary in order to obtain an adequate measure. Also, a test is still necessary on the intercoder agreement on the *temporal location* of eyebrow raises in the dialogues. No resources were available at the time of this investigation to carry out this test.

Examples of eyebrow raises

Examples of eyebrow raising from each of the four participants recorded are provided in a series of figures below. First, it is important to point out that some eyebrow raises are not easily perceived on still images, especially the small, subtle movements. Even if the full sequence of recorded frames is presented from start to end, the brow movement can be difficult to perceive because of the lack of motion and the fact that the images are laid out side by side, instead of superimposed as when the frames advance on a video display. The examples in the figures below are presented in just three video frames for each single brow raise (except for Figures 3.16 and 3.19). In each figure, these three frames show, from left to right, (1) the start of the eyebrow raise, (2) the approximate time at which

[9]The second coder was asked to identify the *number* of eyebrow raises she observed. But the stimuli only differed in the presence/absence of one single eyebrow raise. Thus, the test for intercoder agreement could be considered a measure of agreement on presence/absence of brow raising, more than on the actual number of eyebrow raises

the eyebrows were maximally raised[10], and (3) the end of the brow raise, as annotated in the corpus. Frame numbers have been transformed so that the first image always starts at frame 0, and in the next two images, this is increased by the number of skipped frames between the first image and the current one. The frame sequence is specified in this way below each figure, with its equivalent in seconds in brackets. For instance, "**Frame sequence** (*sec*): **0, 19** (.76), **53** (2.12)" means that the second image was captured 19 frames after the first one (corresponding to a lag of $.76sec$) and the third image was 53 frames after the first one, corresponding to a lag of $2.12sec$ which is the total duration of this brow raise.

In many of the images presented here the participant is looking down at her map. This was quite typical since their task involved the description of a map which was placed on a raised board on their table. When the participant appears looking up (or ahead), she is looking at her interlocutor who is sitting in front of her, as explained earlier in this chapter. Sometimes subjects looked up at some point between the second and third frames shown in the figures. This was the case for instance in Figure 3.13.

The first three examples are from Speaker *A1*. In the first figure (3.4), wrinkles on the speaker's forehead and a clear upwards displacement of the brows illustrate well the eyebrow movement. The next example, Figure 3.5, is a much smaller movement which is not so easily perceived on the still images (it is also much shorter). This type of subtle movement was very frequent in Speaker *A1* and required careful observation of the videos. The fact that subtle movements like this one could be perceived during annotation but are not clearly discernible on the still images emphasises the importance of doing a careful and detailed analysis of the images in motion, reducing the speed as necessary, when studying facial movements in this kind of research. The example in Figure 3.6 is an asymmetrical brow raise, in which only the left eyebrow was raised. Because of the position of her head, the left side of the participant's face is not fully visible. The left eyebrow is in fact only seen partially. Here again, the slight wrinkles on her forehead as the brow goes up help in the identification of the movement. In the

[10]It must be remembered that the point of maximum rise was not labeled on the corpus. Here it is presented for the purpose of the illustration

last example from this participant (Figure 3.7) there seems to be also some asymmetry but less pronounced because both eyebrows were raised in this case, even if not to the same extent.

The next series of images show examples from Speaker *A2*[11]. This participant had thin, blonde eyebrows which were not as easy to see as in the other three speakers. This, however, did not seem to hinder identification of her eyebrow movements, especially when wrinkles formed in the forehead as she lifted her eyebrows. These can be seen clearly in Figures 3.8 and 3.9, with the speaker looking at the interlocutor at the point of maximum rise. In Figure 3.10 the inclination of the head makes it harder to see the displacement of the eyebrows. But notice how the wrinkles above the left eyebrow in the middle figure indicate clearly that this eyebrow was lifted. This asymmetrical movement was very easy to perceive when watching the motion in the video. The next figure (3.11) illustrates an example very similar to that in 3.5 above (in fact it has the same temporal properties). This is another case of a very subtle movement, both in magnitude and duration. Here again, it is very difficult to appreciate any upwards displacement of the brows in the sequence of still images. However, when watching the video this segment was perceived as an eyebrow raise, especially when played in slow motion.

Brow raises from Speaker *B2* are illustrated in Figures 3.12 to 3.15. Of the four participants, *B2*'s eyebrow raises were the easiest to perceive and annotate. On most occasions she had a pronounced movement, with a clear start and end. Notice the short and regular lag between the first and the second image in each example. This indicated a fairly fast and consistent movement from the start to the maximally raised position, something that would have certainly aided identification of the whole movement. Figures 3.12 and 3.13 show a clearly visible raising of the eyebrows, with wrinkles forming on the forehead. The sequence in Figure 3.14 shows another clear lifting of the eyebrows, this time ending in a slight frown. The next example (Figure 3.15) presents a shorter eyebrow raise combined with squinting of the eyes, which was observed in this speaker on several cases.

[11]The lower part of each image has been pixelated to protect the anonymity of this participant

Finally, some sample frames from subject *B1* are provided as well. As explained earlier, when describing the procedure for the annotation of eyebrow raising, the eyebrow movements displayed by this participant presented some difficulties for annotation. Eventually, in the interest of time and to avoid inconsistencies on her annotation, the data collected from *B1* was not included in the analysis. The interesting behaviour from this participant is particularly difficult to illustrate here. In Figure 3.16 *B1*'s eyebrows are in what can be considered their neutral baseline position. Figure 3.17 shows an example of eyebrow raising that was easily identified as such and did not seem very different from the type of behaviour observed on the other participants. Figure 3.18 presents another example identified as a single eyebrow raise. Here the duration is longer than the previous examples from the other three speakers, but it is still within the range observed. The next two images in Figure 3.19 present two instances of the type of brow display that posed problems for annotation. The raised position of the eyebrows here was held for several utterances and the points at which it began and ended were in most cases difficult to identify because of a very gradual movement. It was difficult to decide whether these *states* should be counted as eyebrow raising or not. From this position the eyebrows sometimes were raised further up, as if producing a brow raise from this "raised baseline position". This is illustrated in the sequence of frames in Figure 3.20. The first image there is in fact the same as the second image in 3.19. This brow position had been already held for more than one utterance and then, after the initial frame in Figure 3.20, the eyebrows were lifted further, reaching their maximum rise in the second image before going gradually down again. Other movements not exemplified here posed challenges in the annotation of B1's eyebrow raises. It would be interesting to observe more carefully her facial behaviour in a future study.

Figure 3.4: Speaker *A1*. **Frame sequence** (*sec*): **0, 19** (.76), **53** (2.12)

Figure 3.5: Speaker *A1*. **Frame sequence** (*sec*): **0, 2** (.08), **5** (.20)

Figure 3.6: Speaker *A1*. **Frame sequence** (*sec*): **0, 25** (1), **42** (1.68)

Figure 3.7: Speaker *A1*. **Frame sequence**(*sec*): **0, 11** (.44), **27** (1.08)

Figure 3.8: Speaker *A2*, **Frame sequence** (*sec*): **0, 10** (.40), **37** (1.48)

Figure 3.9: Speaker *A2*. **Frame sequence** (*sec*): **0, 7** (.28), **25** (1)

Figure 3.10: Speaker *A2*, **Frame sequence** (*sec*): **0, 8** (.32), **37** (1.48)

Figure 3.11: Speaker *A2*, **Frame sequence** (*sec*): **0, 2** (.08), **5** (.20)

Figure 3.12: Speaker *B2*. **Frame sequence** (*sec*): **0, 5** (.20), **19** (.76)

Figure 3.13: Speaker *B2*. **Frame sequence** (*sec*): **0, 5** (.20), **46** (1.84)

Figure 3.14: Speaker *B2*. **Frame sequence** (*sec*): **0, 4** (.16), **25** (1)

Figure 3.15: Speaker *B2*. **Frame sequence** (*sec*): **0, 6** (.24), **13** (.52)

Figure 3.16: Speaker *B1* in neutral eyebrow position

Figure 3.17: Speaker *B1*. **Frame sequence** (*sec*): **0, 4** (.16), **39** (1.56)

Figure 3.18: Speaker *B1*, **Frame sequence** (*sec*): **0, 7** (.28), **95** (3.8)

Figure 3.19: Speaker *B1*. Two instances of raised eyebrow position that was held for several utterances

Figure 3.20: Speaker *B1*. **Frame sequence** (*sec*): **0, 97** (3.88), **107** (4.28). Example of brow raising within already raised brow position

CHAPTER 4

Eyebrow raising: discourse structure and utterance function

4.1 Introduction

There have been several studies suggesting a relationship between body movements and **discourse structure** (e.g. McNeill et al. 2001, for hand gestures; Cassell et al. 2001, for postural shifts; McClave 2000, for head movements). About eyebrow movements in particular, Ekman (1979) observed that when describing a series of events, the eyebrows could act as punctuation marks, like a comma. Chovil (1989, 1991a), in her classification of facial displays in dialogue, pointed out that brow raises could mark the beginning and end of a topic, and the continuation of a topic after detracting from it, thus helping to structure the conversation. And it has also been reported that eyebrow raises can have a turn-taking role by signalling a new turn in a conversation (Cavé et al., 2002). In relation to the **function of utterances**, eyebrow raising has been traditionally associated with questioning (e.g. Birdwhistell, 1970; Eibl-Eibesfeldt, 1972). Both Ekman (1979) and Chovil (1989, 1991a) suggested question marking as one of the conversational functions of brow raises. Srinivasan and Massaro (2003) found that both eyebrow raising and head tilting could be used, together with auditory cues, to distinguish echoic questions from statements. But they reported that participants relied most strongly on the auditory cues, even when the visual cues were enhanced.

In summary, observations in previous research suggest eyebrow raising could be related to dialogue structure and utterance function. But these claims need further support. For instance, Ekman's observations (1979), as he explained, were preliminary and did not come from an empirical study. The functions reported by Chovil (1989, 1991a) were derived with an inductive approach and had a very small sample in each group. And Cavé et al. (2002) studied data from French speakers, who may behave differently from English speakers. The study presented in the current chapter investigated a possible relation between eyebrow raising and both discourse structure and utterance function in Map Task dialogues in English.

Following the coding scheme described in Chapter 3 (Carletta et al., 1997), the structure of Map Task dialogues is divided into different levels - *conversational moves*, *games*, and *transactions* - associated with the purpose of the speaker. If eyebrow raises are related to dialogue structure, then they could mark the beginning of one of these levels to announce a shift into a new theme or goal in the conversation. Thus the first hypothesis was:

H1a: Brow raises will be unequally distributed across different levels of the dialogue structure. There will be more brow raises in moves starting new conversational games and transactions.

Games and transactions are composed of different types of moves according to their particular conversational purpose: e.g. to request information, to give an instruction, etc. It was hypothesised that eyebrow raising could contribute to conveying this purpose. In particular, brow raises were expected to occur more frequently in *Instruct* moves because these utterances contain the most important information to advance the dialogue and complete the task (see section 3.3.1). That is, in theory *Instruct* moves alone could enable the speakers to complete the task of reproducing the map route on the *Instruction Follower*'s map. Brow raises were expected to help convey the importance of this key information. Additionally, as we have seen, the literature mentioned above suggests that we raise our eyebrows when we ask a question. Thus, *Query* moves in the dialogues under investigation were also predicted to have more frequent eyebrow raising. The second hypothesis was stated as:

H1b: Brow raises will be unequally distributed across moves with different purposes. *Instruct* moves and *Query* moves will have more brow raises than other types of move.

If, however, eyebrow raises are not related to the structure of the dialogue or to the purpose of the utterance and they simply occur at random, then we would expect only effects of the opportunities for brow raises, which would vary as a function of the duration of the sampled unit:

H0: Brow raises are a random phenomenon determined only by utterance length. Long moves will have more brow raises than short ones but uptake of opportunities will not depend on type of move.

So far this discussion has considered only the occurrence of brow raises, as if all had the same duration. But, of course, they do not. What may be associated with discourse then is the *duration* of eyebrow raising, not its frequency. Thus, another set of hypotheses was made, with similar predictions, about brow raise duration:

H2a: Total brow raise duration (per move) will be unequally distributed across different levels of the dialogue structure. There will be longer eyebrow raising in utterances starting new conversational games and transactions.

H2b: Total brow raise duration (per move) will be unequally distributed across utterances with different purposes. *Instruct* moves and *Query* moves will have longer eyebrow raising than other types of move.

And again, the null hypothesis would state that:

H0: Brow raises are a random phenomenon only determined by utterance length. Long moves allow longer total brow raise duration than short utterances.

4.2 Method

4.2.1 Materials

The materials came from the collected corpus of task-oriented dialogues described in Chapter 3. The data in this study corresponded to six dialogues in

which each of three speakers (*A1, A2, B2*) participated twice as *Instruction Giver* (*IG*) to two different *Instruction Followers* (*IFs*). The dialogues had an average duration of $369sec$. The utterances and brow raises analysed belonged to speakers in the role of *IG* only, but the *IF's* utterances were obviously considered when determining utterance function and discourse level in the *IG's* speech.

Dialogue structure

As described in Chapter 3, the structure of each dialogue was annotated according to the Conversational Games Analysis coding scheme (Carletta et al., 1997), yielding three conversational levels: moves, games, and transactions.

Conversational moves were segmented as portions of the digitised speech signal, with labeled start and end times. They were classified by type according to their purpose in the communicative task. As explained in section 3.3.1, for the current analysis five categories were used in the classification of moves, taken from a larger set of move types in Carletta et al. (1997): *Instruct*, *Explain*, *Query*, *Reply*, and *Acknowledge*. A sixth category, *Unclassifiable*, was used to label moves whose purpose was unclear because for instance they were interrupted, and moves within this class were excluded from the analysis. The conversational move was the unit of analysis in this study.

Conversational games were also labeled as portions of the speech signal. The start label was aligned with the start of its first constituent move, and the end label was aligned with the end of the last move.

Transactions were labeled as the highest level of the structure of the dialogue. The start time was aligned with the start of the first game (and move) within it. The end of a transaction finished with the start of the next one, and therefore only start timestamps were marked at this level.

Across the six dialogues *IG* speakers produced a total of 682 moves (excluding 30 *Unclassifiable*). As for games and transactions initiated by the *IG*s, those moves were included into 185 games, which were in turn combined into 104 transactions.

Eyebrow raises

The start and end of brow raises were recorded on the timeline of the dialogue, together with the conversational events described above. Among the six dialogues, there were 274 brow raises produced by the *IG* speakers. Four of these occurred without accompanying speech by the same speaker: they started and ended in an inter-move interval (IMI), i.e. after the end of a move and before the start of the next move. Since the unit of analysis was the move, those four cases were excluded, leaving a total of 270 eyebrow raises.

How were brow raises associated with a particular move? The eyebrows can be raised and lowered at any point in the dialogue structure and, therefore, it was necessary to establish some criteria that would determine when a brow raise 'belonged' to one move or another in this analysis. Brow raises that started in a move and ended within that move posed no problem. But a brow raise could start in a move and end in the next one. Or it could start in a move and end outside it but before the next move, in the IMI. Brow raises could also start in an IMI and end in the next move, or could go on across several moves. To associate a brow raise with a particular move, then, the first step was to always look at the point in which the brow raise started. This was not a random choice. The annotation of the data described in Chapter 3 showed that the start was generally more marked and perceptually clearer than the end of a brow raise. As a change in the behavioural flow, the start of the rarer event (eyebrow raise) should have more significance than its end, which coincides with the resumption of the default (no brow raise). Thus, **a brow raise was associated with the move in which it started**. When a brow raise started in an IMI (between moves) and finished after the start of the next move, then it was associated with that move. And as explained above, four cases that started and finished in the same IMI were excluded from the analysis. These criteria were used to count the *number of brow raises* per move for the statistical analyses reported below.

As mentioned earlier in 4.1, the association between brow raises and moves was considered from a different point of view as well. A measure of the *total duration of eyebrow raising* per move was calculated by adding up the portions of that move that were accompanied by raised eyebrows, regardless of where those

brow raises started and ended. This meant that an initial or final brow raise portion that fell outside the move did not add up to the total duration of eyebrow raising in that move.

4.2.2 Statistical analysis

The unit of analysis in this study was the conversational move. **Multiple regression** analyses were carried out in order to examine a possible relationship between brow raises and both dialogue structure and utterance function, according to the above hypotheses. The independent contribution of the predictors was evaluated and diagnostics were performed to detect possible multicollinearity between the variables. The dependent variables and the predictor variables included in the multiple regression analyses are listed below. The abbreviation 'BR' will be used in places to refer to 'eyebrow raise'.

Dependent variables

Two separate dependent variables were used, one at a time, measured with the criteria described in the previous section:

DV1. **Number of BRs per move**
DV2. **Total BR duration per move**

Predictor variables

The following predictor variables were used:

- *Move type*: *Instruct, Explain, Query, Reply, Acknowledge*
- *Discourse position*: *Transaction initial, Game initial, non-initial*
- *Speaker*: *A1, A2, B2*
- *Move length* (number of words)

In multiple regression each predictor is assessed for its ability to account for variance in the dependent variable in a situation where the values of the other predictors are statistically held constant. The predictor variables above were entered into the equation in blocks by a stepwise procedure in the order listed. Categorical predictor variables were entered as *dummy variables*. That is, within each

categorical variable (*move type*, *discourse position*, and *speaker*) a new variable was created for each group and was coded with zeros and ones. Then, all groups minus one, within each category, were entered into the regression analysis. The group left out in each categorical variable was the reference group to which the other groups within that variable were compared. For *move type*, the reference group was *Instruct* and *Query*, one at a time as described below, to which the other move types were compared. *Discourse position* was entered as two separate binary variables, *transaction initial* and *game initial*[1] that were compared to *non-initial* position in each case (non-initial in transaction and non-initial in game)[2]. As for the categorical variable *speaker*, preliminary observations of the data seemed to indicate that speaker *A1* raised her eyebrows more frequently than the other two speakers, and so she was selected as the reference to which speaker *A2* and *B2* were compared. The speakers were included in the analysis to evaluate speaker variability and the contribution of the other predictors independently from this. Similarly, *move length* was included to assess the predicting value of the other variables independently from the number of words in the utterance.

4.3 Results

4.3.1 Some descriptive statistics

Before reporting the results of the multiple regression analyses, tables 4.1, 4.2, and 4.3 are presented below to describe the data set.

[1]The category *game initial* did not include any moves that were also the first move in a transaction. This was to maintain independence between the *transaction initial* and *game initial* variables

[2]In this sense discourse position differed slightly from the way the categorical variables *move type* and *speaker* were coded, since it was coded as two binary variables compared to a different reference group each

		Speaker			
Move type	(%)	*A1*	*A2*	*B2*	Total
Instruct	(41.6%)	73	102	109	284
Explain	(8.7%)	11	21	27	59
Query	(13.8%)	20	42	32	94
Reply	(19.8%)	62	33	40	135
Acknowl.	(16.1%)	37	41	32	110
Total	(100%)	203	239	240	682

Table 4.1: Number of conversational move types by speaker

		N of words		*N of BRs*		*BR dur*(sec)	
Move type	(N)	Mean	SD	Mean	SD	Mean	SD
Instruct	(284)	10.58	5.54	.65	.821	.834	1.10
Explain	(59)	9.10	4.69	.44	.702	.376	.612
Query	(94)	6.40	3.47	.29	.500	.346	.564
Reply	(135)	3.11	3.57	.21	.447	.157	.386
Acknowl.	(110)	1.25	.747	.02	.134	.007	.04
Overall	(682)	6.90	5.74	.39	.675	.460	.848

Table 4.2: Mean and SD for *move length, N of BRs,* and *Total BR duration,* by *move type*

		N of words		*N of BRs*		*BR dur*(sec.)	
Disc. Pos.	(N)	Mean	SD	Mean	SD	Mean	SD
Trans-init	(104)	10.79	5.48	.75	.76	.771	.971
Game-init	(185)	9.43	5.63	.49	.716	.517	.760
Non-init	(393)	4.67	4.75	.25	.586	.349	.832
Overall	(682)	6.90	5.75	.39	.675	.460	.848

Table 4.3: Mean and SD for *move length, N of BRs,* and *Total BR duration,* by *discourse position*

4.3.2 *Number of brow raises*

The resulting final model, from the first analysis, is reported below, with the independent contributions of the significant predictor variables presented in Table 4.4:

$R = .504, R^2 = .254$

$F_{(8,673)} = 28.655, p < .001$

The standardised regression coefficient β is a measure of the independent contribution of a predictor variable to the variance in the dependent variable, with the

Predictor	β	Sig.
Acknowledge	$-.105$	.020
Query	$-.095$	.012
Trans. initial	.101	.006
Speaker A2	$-.178$	$< .001$
Speaker B2	$-.204$	$< .001$
Move length	.379	$< .001$

Table 4.4: Independent contribution of the significant predictors of *Number of BRs* (move types compared to *Instruct*)

other variables statistically held constant. And because β values are expressed in standardised units they can be directly compared across predictors. Table 4.4 shows the predictors that make a significant contribution to the model. For each predictor variable, β indicates how much the number of BRs will change with a change of one standardised unit in that predictor variable. The values for the individual move types and the individual speakers are in comparison to the reference groups: *Instruct* type and speaker *A1*, respectively. Negative β values for those categorical variables indicate that the predictor in question has significantly *fewer* BRs than the reference group.

Overall, the resulting model accounts for 25% of the variance in the number of BRs per move. In order of absolute values of β (strength of prediction), the influences from the significant predictors on the number of BRs are as follows[3]:

- *Move length* ($\beta = .379$, $p < .001$), with more BRs as the number of words increases
- Speaker *A1* produces more BRs per move than *B2* ($\beta = -.204$, $p < .001$) or *A2* ($\beta = -.178$, $p < .001$)
- *Instruct* moves have more BRs than *Acknowledge* moves ($\beta = -.105$, $p < .05$) or *Query* moves ($\beta = -.095$, $p < .05$)
- *Transaction initial* moves ($\beta = .101$, $p < .05$) have more BRs than moves which are in non-initial position

[3]Notice that, if following a strict order, the last β value reported here should be that of *Query*. However, for ease of exposition and because it is only slightly smaller than the value for *Transaction initial* it is reported next to *Instruct*, the other significant predictor within the move type categories

Considering a possible relation between BRs and utterance function, moves of type *Query* in this study were predicted to have more BRs than other moves. The results above showed that *Query* moves had fewer BRs than *Instruct* moves, but this analysis did not compare *Query* with the other move types. Thus, another multiple regression analysis was done, with the same variables as the previous one, but this time making *Query* the reference group to which other move types were compared. The results are reported below, with the significant predictors in Table 4.5.

$R = .503$, $R^2 = .253$

$F_{(7,674)} = 32.639, p < .001$

Predictor	β	Sig.
Instruct	.108	.015
Trans. initial	.092	.009
Speaker A2	$-.184$	$< .001$
Speaker B2	$-.208$	$< .001$
Move length	.372	$< .001$

Table 4.5: Independent contribution of the significant predictors of *Number of BRs* (move types compared to *Query*)

These results are very similar to the previous ones above. The same relation to *Instruct* type appeared, this time expressed in reverse: *Instruct* moves had significantly *more* BRs than *Query* moves ($\beta = .108, p < .05$). What is new is that *Query* moves did not appear to have more BRs than any other move type. That is, it was *Instructs*, all else being equal, which attracted BRs, not *Queries*. Moreover, *Instructs* differed from *Queries* but *Acknowledges* did not.

In both analyses *move length* is by far the best predictor of the number of BRs, whereas *move type* and *discourse position* contribute much less to explaining the variance in that dependent variable. Nevertheless, their contribution is significant. This was confirmed by the general linear test statistic. In this approach (e.g. see Neter et al., 1996) the individual contribution of a predictor can be assessed by comparing a full model, containing all the predictor variables, with a reduced model in which the predictor under evaluation is excluded. So first we fit a full model with all predictors and obtain the error sum of squares ($SSE(F)$), that is, the sum of the squared deviations of each observation Y_i around its estimated

expected value. Then we fit a reduced model with the predictor of interest excluded and we obtain the error sum of squares ($SSE(R)$). Finally, the following test statistic is applied to compare the two error sums of squares:

$$F = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

where df_R and df_F are the degrees of freedom associated with the reduced and the full model error sums of squares, respectively. The significance of $F_{(df_R - df_F, df_F)}$ is then assessed. This test confirmed the significance of the increase in R^2 when adding the variable *move type* to a reduced model without it ($F_{(3,673)} = 2.920, p < .05$), and when adding the predictor *discourse position* to a reduced model without it ($F_{(1,673)} = 7.612, p < .001$).

4.3.3 *Total brow raise duration*

The second set of hypotheses made predictions about the duration of eyebrow raising per move. Multiple regression analyses were carried out, with the above predictor variables entered in the same way, to assess how much variance in this duration could be explained by those predictors. The resulting model and table of significant predictors are reported below, first comparing move types to *Instruct* type (Table 4.6).

$R = .521$, $R^2 = .272$
$F_{(6,675)} = 41.940$, $p < .001$

Predictor	β	Sig.
Acknowledge	$-.097$	.027
Reply	$-.091$	.032
Query	$-.088$	.017
Explain	$-.123$	$< .001$
Move length	.430	$< .001$

Table 4.6: Independent contribution of the significant predictors of *Total BR duration* per move (move types compared to *Instruct*)

Here the resulting model accounts for 27% of the variance in the total BR duration per move. From the independent contributions of each significant predictor

in Table 4.6 we can see that the best predictor again is *move length* ($\beta = .430$, $p < .001$). And the next is move type, where the reference *Instruct* had significantly longer BRs than any other move type, even when move length was controlled, in the following order of strength: *Explain* ($\beta = -.123, p < .001$), *Acknowledge* ($\beta = -.097, p < .05$), *Reply* ($\beta = -.091, p < .05$), and *Query* ($\beta = -.088$, $p < .05$). Discourse position and speaker identity were not significant predictors of BR duration per move.

Another test was done again using *Query* as the reference group for the other move types in order to assess its predictive power. The results are reported below.

$R = .519, R^2 = .270$
$F_{(5,676)} = 49.970, p < .001$

Predictor	β	Sig.
Instruct	.163	$< .001$
Move length	.421	$< .001$

Table 4.7: Independent contribution of the significant predictors of *Total BR duration* per move (move types compared to *Query*)

Again results show that *Query* moves did not have longer eyebrow raising than other move types, and in fact had significantly shorter brow raising than *Instruct* moves ($\beta = .163$, $p < .001$), even when move length was taken into account. Also, *move length* was again the best predictor ($\beta = .421, p < .001$).

The general linear test statistic was applied as described above (4.3.2) to confirm the significant contribution of *move type* to the model. And again, *move type* was found to contribute significantly to the explanation of the variance in the duration of brow raising ($F_{(2,675)} = 3.751, p < .05$).

4.3.4 Multicollinearity diagnostics

A common problem in multiple regression analysis is the presence of multicollinearity, a strong correlation between two or more predictor variables which can have important consequences for the interpretation and use of a fitted regression model (see Neter et al., 1996; Field, 2005). Multicollinearity can limit the

size of R because several predictors are accounting for the same variance. It can also result in individual regression coefficients being statistically not significant even though a definite relation exists between those predictor variables and the dependent variable. Also, those coefficients become unstable and it may be difficult to assess the relative importance of the predictors. As is often the case in the social sciences, in the current study there was some correlation between some of the predictors. Formal and informal diagnostics were used to check whether there was multicollinearity between those predictors.

An informal method to detect the presence of multicollinearity is to inspect the correlations between the predictor variables and look for very high correlations (greater than .80) between them. Bivariate correlations are reported below in Table 4.8 (point biserial correlation for pairs of interval and nominal dichotomous variables) and in Table 4.9 (Phi coefficient for pairs of dichotomous variables). All the correlations (or associations) except one are statistically significant. But it is the strength of the correlation, not its significance, what may indicate whether multicollinearity may be biasing the result of the multiple regression analysis. The strongest correlation was between *move length* and *Instruct*. This is not surprising since, as we saw earlier (Table 4.2), *Instruct* was the longest type of move in the data. But the correlation is not too strong, and there is no suggestion of multicollinearity. To confirm this, further diagnostics are reported below.

		move length	Sig.(2-tailed)
r_{pb}	*Instruct*	.543	.000
	Trans. initial	.287	.000
	Speaker A1	.125	.001

Table 4.8: Point biserial correlation between *move length* and *Instruct, Trans. initial* and *speaker A1*

		Instruct	*Trans. initial*	*Speaker A1*
Phi coeff.	*Instruct*	1.000	.204	.075
	Trans. initial	.287	1.000	.009
	Speaker A1	.075	.009	1.000
Sig.(1-tailed)	*Instruct*	–	.000	.025
	Trans. initial	.000	–	.412
	Speaker A1	.025	.412	–

Table 4.9: Association (Phi coeff.) between *Instruct, Trans. initial* and *speaker A1*

A formal method to diagnose multicollinearity between predictors is the variance inflation factor (VIF) (Neter et al., 1996, pp. 408–411). The VIFs measure how much the variances of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related. The largest VIF value among all predictor variables is used as an indicator of the severity of multicollinearity. A maximum VIF value in excess of 10 is often taken as an indication that there is a problem of multicollinearity (Neter et al., 1996), though some researchers suggest a value of 4 as a cutoff point to determine serious multicollinearity (Miles and Shevlin, 2001). Another closely related diagnostic is the tolerance value (equivalent to 1/VIF). This value varies between 0 and 1, where 0 indicates perfect collinearity. Here, an arbitrary cutoff point often used is .1 (Myers, 1990), that is, tolerance values $\leq .1$ would indicate multicollinearity between predictors. Some researchers suggest a more conservative cutoff value of .2.

Both the VIF and tolerance values in the regression analyses presented above were acceptable, even considering the more conservative cutoff values just mentioned. Therefore there is no indication that multicollinearity may be causing problems for the interpretation of the regression model. Table 4.10 (for the analysis with *Number of BRs* as dependent variable) and Table 4.11 (with *BR duration* as dependent variable) show that both indicators were far from the customary cut-off points in all cases.

Predictor	Tolerance	VIF
Acknowledge	.552	1.811
Query	.775	1.291
Trans. initial	.831	1.203
Speaker A2	.682	1.466
Speaker B2	.671	1.489
Move length	.539	1.855

Table 4.10: Collinearity Statistics: VIF and Tolerance values for the predictors of *Number of BRs*

Predictor	Tolerance	VIF
Acknowledge	.559	1.790
Reply	.605	1.654
Query	.795	1.258
Explain	.896	1.116
Move length	.557	1.794

Table 4.11: Collinearity Statistics: VIF and Tolerance values for the predictors of *BR duration*

4.4 Discussion

In this study Conversational Games Analysis (Carletta et al., 1997) was applied to six Map Task dialogues in order to investigate whether eyebrow raises produced by speakers were related to the structure of the dialogue and to utterance function. One of the predictions was that brow raises would occur more frequently in moves starting conversational transactions and conversational games than in other positions in the structure of the dialogue (H1a). Also, brow raises were predicted to occur more frequently in *Instruct* and *Query* moves than in other types of move (H1b). Similar predictions were made about the duration of eyebrow raising: the first move in a transaction and the first move in a conversational game were predicted to have longer total brow raising than other moves in the dialogue (H2a), and *Instruct* and *Query* moves were also predicted to have longer brow raising than other types of move (H2b). These hypotheses were partially supported by the results of multiple regression analyses.

Brow raises were found to relate most strongly to the length of the utterance. As the number of words in a move increased, so did the number and total duration of brow raises in the move. If this had been the only relationship found, eyebrow raising would have seemed a random phenomenon with simply more opportunities to occur in long utterances. However, other relations appeared that were independent of move length. Supporting H1a partly, speakers raised their eyebrows more frequently in transaction-initial moves than in non-initial moves. This seemed to indicate that they used eyebrow raising when starting a new task-related section of the discourse. This tendency was not present at a lower discourse level: game-initial moves that were not also transaction-initial

did not have more eyebrow raises than non-initial moves. As for utterance function, brow raises were also found to occur more frequently in *Instruct* moves than in *Query* and *Acknowledge* moves[4], lending some support to H1b above. *Query* moves had also been predicted to have more eyebrow raises than other move types, but no evidence was found for this relation in the current data. In fact, interestingly, the only relation between *Query* and other move types was, as we just saw, that speakers raised their eyebrows *less* frequently when asking questions than when giving an instruction. Utterance function also explained some small variance in the *duration* of eyebrow raising per move: supporting part of H2b, speakers were found to produce longer eyebrow raising while giving instructions than when asking a question. And again, contrary to the second part of the prediction, their queries were not accompanied by longer brow raising than any other type of move was. Thus, *Instruct*, rather than *Query*, had longer and more frequent eyebrow raising. Finally, in relation to discourse position, duration and frequency of eyebrow raising behaved differently at the highest level of the dialogue structure: against the prediction (H2a), eyebrow raising was not longer in utterances starting new transactions. At the level of conversational games, once again no differences were found between the start of the game and other positions.

Speakers did not differ significantly in terms of how long they raised their eyebrows for during their utterances. However, in terms of *number of brow raises* the speaker identity was a better predictor than either the type of utterance or the position of that utterance in the discourse structure. One speaker produced more eyebrow raises per conversational move than the other two speakers did. Large variability between participants is very often found in this type of research and can actually be a problem for the interpretation of findings. In this study, the influence of one speaker on the frequency of brow raising in the data set was stronger than the influence of the variables we were interested in. However, the reported statistical significance found in the latter (discourse position and utterance function) is still valid, because the contribution of each variable was assessed independently of the contribution made by the others. And so, those variables did influence the frequency of brow raising, even if the influence of the

[4]The other move types, *Explain* and *Reply*, showed the same pattern, i.e. they had less brow raises than *Instruct*. But this tendency did not reach significance

speaker identity was larger. Similarly, the influence of *move length* was statistically controlled when evaluating the other potential predictors. As mentioned above, the number of words in an utterance was always the strongest predictor of the frequency and duration of brow raising produced by the speaker in that utterance. If this had been the only influence found, or this plus the effect of speaker identity, then brow raising would have seem just a random behaviour. But because some linguistic phenomena had an influence on it too independent of the duration of the utterance, even if this influence was much smaller, we can claim with some confidence that eyebrow raises bear some relation to the linguistic message accompanying them. It is important to point out though, that the predictive power of the whole model was not very strong. Putting together the influence of discourse position, utterance function, speaker identity and duration of the utterance, only 25 to 27% of the variance in brow raising was accounted for. And therefore, if the majority of this accounted variance is explained by the duration of the utterance and by the identity of the speaker, then the influence of the type of utterance and its position across the discourse is significant but relatively very small. Being aware of this limitation, we could interpret the results as described below.

Discourse structure

Speakers in the dialogues under investigation raised their eyebrows more frequently in the first utterance of a transaction than elsewhere in the dialogue. In Map Task dialogues, an utterance initiating a transaction marks the start of a new section in the description of the map route. And this new section corresponds with a new segment at the highest level in the structure of the dialogue as described by Carletta et al. (1997). Looking at the research literature, this has some similarities with findings by Chovil (1989, 1991a), who reported that in her recordings speakers' facial displays (often brow raises) sometimes marked the start of a new topic in the conversation. The reported frequency was very small, with only five cases representing 2% of the syntactic displays (see 2.5.2). Also the dialogues in her study were of a different nature than the current corpus and they would have had a different structure. However, the beginning of a new story in those conversations could be compared to the beginning of a new transaction in the current corpus. In both cases the speaker is introducing a high-level segment with a new "theme". In Chovil's data this is a new story or topic in the speaker's

narration, and in the Map Task dialogues described here it is a new portion of the route normally described around a particular landmark in the map.

The finding that brow raises appeared with greater frequency at the beginning than elsewhere in a transaction could also be compared to findings in studies of other body movements, such as head movements and other posture shifts. In two conversations microanalysed by McClave (2000) speakers often changed the orientation of their heads as they switched from indirect to direct speech. McClave concluded that one of the functions of head movements, then, would be to mark the beginning of quotes. The beginning of a quote cannot be directly compared to the beginning of a transaction in the current dialogues, but they both represent a change into a distinct segment within the dialogue structure. In this sense both head movements and eyebrow raises would share a similar discourse function. Cassell et al. (2001) also studied position changes for parts of the body excluding hands and eyes. She found that these posture shifts occurred more frequently and were more energetic at the start of high-level discourse segments than within those segments. Cassell et al. labeled the start of high-level discourse segments at the point in which the speakers started a new task topic from the ones they had been assigned: describing rooms, giving directions, and generating an idea for a project. The start of those segments can again be compared to the start of transactions in this study. Cassell et al. also found that posture shifts were more frequent at a turn boundary than within a turn and they concluded that posture shifts can signal *boundaries* of units.

Marking turn boundaries, in particular their start, was a function suggested for brow raises by Cavé et al. (2002). This would suggest that body movement marked boundaries not only at high-levels in the structure of a conversation. However, in the dialogues investigated in this thesis, eyebrow raising did not mark the start of low-level structure units. Within transactions, utterances at the start of a conversational game were not accompanied by eyebrow raising more frequently than other utterances within the game. This is probably because the change from one game to the next is not as marked as a change from one transaction to another. The start of a transaction clearly introduces a change in the conversation by moving to a new portion of the route. Often the *IG* speaker has just finished conveying the description and instructions on a previous portion of the route, and this has been followed by the participant reproducing the route

on her map. Having negotiated that part, the *IG* now moves on to deal with the next part of the route and starts a new transaction in the conversation. On the other hand, the start of a game marks the initiation of a new purpose in the conversation, such as, to provide some instruction or some information, or to acquire some information from the other participant. When the goal of this initiation has been fulfilled or abandoned in the following move(s), another game may start with a new goal. While this implies a change in the conversation, those games are still linked by a coherent "topic" within the same transaction, that is, they have in common the fact that they discuss or negotiate the same part of the route[5]. Also, unlike transactions, games can be embedded within other games to which they are obviously linked. Thus, marking a change into a new game could potentially interfere with the coherence of the larger unit. This would explain why the participants in the current corpus did not use eyebrow raising at the start of a new game.

The frequency of speakers' eyebrow raises, as we have seen, is affected by whether the speaker is starting a new high-level unit in the dialogue or not. On the contrary, and against our prediction, the *duration* of eyebrow raising did not show any relation with the location across the different discourse levels. This could mean that eyebrow raises in transaction-initial moves are frequent but short, and so the total (cumulative) duration of eyebrow raising is not significantly longer there than in non-initial moves. Alternatively, it could be that those brow raises tend to start earlier than the transaction-initial move, i.e. they may start in the long inter-move interval (IMI) that normally precedes a new transaction. If this is the case, then, following the criteria described in section 4.2, that brow raise would be counted as 'belonging' to that initial move. But because its start preceded the start of the move, the total brow raise duration *within* that move would be reduced.

Considering previous findings in the literature and those here, a general conclusion could be made that a change in body movement can signal a change from one segment of the discourse to another. And more in particular, we could conclude that in the task-oriented dialogues under investigation speakers' eyebrow

[5]A game can of course start the negotiation of a new route section when that game is the first one in a new transaction. But recall that the first move in a game that was also first in a transaction was excluded from the game-initial category

raises seemed to have a discourse function by marking the *start* of *high-level* discourse units in the conversation.

Utterance function

In relation to utterance function, *Instruct* moves had more brow raises than *Acknowledge* and *Query* moves, and longer eyebrow raising than any other type of move. But, contrary to what had been predicted, ***Query* moves** did not have more or longer eyebrow raising than other types. This is an interesting finding because in previous research it has been claimed that eyebrow raises can have a questioning function. Some studies have been only descriptive and have presented their observations without empirical investigation and supporting data (e.g. Birdwhistell, 1970; Eibl-Eibesfeldt, 1972; Ekman, 1979). Chovil (1989, 1991a) and Srinivasan and Massaro (2003) did present empirical data suggesting a possible questioning function for eyebrow raises, but strong claims could not be made. In her inductive study, Chovil claimed that 14% of what she called "syntactic displays" had a question marking function, and that these displays consisted mainly of eyebrow movements. Srinivasan and Massaro (2003) carried out a series of perception experiments to find out whether visual and auditory cues could distinguish echoic questions from statements. They found that eyebrow raising and head tilting could mark the echoic question and distinguish it from its equivalent statement form in synthetic speech. But the effect of these visual characteristics was much weaker than the effect of the auditory cues (F0, duration and amplitude), even when the former were enhanced.

In spite of this lack of strong empirical evidence for the use of eyebrow raises as questions markers, the intuition that we often raise our eyebrows to ask a question seems generally accepted. In this study I investigated whether eyebrow raises would indeed be used in this way in task-oriented dialogues. Interestingly, and contrary to the expectation, the speakers in the six dialogues studied did not use eyebrow raising more often in queries than they did in any other type of utterance. Furthermore their queries were characterised by significantly fewer and shorter eyebrow raising than their instructions. This seems to contradict the findings by Chovil (1989) and Srinivasan and Massaro (2003), but it would agree with the weak effect found by the latter and the fact that their participants relied most strongly on auditory cues when discriminating questions and statements.

The current findings could be interpreted in the following ways. First, it could be that eyebrow raising does not have a questioning function. Or that it is not used with this function in the type of interaction that takes place in Map Task dialogues. Alternatively, it could be that eyebrow raises can add a question meaning to an utterance that would have a different function if only listening to the speaker. In line with this, it has been suggested (Ekman, 1979; Chovil, 1989) that eyebrow raises are more likely to be used as question markers when the syntactic form of the utterance does not make it clear that a question is being asked. In the current investigation, utterances classified as *Query* did not necessarily have the syntactic form of a question. That is, as long as the utterance was perceived by the annotator as a question within its context, then it was classified as such. But as explained in the previous chapter (section 3.3.1), the function of utterances was labeled by only listening to the dialogues and not looking at the speakers' faces. If it was the case that some utterances could only be interpreted as questions by looking at visual cues such as eyebrow raises, then in the current study this interpretation would have been missed. However, considering the report by Srinivasan and Massaro (2003) that visual cues had a much weaker effect than auditory cues in their perception experiments, it seems unlikely that the finding in the current study would be due to wrong assignment of some questions to a category other than *Query*.

As for ***Instruct* moves**, the fact that these had longer and more frequent eyebrow raising than other types of utterance could be due to the importance of these moves in the dialogues. Map Task dialogues, as explained earlier, are driven mainly by the instructions provided by the *IG* speaker and it is mainly these instructions that allow the *IF* to draw the route on her map. Therefore, instructions must be conveyed clearly and efficiently by the speaker in order to succeed in the completion of their task. Eyebrow raises may play a role here by reinforcing the content of these utterances and setting them apart. In a different kind of dialogue, another type of utterance could carry the key information that would be marked by eyebrow raising. Providing emphasis is certainly a role that is intuitively linked to eyebrow raising, and it seems natural that it would be used to a larger extent in utterances carrying the most important information.

Another interpretation, in connection with the discussion above, is that the reason why *Instruct* moves were associated with eyebrow raising was to add a

questioning meaning to the instructions, as if simultaneously asking *'ok?'*, *'are you with me?'*. It is possible that this function, of checking that the interlocutor is following the conversation, can be achieved sometimes by an explicit *Align* move (see Chapter 3 for a description of *Align*), and other times by means of eyebrow raising accompanying the instructions. It would be interesting, in future research, to study the interlocutor's (*IF*) behaviour, to see how many times they produced an *Acknowledge* or *Reply* move immediately following a brow raise by the *IG* speaker in a non-query move, as if the *IF* had felt prompted to provide a reply or a sign that a message had been successfully conveyed.

We have seen that eyebrow raises in the current data did not mark *Query* moves. But it is not possible to say whether they added a questioning meaning to other utterances that were not classified as queries in the dialogues What brow raises did mark were utterances classified as *Instruct*. It is likely that speakers used eyebrow raising to emphasise words or phrases in order to communicate their message in the most efficient way. This idea, which we could call the "emphasis hypothesis", cannot be tested in the current analysis but will be explored in Chapter 5.

To conclude, taking into account the weakness of the reported results and the variability between the speakers, the findings in this study are still of considerable importance. They provide tentative evidence of a relation between eyebrow raising and the linguistic message, which can be interpreted as an indication that eyebrow raises have conversational functions. These functions would be to signal the beginning of high-level discourse segments and to emphasise information in the utterances with the most important function in the dialogue.

CHAPTER 5

Eyebrow raises and pitch accents

5.1 Introduction

In Chapter 4 eyebrow raises were studied in relation to dialogue structure and utterance function and a small tendency was found for them to occur at the start of high-level discourse segments (transactions) and in *Instruct* moves. What we have not yet considered is *where within utterances* brow raises occur. As we saw in Chapter 2, eyebrow raising has been associated with intonation. Much of this research has been of a descriptive nature (see 2.5.1), but there has also been some, though limited, empirical research. Studies with synthetic animations have reported a preference for short eyebrow raises to be aligned with pitch accents in Dutch, where these movements could influence the perception of prominence (Krahmer et al., 2002a; Krahmer and Swerts, 2004). A similar result has been reported for Swedish (House et al., 2001). Brow raises could also contribute to the marking of information in focus (contrastive information) in Dutch (Krahmer et al., 2002b). And in French, brow raises have been reported to occur frequently with accentuating rising pitch contours (Cavé et al., 1996, 2002).

The purpose of the study in this chapter was to investigate a possible alignment between eyebrow raises (regardless of their length) and pitch accents in English. Since pitch accents have roles in discourse, like marking new information or achieving contrastive focus, such an alignment might be the mechanism whereby brow raises play a linguistic role. On the basis of the existing literature, we might hypothesise that:

H1: A brow raise will occur in alignment with a pitch accent

H2: The properties of some but not other pitch accents will attract brow raises. Attractor accents might be distinguished from non-attractors in terms of:

a) Information structure

If brow raises mark new information (see 2.3.3), pitch accents on first mentions of map landmarks should attract brow raises proportionally more often than accents on second mentions.

b) Position in move

Attractor pitch accents will tend to occur later in the dialogue move than non-attractors. The hypothesis here is based on the fact that new and important information tends to appear late in the utterance. This prediction is related to information structure as well, but in contrast with a) above, it is not restricted to referring expressions and affects the whole utterance.

c) Pitch accent type

Some types of accents will attract brow raises more frequently than others. Two different predictions were made here:

1. Primary accents will attract brow raises more frequently than secondary accents, which are weaker than the former.
2. In a group of accents affected by downstep (see 2.3.2), the first accent will behave as attractor more frequently than the following descending accents in the group.

As mentioned in 3.3.2, the final accent in a downstep group is a primary type accent, and the preceding ones are secondary. Thus, in downstep groups, 1. and 2. would predict different relations between accent type and brow raising. The motivation was purely exploratory to find whether (1.) brow raising would relate to a contrast between weak/strong intonational events, and (2.) in downstep groups it would be associated to prosodic structure.

d) Move type

There will be proportionally more attractors in *Instruct* moves than in other types of move. This prediction is different from a), b), and c) in the sense that it is not about where *within* an utterance brow raises occur. It was motivated by results from Chapter 4 suggesting that *Instruct* moves tend to have more brow raises than other types of

move. There, it was hypothesised that brow raises emphasise the key information presented in instructions. Alignment with pitch accents in these moves would provide some support for this hypothesis.

H0: Contrary to the two hypotheses stated above, if brow raises were not related to pitch accents, then they would not be temporally aligned with them, and their nearest accent would not have different properties from other accents in the dialogue.

5.2 Method

5.2.1 Materials

The materials in this study came from the same dialogues used in Chapter 4. The annotation scheme and procedure for brow raises, pitch accents, and information structure were fully described in Chapter 3.

Eyebrow raises

The start and end times of all brow raises were recorded on the dialogue timeline. There were a total of 274 brow raises (mean duration $= 1.47sec$; $s.d. = 1.61$) . Three of them, with a duration longer than $8sec$, were excluded from the analysis, leaving 271 brow raises with a mean duration of $1.379sec$ ($s.d. = 1.321$). For every brow raise, associated features from the move in which it occurred were also coded.

Pitch accents

There was a total of 1893 pitch accents annotated on the dialogues. Those in or around the three unusually long brow raises mentioned above (longer than $8sec$) were excluded. This left 1858 pitch accents for the analysis. To test the first hypothesis, it was necessary to know the location of PAs in the dialogues. Pitch accent codes from the dialogue transcripts of the *IG* speakers were recorded as codes for portions of the digitised speech signal, as described in 3.3.2. To test the second hypothesis, and its subset of predictions, the following features were included in the pitch accents data set:

a) Information status of referring expressions
   Pitch accents on mentions of map landmarks were labeled as occurring on a first or second mention as described in 3.3.3. First mentions were considered *new* information, and second mentions were considered *given* information[1].
b) Position in move
   Each move was divided into four equal quarters of its length in seconds, and pitch accents were labeled as occurring on the first, second, third, or fourth quarter of a move.
c) Pitch accent type
   Each pitch accent was classified into one of five categories: *primary, secondary, downstep-initial, downstep-medial,* or *downstep-final.*
d) Move type
   A note was made of the type of move in which the pitch accents occurred.

5.2.2 *Statistical analysis*

Alignment between eyebrow raises and pitch accents

Brow raises may be aligned with a previous or with a following pitch accent. In any case, perfect synchrony is not expected since it would be physically impossible for speakers to produce and technically difficult for a researcher to measure. In order to decide whether brow raises and pitch accents occurred close enough to consider them aligned, a frequency distribution was plotted for the distance between brow raises and their nearest accent. The reference points considered to calculate that distance were the start of the events. For brow raises, as explained in 4.2.1 above, their start was regarded as more significant than their end, and was also perceptually more salient. For pitch accents, the start of the accented syllable was chosen on the basis of findings by Ladd and Schepman (2003), who showed that the pitch excursion normally begins very close to the start of the syllable. For simplicity, I will generally refer to 'brow raise' (BR) and 'pitch accent' (PA), rather than '*start* of brow raise' and '*start* of pitch accent'.

[1]Some first mentions did not have a second mention. Also, second mentions were generally accented but there were six instances (out of fifty-seven) that were unaccented. In these cases, the next accented mention was considered for the analysis. For simplicity and for the purpose of the analysis these are also referred to as *second mentions*

The distances, or delays, plotted for the frequency distribution can be better explained with reference to the following symbolic representation:

–PA1————————-BR——PA2–

where the horizontal line represents the timeline of the events, PA1 is the accent preceding the start of the BR, and PA2 is the one following the start of the BR. In this case, the closest accent to the BR is PA2, and so the distance plotted in the distribution graph would be the distance from BR to PA2. Evidence of alignment would be provided by a distribution peaking at zero.

If PAs always occurred very close together, then BRs during speech would necessarily start very close to a PA even if they were randomly distributed. To evaluate whether a potential alignment between the two events was simply forced by short distances between PAs, a second distribution was graphed of the distances between the two consecutive accents surrounding the start of BRs, i.e. between PA1 and PA2 in the representation above. The question of whether the BR occurred significantly closer to one of the two surrounding accents was addressed with a pairwise t-test (two-tailed) comparing the mean distance between PA1 and BR with that between BR and PA2. A significant difference between the means would provide some evidence that the BR did not start just randomly between two accents and tended to occur closer to one of them. Additionally, to evaluate possible large subject variability, a one-way ANOVA was performed to test the hypothesis that the mean distance from BR to PA2 differed between subjects.

Finally, in order to show whether the pattern of alignment between BRs and PAs would vary depending on the duration of the BR, separate frequency distributions of that distance were made for *short* and *long* BRs. The criteria used to select BRs as short/long, was as follows: ordered in terms of their duration, the first quartile of the BR data set ($\leq 0.4sec$ long) and the fourth quartile ($\geq 1.92sec$) were classified as *short* versus *long* BRs, respectively.

Properties of pitch accents attracting eyebrow raises

The second question addressed in this study was whether there is a subset of PAs with special properties that *attract* BRs. The properties considered were in terms

of:

a) Information structure
b) Position in move
c) Pitch accent type
d) Move type

Pitch accents were classified as *attractors* and *non-attractors*, according to whether they were the nearest PA to a BR or not, respectively. The ratio of PA attractors to non-attractors was compared by means of Chi-Square tests in the groups below (possible interactions were also evaluated):

a) *First mentions* vs. *second mentions* of map landmarks
b) Each quarter of the move length vs. one another
c) *Primary* vs. *secondary*, and *downstep-initial* vs. *non-initial*
d) *Instruct* vs. all other move types (grouped as *non-instruct* type)

An important clarification must be made at this point about the issue of independence in the data used here. The size of the data set was quite large and each item contributed to only one cell of the contingency table in the Chi-Square statistic. However, because the items were collected from only three participants, the data within cells would not be independent of speaker. One participant might have contributed a very large number of pitch accents for one cell, and this might influence the measure of the relationship between the variables. This can be a serious problem for the interpretation of the results and should be avoided by collecting more data from a larger number of participants. With this clarified, the results are presented below and will be interpreted as suggestive of a pattern that might be confirmed if more data from more participants were collected.

5.3 Results

5.3.1 Alignment between eyebrow raises and pitch accents

Figure 5.1 shows the distance in seconds between the start of a BR and the start of the nearest PA (preceding or following). Zero on the X axis represents the start of the PA, and the Y axis shows the frequency of distances from BRs to

PAs (negative values), or from PAs to BRs (positive values)[2]. The mean distance between the two events is $-0.063sec$ ($s.d. = 0.458$) and, as the graph shows, delays from PA to BR or BR to PA cluster around zero. This is the pattern which was expected and which provides some evidence of alignment.

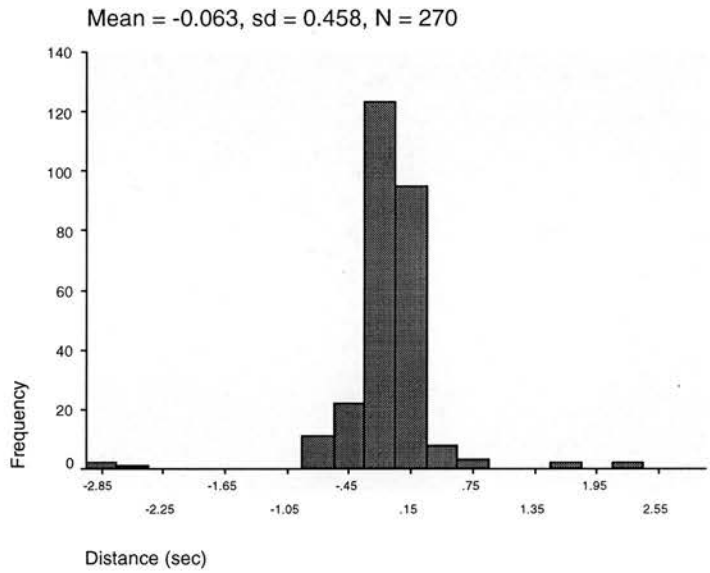

Figure 5.1: Distance between BRs and nearest PA

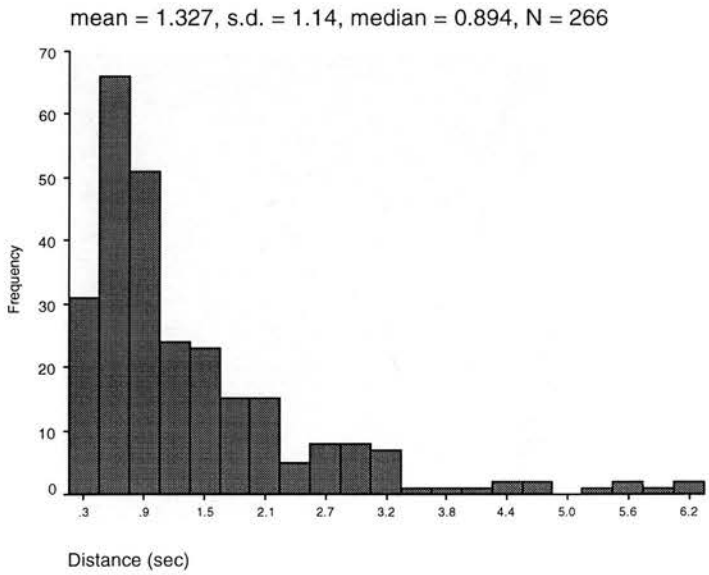

Figure 5.2: Distance between two PAs surrounding the BR start

[2]Comparing this to the symbolic representation in section 5.2.2, negative values in Figure 5.1 correspond to the distance between BR and PA2 and positive values correspond to the distance between PA1 and BR

In Figure 5.2 we can see the frequency distribution of distances between two PAs: the one preceding and the one following a BR start. For two BRs in the data set that distance could not be calculated, because one occurred at the start of the dialogue and did not have a preceding PA and the other one was at the end of the dialogue and did not have a following PA. Two additional extreme values (greater than $8sec$) were excluded. These corresponded to distances between PAs which belonged to different turns. The distribution is rather skewed, as expected (mean $= 1.327sec$; $s.d. = 1.14$; median$= 0.894$).

A pairwise t-test (two-tailed) comparing the mean distance between a BR and its preceding PA with the mean distance between that BR and its following PA showed a significant difference ($t = 2.381, df = 271, p < 0.05$). BRs are significantly closer to their following PA than to their preceding PA[3]. A one-way ANOVA evaluated the effect of the participants' identity on how close the BR started to its following PA. No significant differences were found between the participants ($F_{(2,270)} = 0.552, p = .57$).

Next, Figure 5.3 presents the frequency distribution of distances between short BRs ($\leq 0.4sec$) and their nearest PA. Figure 5.4 shows the distribution for long BRs ($\geq 1.92sec$). These graphs have the same format as the one in Figure 5.1 and show a similar distribution shape. Notice that the average distance to nearest PA is smaller for short BRs (mean$= -0.024sec$, $s.d. = 0.491$) than for long ones (mean $= -0.126sec$, $s.d. = 0.508$).

5.3.2 *Properties of pitch accents attracting eyebrow raises*

As explained above, attractor PAs were those nearest to a BR, and non-attractors were all other PAs. A comparison of their properties was made to investigate what may be attracting BRs. Chi-Squares comparing the ratio of attractor to non-attractor PAs in the groups below gave the following results (with Yates' correction factor).

[3]The difference between means was $0.203sec$ and the 95% confidence interval was between $0.035sec$ and $0.370sec$

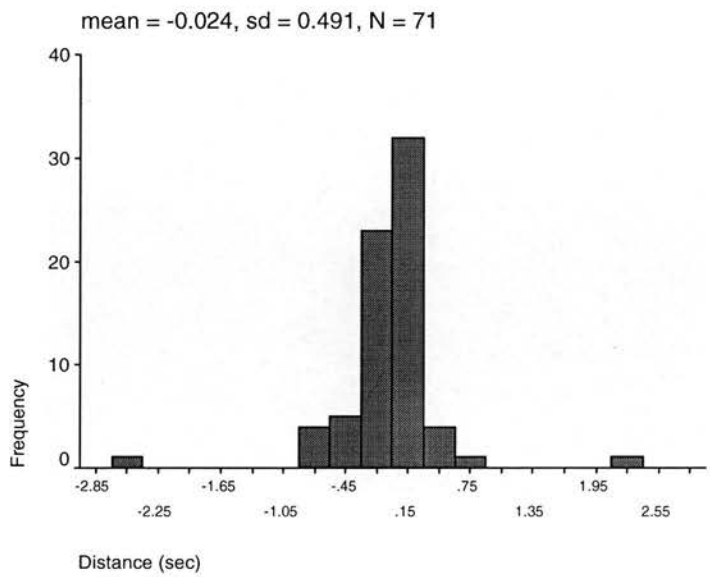

Figure 5.3: Distance between short BRs and nearest PA

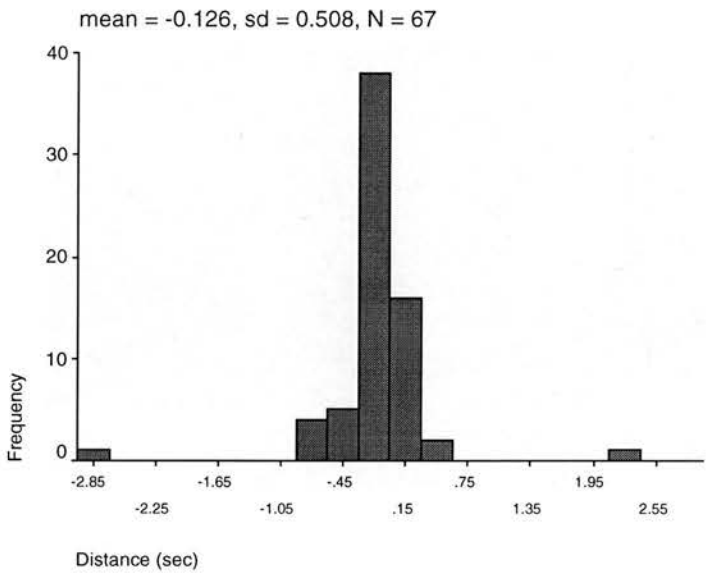

Figure 5.4: Distance between long BRs and nearest PA

a) Information structure

Accented first and second mentions of map landmarks showed no significant difference in the ratio of attractor to non-attractor PAs (Table 5.1). The percentage of attractors versus non-attractors in first mentions was higher than in second mentions, as predicted, but this difference was far from being significant.

	1st mention	2nd mention
non-attractor	53(80.3%)	47(82.5%)
attractor	13(19.7%)	10(17.5%)
Total	66(100%)	57(100%)

Table 5.1: Frequency of attractor/non-attractor PAs in *first/second* mentions

b) Position in move

No significant difference was found among PAs in different quarters of the move length (Table 5.2).

	First quart.	Second quart.	Third quart.	Fourth quart.
non-attractor	508(86.5%)	405(86.5%)	385(84.6%)	296(85.1%)
attractor	79(13.5%)	63(13.5%)	70(15.4%)	52(14.9%)
Total	587(100%)	468(100%)	455(100%)	348(100%)

Table 5.2: Frequency of attractor/non-attractor PAs across move length quartiles

c) Pitch accent type

There was no significant difference between primary and secondary accent types (Table 5.3). In downstep groups, however, the position of the accent in the group (initial vs. non-initial) was associated with the accent being an attractor/non-attractor accent. Downstep-initial accents attracted BRs more often than non-initials did ($\chi^2 = 5.34$, $df = 1$, $N = 525$, $p = 0.021$). See Table 5.4 and Figure 5.5.

	Primary	Secondary
non-attractor	1062(85.6%)	80(86%)
attractor	178(14.4%)	13(14%)
Total	1240(100%)	93(100%)

Table 5.3: Frequency of attractor/non-attractor PAs in *primary/secondary* type

	D. initial	D. non-initial
non-attractor	191(82%)	261(89.4%)
attractor	42(18%)	31(10.6%)
Total	233(100%)	292(100%)

Table 5.4: Frequency of attractor/non-attractor PAs in *downstep initial* vs *non-initial* position

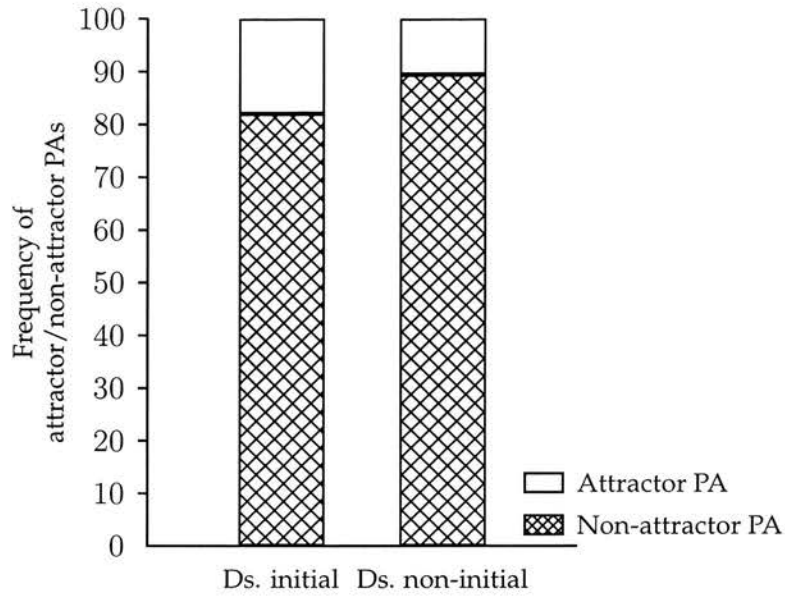

Figure 5.5: Ratio attractor/non-attractor in *downstep initial* vs *non-initial* PAs

d) Move type

The ratio of attractors to non-attractors was significantly higher in *Instruct* moves than in other move types (non-instruct), as we might expect ($\chi^2 = 6.5$, $df = 1$, $N = 1858$, $p = 0.011$). See Table 5.5 and Figure 5.6.

	Instruct mv.	Non-instruct mv.
non-attractor	939(84.1%)	655(88.4%)
attractor	178(15.9%)	86(11.6%)
Total	1117(100%)	741(100%)

Table 5.5: Frequency of attractor/non-attractor PAs in Instruct vs non-instruct moves

No interactions were found between all the groups compared.

5.4 Discussion

In Chapter 4 brow raises were analysed with respect to dialogue structure and utterance function. In the current chapter they were studied in relation to intonational events, namely pitch accents, to investigate possible linguistic functions of brow raises at this level. The approach was to first find out if there was temporal coordination between the brow raises and the accents in the dialogues. Once this

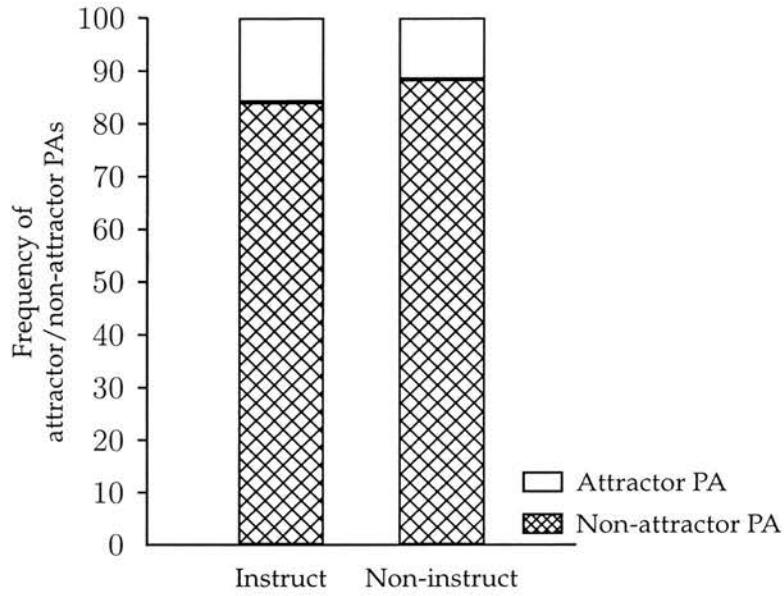

Figure 5.6: Ratio attractor/non-attractor PAs in *Instruct* vs. *Non-instruct* moves

was established, the second step was to look at properties of the accents coordinated with the eyebrow raises in an attempt to determine what may be causing this temporal association and what we may infer from it.

5.4.1 *Alignment between brow raises and pitch accents*

Are eyebrow raises aligned with pitch accents, as the first hypothesis predicted? Figure 5.1 shows that brow raises did occur remarkably close to an accented syllable. Eighty-seven percent of the brow raises started less than $0.350sec$ away from the nearest accent, and in total, brow raises started an average of $0.063sec$ earlier than the start of the accent ($s.d. = 0.458$). This provided some support to the alignment hypothesis. To assess the relevance of this alignment, it was necessary to investigate whether it was simply forced by short distances between pitch accents. If this was the case the alignment would not necessarily imply a relation between the events, since there simply was no room for longer distances between them. Figure 5.2 shows this was not the case. The mean distance between the pitch accent preceding and the one following the start of the brow raise was $1.327sec$ ($s.d. = 1.14$). In other words, between the two accents there was an average window of $1.327sec$ in which the brow raise onset occurred. This was wide enough for the brow raise to start further away from any accent than the average distance found between them in Figure 5.1, and it suggested a true

relation between the two aligned events. Indeed, statistical evidence was found to support this: brow raises began significantly closer to one of the two pitch accents around its onset than they would have done if they simply occurred at random ($t = 2.381, p < .05$). And interestingly, they occurred closer to the *following* accent than to the preceding one. To the best of my knowledge, this is the first study that provides evidence of alignment between brow raises and pitch accents in English.

To characterise the alignment further, the distance between brow raises and pitch accents was compared for short versus long brow raises. Figures 5.3 and 5.4 suggest that although the mean distance to the nearest accent was closer for short brow raises ($-0.024sec$, $s.d. = 0.491$) than for long ones ($-0.126sec$, $s.d = 0.508$), both groups appeared aligned with an accent and tended to precede this accent. This is an important contribution to the literature, especially because previous studies in other languages concentrated mainly on short brow raises.

Several works in the research literature have suggested that brow raises are associated with intonation, and with pitch accents in particular (see section 2.5). But very few empirical studies have looked at the temporal relation between brow raises and pitch accents, and those which have did not study this relation in English but in other languages. In a perception study with synthetic stimuli in Dutch (Krahmer et al., 2002a), participants preferred animations in which a short brow raise ($0.300sec$ long) and a pitch accent were synchronised on the same word in a two-word phrase (other findings from similar perception studies will be referred to further below). This could suggest that in natural conversations speakers raise their eyebrows in alignment with an accented word. However, a strong conclusion about natural production cannot be derived from results on such short synthetic phrases. As the authors pointed out, investigations on natural interactions with real speakers are necessary to gain more insight into this. Some research on natural conversations has been done in French. Brow raises there seemed to occur most often with accentuating intonation contours (mostly "rising-falling") than with non-accentuating contours (Cavé et al., 1996, 2002, see 2.5.2 above). But the authors did not report what exactly the temporal relation was between the brow raises and these contours, or how they determined co-occurrence.

The findings in this chapter make an important contribution to this area by providing evidence of alignment between natural brow raises and pitch accents in English and also a measure of the temporal characteristics of this alignment. The results were based on careful analysis of natural dialogues. The sample data size was reasonably large and was not restricted to short brow raises. Furthermore, care was taken to evaluate the possibility that this alignment was due to a narrow temporal spacing between accents in the dialogues which forced the brow raise to always occur very close to one of them. This possibility was ruled out (see Figure 5.2 and reported t-test), and so we can conclude that the alignment between brow raises and pitch accents in the data was not random and may serve some linguistic purpose. A caveat of the sample data is that it only came from three participants, which means that generalisations cannot be made. As for variability between the participants, a comparison of the mean distance from brow raises to their following accent for each participant did not result in significant differences. Thus, there was no evidence of large differences between them in the general pattern of alignment found in the data set.

Regarding the characteristics of the alignment found, the tendency for the start of the brow raise to precede the start of the accented syllable is noteworthy. This finding agrees with several observations in the literature of body movement, in which researchers have reported a tendency for the movement to precede the word(s) with which it is associated. This was observed by Kendon (1972, 1980) and later by McNeill (1992) and was termed by the latter the "phonological synchrony rule": the stroke phase of a gesture (the moment of most accented movement) is completed before or at the accented syllable of the accompanying speech. Loehr (2004) found very good alignment between hand gestures (in particular the apex of the stroke) and the nearest accent, with the hand gesture generally preceding the pitch accent. He reported a mean delay of $0.017sec$ from the gesture to the accent, which is remarkably short. In the current study, there was a slightly longer delay from the brow raise to the accent: on average, brow raises started $0.063sec$ earlier than the accent. But the similarity in the pattern of alignment, in which the movement precedes the intonational event, is striking, especially since hand movements and eyebrow movements are very different in shape and magnitude.

For eyebrow movements in particular, associations with intonational events have been made, but details of a temporal relation between the two have not been reported for natural interactions. Manipulations with synthetic stimuli in perception experiments can provide some information about what timing relations seem more natural to human observers. For instance, House et al. (2001) reported a preference for both head movements and eyebrow raises to be synchronised with an accented word in Swedish, but also that perfect synchrony was not necessary for the integration of the visual and auditory stimuli (perceptual sensitivity to timing was in the order of $100ms.$). Other findings in Dutch would seem to agree with the trend in the current thesis for eyebrow raises to precede their nearest accent. Krahmer et al. (2002a) and Krahmer and Swerts (2004) reported that when both words carried a pitch accent in their synthetic stimuli in Dutch, participants preference for a single brow raise was on the first word. In the light of the current findings their preference might be interpreted as follows: a brow movement on the second word would have immediately followed the first accent, an this sequence may not be as natural as a brow raise on the first accent and preceding the second one. Furthermore, when only one of the two words was accented participants preferred the brow raise to be synchronized with the accent, and this preference was clearest when the accent fell on the first word (75% of the cases) than when it fell on the second one (62%). Again, the latter could reflect a tendency for natural eyebrow raises to start before an associated accented word, which would have made the sequence 'brow – accent' somewhat more acceptable than the reverse, 'accent – brow'. Another perhaps more likely interpretation of participants' preferences in those studies would be that brow raises could mark the start of linguistic units, and so they were preferred on the first word of the short stimuli phrases. This possible function of eyebrow raising will be mentioned again below, when interpreting the results found in relation to groups of accents linked by downstep.

5.4.2 Properties of pitch accents attracting eyebrow raises

In the investigation of the question of alignment in this chapter, the first hypothesis was supported: brow raises occurred in alignment with a pitch accent. But obviously there are many more accents than there are brow raises. So, brow raises were associated with an accent, but not all accents were associated with a

brow raise. We can speculate that those accents, or their context, *attracted* brow raises and this alignment served some linguistic purpose. Thus, by exploring differences between accents that are aligned with brow raises and those that are not, we might find an explanation of what linguistic function, if any, brow raising may have in dialogue. The second hypothesis, then, predicted different properties between attractor and non-attractor accents (i.e. between accents nearest to a brow raise and others, respectively). Differences were predicted in terms of: *a)* information status of referring expressions, *b)* position of the accent along a conversational move, *c)* type of pitch accent, and *d)* type of move. Groups of pitch accents classified in terms of *a), b), c)*, and *d)* above were compared for their ratio of *attraction* versus *non-attraction* of a brow raise, where *attraction* means that a brow raise started next to them. Analyses using the Chi-Square statistic were carried out to test whether the differences in those ratios between the groups were statistically significant. However, as explained above, in spite of the relatively large number of items in the analyses, there were only three participants from which these items were collected, which means that the data was not independent of speaker. Thus, strong conclusions cannot be made. Nevertheless, the results are suggestive and illustrate a pattern that might be confirmed if more data from a larger number of participants were collected.

Of all the comparisons made between the groups, only two showed differences. The first one was related to groups of accents affected by **downstep**. The first accent in a downstep group attracted brow raises more often, proportionally, than the rest of the accents in the group (see Figure 5.5). If confirmed with more data, this would be a new finding. As explained in section 2.3.2, downstep is a phenomenon affecting a series of two or more similar accents (usually High tones) where F0 is gradually lowered from one to the next to an extent that cannot be accounted for by background declination. This group of accents stands out as a cohesive intonation unit. The result of the analysis here could suggest that an eyebrow raise may signal the start of that prosodic unit, and that the brow raise is unlikely to start somewhere else in the group perhaps because it would interfere with the prosodic cohesion of that unit. We could speculate then that eyebrow raising is related to prosodic structure. This would fit with previous observations that claimed a close relationship between the prosodic structure and the structure of body movement, especially of hand movements (e.g. Kendon, 1972, 1980). In

the current thesis, only the start of the brow raise was considered. In future research it would be interesting to find out whether the end of the brow raise and the end of the downstep group are also aligned, which would indicate a stronger association between the two events and a stronger parallel with earlier observations on body movement in general.

As we said earlier, exploring relations between eyebrow raising and the verbal channel may bring some insight into possible linguistic functions of eyebrow movements. If the downstepped accents conveyed a meaning of finality in the utterances, then they would have a discourse function, and if eyebrow raising was associated with these accents it may share this function. The function or meaning of the phenomenon of downstep have not been widely investigated. Perhaps further research into eyebrow raising in the dialogue and in these groups of accents in particular, might shed some light on this as well. A more general function that could be attributed to eyebrow raising, and that connects with findings from Chapter 4, is marking the start of coherent units in the linguistic structure (prosodic structure, in the current chapter, and discourse structure in the study in Chapter 4).

In relation to linguistic structure, contrary to what had been predicted, accented first and second mentions of map landmarks did not show a significant difference in the frequency with which they attracted brow raises (see Table 5.1). This indicates that speakers did not use eyebrow raising when referring to a map landmark as *new* information, as opposed to *given* information. Thus, the observation that some hand gestures (beats) are often used when introducing new entities into the discourse (McNeill, 1992) was not supported for eyebrow raises in the current corpus. In relation to information structure, another contrast may appear at sentence level between two parts of an utterance that have been referred to as "theme" and "rheme" (Halliday, 1967a). In simple terms, the "rheme" is the part of the utterance that adds something new to the "theme", which is the part that connects with the previous discourse and specifies what the utterance is about. Eye movements have been related to the thematic structure of an utterance in previous research. Cassell et al. (1999) found that speakers gazed at their interlocutor at the beginning of the rheme, whereas at the beginning of the theme they looked away. The theme/rheme structure of utterances was not annotated in the current data and could not be analysed. A related comparison

was made between accents across the four quartiles of the utterance length in seconds. This was motivated by the idea that the most relevant information in an utterance tends to appear at the end. But again, there was no significant difference across those sections of the utterance, even when only the first and last quartile were compared: brow raises did not align with accents at the end of the utterance more frequently than with earlier accents. This is another indication that in the dialogues under investigation, speakers did not seem to mark contrasts in information structure by means of eyebrow raising. Impressionistically, however, it seems as if eyebrow raising does contribute to establish some kind of contrast between units of the linguistic message. Experiments with synthetic animations in Dutch (Krahmer et al., 2002b, see 2.5.3 above) have shown that brow raises could mark information in focus (contrastive information) in two-word synthetic phrases, although pitch accents had a much greater effect. It could be the case that in the Map Task dialogues in this thesis, a contrast other than the ones explored above was visually marked. Preliminary observations suggest that eyebrow raises sometimes accompanied words of direction or location, perhaps marking contrasts such as *up/down, right/left, above/below*. Further investigation is required here.

Another test that did not show differences between groups of accents was the comparison between primary and secondary accents (see Table 5.3). As a purely exploratory hypothesis, brow raises had been predicted to occur more frequently in alignment with a primary accent than with a secondary one. But results showed that eyebrow raising did not relate to the strong/weak phonological relation between these accents.

A second comparison that suggested an important difference between groups had to do with the type of move in which the accents occurred. As expected, pitch accents in ***Instruct*** moves seemed to attract brow raises more often than accents in other types of move (see Figure 5.6). This provided some support to a previous finding in Chapter 4, where *Instruct* moves appeared to have more brow raises than at least *Query* and *Acknowledge* types. As explained in that chapter, in Map Task dialogues *Instruct* moves contain the most important information to advance the dialogue and complete the task. Therefore they must be presented efficiently, highlighting the important bits of information within the utterance. This can be achieved by making certain words prominent. Pitch

accents alone can provide prominence, but when more emphasis is needed, it might be necessary to reinforce a word in some other way as well. Eyebrow raising may be one of the mechanisms that can achieve this purpose by adding to the salience of words. This idea seems to be supported by previous findings in the research literature. For instance, eyebrow raising was found to cue prominence in a Swedish animated talking head (House et al., 2001). Similarly, in synthetic animations in Dutch, brow raises boosted the perceived prominence of accented words and scaled down that of unaccented ones (Krahmer et al., 2002a). And in a more recent production test (Krahmer and Swerts, 2004) some speakers of Dutch spontaneously used eyebrow raising when reading words in which they had to emphasise a specific syllable.

To summarise, the analysis of six Map Task dialogues in this study showed that when speakers raised their eyebrows they did so in alignment with a word carrying a pitch accent. Specifically, the eyebrow raise started an average of $0.063sec$ earlier than the accented syllable, resembling the pattern found in studies of body movement where the movement preceded the associated speech. This pattern appeared to persist in both short and long brow raises, though the alignment was slightly closer for the former, and across participants. It was hypothesised that this alignment resulted from shared linguistic functions of the two events. From the different possibilities explored, eyebrow raises were not associated with strong versus weaker accents, and they did not mark contrasts in information structure. There was an indication, however, that brow raising may have had some linguistic function. First, they occurred more frequently at the start of groups of accents affected by the phenomenon of downstep than later in the group. Second, they were more frequently aligned with accents in instructions than with those in other utterances. From this we could speculate that eyebrow raising had some prosodic function in the dialogues: namely, signalling the start of certain segments in the prosodic structure, and providing emphasis to words.

In conclusion, some association was suggested here between verbal and non-verbal behaviours in dialogue. The most important finding was that when speakers raised their eyebrows they always did so in close alignment with an accented syllable in their speech. It was speculated that eyebrow raising had prosodic

functions, such as marking the start of certain prosodic units and emphasising information. These results encourage further research including a larger number of subjects to confirm these suggestions and explore other possible relations.

CHAPTER 6

General discussion and conclusions

6.1 Introduction

Eyebrow raising, like many human behaviours, is something we do mostly without conscious control of it. However, it is clearly not a random phenomenon and seems to be linked to other behaviours. For instance, the fact that we raise our eyebrows in a reaction of surprise indicates that eyebrow raising is related in some way to certain emotions. The fact that we also raise our eyebrows as we are talking and not necessarily feeling a particular emotion indicates a relation with communication and with speech. As we saw in Chapter 2, the connection between facial expression and emotion has received a lot of attention. On the other hand, we still know comparably little about how eyebrow raises may relate to speech. Yet, we use spoken language as a communication tool in daily life very frequently, and on most occasions what we are communicating is not emotional states. Thus, there would be important advantages in a better understanding of how eyebrow raising may be related to verbal communication. One area in which this is specially true is the development of multimodal dialogue systems in which visual information from the face can be part of the communication.

This thesis was motivated partly by an intuition that eyebrow raising is connected to the production of spoken messages, and partly by previous research supporting this idea. Two basic questions were asked: *when* do we raise our eyebrows in conversation? and *why*? These questions are not only interesting from

a psycholinguistic and cognitive point of view but they could also be a key to efficient communication in multimodal dialogue systems that make use of conversational animated agents. In order to investigate these questions, it is necessary to study spontaneous, but controlled, human behaviour in interactive communication. The Map Task (Anderson et al., 1991) and Conversational Game Analysis (Carletta et al., 1997) provided the basis needed for such an investigation.

In the following sections, I will first summarise the principal findings of this thesis (6.2.1) and how these can be related to previous work in the research literature (6.2.2). The following section (6.3) examines reasons *why* eyebrow raising occurs during dialogue. Methodological issues from this investigation will be addressed in 6.4. I will then describe practical applications of this kind of research to the area of Embodied Conversational Agents (6.5). Finally, future research will be suggested before concluding the chapter.

6.2 **When** do we raise our eyebrows in conversation?

6.2.1 Principal findings

Introspection suggests that, we tend to raise our eyebrows more when we are engaged in conversation than when we, for instance, read to ourselves. This suggests eyebrow raising is linked to *interactive* aspects of linguistic communication. Physical co-presence of an interlocutor does not appear necessary from the fact that we still raise our eyebrows when we speak to someone on the phone. Similarly, in the corpus collected here and described in Chapter 3, a fair amount of eyebrow raising was observed in the recordings where one speaker alone was giving instructions to a camera. This was probably because, although there was no interlocutor present, the speaker had in mind an 'imaginary' receiver of the instructions. Additionally, in a conversation, we often use eyebrow raising much more when we are speaking than when we are listening. This was certainly the case in the dialogues investigated here. Of the 274 eyebrow raises in the dialogues, 270 were produced while speaking. And the other four occurred in a pause between utterances or at the end of a dialogue. Thus eyebrow raising in this context is associated with *interactivity* and with the act of speaking.

In order to find out which particular aspects of a conversation are linked to eyebrow raising, in the current study the location of eyebrow raises was investigated in relation to dialogue structure and utterance function in Chapter 4, and in relation to prosodic events, namely pitch accents, in Chapter 5. The following findings were obtained in the study reported in Chapter 4:

- The length of an utterance was the best predictor of total eyebrow raising duration in it and of the number of brow raises. The latter was also influenced by the speaker identity: one speaker produced more eyebrow raises than the other two. But there were also other relations, as listed below.
- Speakers raised their eyebrows more frequently at the start of high level discourse segments (transactions) than they did elsewhere in the dialogue.
- They also raised their eyebrows more when giving instructions than when asking a question or acknowledging the receipt of some information. In addition eyebrow raising was longer in instructions than in any other kind of utterance in the dialogue.
- Speakers did not use eyebrow raising in questions more often than in other utterances.

In the study in Chapter 5 it was found that:

- The start of eyebrow raising was aligned with a pitch accent. Usually the start of the brow raise immediately preceded the start of the accented syllable, and this pattern was the same for long and short brow raises and for all three speakers.
- When speakers mentioned a new landmark in the dialogue they did not raise their eyebrows more often than when they mentioned that landmark for the second time. Nor did they use brow raising more frequently in the last part of an utterance, where new information is more likely, than in earlier parts.
- In a series of two or more pitch accents affected by downstep, brow raises were aligned with the first accent from which the rest descended.
- On the other hand, the phonological difference between primary/secondary accents did not influence their alignment with the brow raises (i.e., brow raises were not aligned with primary accents more than with secondary accents).

- As for type of utterance, brow raises were aligned with pitch accents in instructions more frequently than in other types of utterances.

In summary, from those findings we can tentatively suggest that eyebrow raising in Map Task dialogues is associated with some aspects of discourse structure and utterance function. It also is aligned with some prosodic events, namely pitch accents. And there are preliminary indications that this alignment occurs more frequently at the start of certain prosodic groups, namely downstepped accents, and in a type of utterance that carries important information to advance the dialogue. On the other hand, eyebrow raising does not seem to relate to information structure or to the phonological contrast between strong and weak accents.

6.2.2 Relation to previous research

In this thesis emotion or social aspects of body movement were not addressed. Instead, the focus was on the linguistic aspects of communication. In this sense, this thesis shares a general theoretical framework with earlier studies that have related body movements to the verbal channel and have claimed these movements are an integral part of the linguistic message. Within this theoretical framework, movements of other body parts, such as hand gestures, have been studied to a much larger extent than eyebrow movements, as we saw in Chapter 2. Eyebrow movements cannot be directly compared to for instance hand movements, and this was certainly not the aim of this investigation. To begin with, in contrast with the hands and arms, the eyebrows are very limited in their movement capacity, not only in the magnitude of movement but also in shape. The imagistic properties available in hand movements are not present in eyebrow movements alone. In this thesis, the interest was not on iconic aspects of movement but on its apparent temporal alignment with the linguistic signal and on what purpose, if any, this may have in communication. In this sense, the current study fits in with the goals of previous studies on movements of the head, eyes, hands and arms, and body posture, described in section 2.4, and more in particular with previous research on eyebrow raises presented in section 2.5. Within this theoretical approach body movement has been associated with different aspects of the linguistic message that were addressed in this thesis: discourse structure, utterance function, information structure, prosodic structure and intonational prominence.

The specific current findings listed above can be related to observations in those earlier studies as discussed in Chapters 4 and 5 and summarised below.

First, the finding that eyebrow raises appear with higher frequency at the start of transactions, was related in Chapter 4 to claims by Chovil (1989) on facial displays, and by Cassell et al. (2001) on posture shifts. Facial displays (often eyebrow movements) and postural shifts were observed at the start of new topics in the conversations in those studies. Chovil's claim was based on a very small number of cases. Also the narrative nature of the conversations in both studies was different to the task-oriented dialogues in this thesis. Nevertheless, segments at high levels of the discourse structure could be compared across those studies, and thus the current finding would seem to support the idea that body movement can signal the start of a new segment associated with a new 'topic' in a conversation. This is further related, though not so closely, to head movements marking the start of quotes in dialogues (McClave, 2000, see 2.4.3 and 4.4 above).

In terms of utterance function, results on the analysis of queries in this thesis disagree with previous observations (e.g. Birdwhistell, 1970; Eibl-Eibesfeldt, 1972; Ekman, 1979; Chovil, 1989). Speakers in the Map Task dialogues investigated did not use eyebrow raising when asking questions. The earlier observations had been presented without supporting data (with the exception of Chovil, who presented a very small percentage of her facial displays as marking queries). And the only study that did present empirical data and a thorough quantitative analysis (Srinivasan and Massaro, 2003) had found that eyebrow raising and head tilting had a small influence on the perception of an utterance as a question, while the influence of auditory cues was far larger. As explained in Chapter 4, we cannot say without further analysis of the data, whether perhaps the speakers raised their eyebrows to add a questioning meaning to other utterances. Without further investigation the only conclusion we can make is that in the Map Task dialogues studied here the delivery of questions was not associated with eyebrow raising on the part of the speaker.

The relation found between eyebrow raising and intonational events, namely pitch accents, can also be linked to previous studies in which body movement

appeared to be aligned with the prosodic structure. For instance, several researchers have reported that hand gestures, particularly *beats*, coincide with accented syllables (Kendon, 1972, 1980; McNeill, 1992; McClave, 1998; Loehr, 2004). And in a striking parallel to these studies in which the stroke phase of the movement typically preceded the accented syllable, brow raises here started by an average of $0.063sec$ earlier than the start of the syllable. Very few previous studies have investigated natural production of eyebrow raising, as in this thesis, but there is some relevant research on French in which rapid eyebrow movements were found to correlate with accentuating intonation contours (Cavé et al., 1996, 2002). There have also been some observations from perception studies that found eyebrow raises were preferred in alignment with an accented word than with a non-accented one in Swedish and Dutch (House et al., 2001; Krahmer et al., 2002a). These studies used very short synthetic stimuli which limited their interpretation of results to some extent. The current thesis provides some supporting evidence of a relation from natural production in the large context of a dialogue.

As for whether this alignment (brow raises and pitch accents) may be associated to some linguistic function, the finding in this thesis suggesting that it was not related to a contrast in information structure was in contradiction with earlier observations of body motion (McNeill, 1992; Cassell et al., 1999; Krahmer et al., 2002b). Krahmer et al. (2002b) reported stronger effects of auditory cues (pitch accents) than visual ones (eyebrow raises) on the perception of information in focus. It must be remembered that in the current research instances of new/given information (first/second mentions) were all acoustically accented. Brow raises added no contrast to this distinction. It would be interesting to look at the form of the referring expressions, to see if syntactic reductions in second mentions marked a contrast with the first mention, thus leaving less room for a brow raise to mark this contrast as well. On the other hand, other preliminary findings on the alignment between brow raises and pitch accents would seem to agree with previous claims. In particular, the association of eyebrow raising with the start of groups of downstepped accents, if confirmed, would agree with general observations on the alignment between body movement and the prosodic structure (e.g. Kendon, 1972, 1980). Similarly, indications that it may also be associated

with the delivery of instructions would also agree with the more or less general view that brow raises are used to provide emphasis (see section 2.5).

In conclusion, the findings in Chapters 4 and 5 and their relation to previous studies, suggest two *themes* in the association of eyebrow raising to the verbal message: emphasis and structure (both discourse and prosodic structure). These will be discussed below.

6.3 **Why** do we raise our eyebrows?

The findings presented above, about existing correlations between eyebrow raising and the linguistic message, provide some answers to the question of *when* we raise our eyebrows in conversation. The next question is *why* we raise them. This is a more difficult question, especially because this area of research is still at a preliminary stage. On the grounds of the findings in this thesis, we can speculate that brow raises may have different communication roles associated to them. These roles could be summarised into two hypothetical functions: structuring and emphasising.

First, it seems that by means of eyebrow raising, speakers can add a visual marker to the start of groups of utterances (i.e. transactions) and groups of words prosodically linked (e.g. by downstep). Both these groups represent coherent linguistic units in the structure of the conversation: discourse structure and prosodic structure. And by signalling the start of these units the eyebrow raises may contribute to convey this structure and maintain the coherence. Second, apart from a structuring function in dialogue, we can hypothesise that eyebrow raising has an emphasising function. This is supported by the fact that brow raises were aligned with pitch accents in the dialogues under investigation. Pitch accents can lend acoustic prominence to words that need to be emphasised. But some words may require greater emphasis. In these cases, a cue on a different channel of communication, such as body/facial movement, may add an extra signal that reinforces that segment. In this case brow raises and pitch accents would share the same linguistic goal. The pressure to satisfy this goal would sometimes give rise to covariation of both behaviours. The emphasis hypothesis is consistent with the fact that in the dialogues of this corpus, instructions were accompanied by eyebrow raises more often than other type of

utterances. As explained in Chapter 4, in the task performed by the participants of these dialogues, the instructions carried the information most important to the task's goal. The delivery of these instructions, therefore, would need to be very clear and this could be seen to warrant extra emphasis on certain bits of information. It is important to notice that this reinforcing of instructions is not intrinsic to the type of utterance then, but is related to the important function that those utterances had in this type of dialogue. Thus, in different contexts different types of utterance might be accompanied by eyebrow raising, where the brow raises would emphasise words within the utterance and at the same time, to a certain extent, they would serve to identify its function. It would be interesting for future research to investigate other types of task-oriented dialogue in order to test this hypothesis.

We have talked about eyebrow raises as *signaling* different linguistic phenomena. A question arises as to whether this is an intended signal addressed to the listener or if it is not intended and merely produced for the benefit of the speaker. There has been a long debate about this in the field of gesture studies (see Loehr, 2004, for a good discussion on this subject). On one side of the debate researchers believe that gesture has a communicative role and can add meaning to a linguistic message (e.g. McNeill, 1992). On the other side, gestures are believed to aid speech production but to add little or no meaning to it (e.g. Krauss et al., 1996). The two views are not completely antagonistic, and indeed researchers on each side accept some of the arguments of the other. I believe that eyebrow movements may be explained with arguments from both sides. Some brow raising could be a by-product of the speaker's processing effort in organising and delivering her message. Other brow raises may be intended signals to attract the attention of the interlocutor to certain parts of the message. The former would perhaps agree most with the hypothesised structuring function of eyebrow raising, whereas the latter would be more on the line of the emphasis hypothesis. We could further speculate that these two functions would be reflected on the magnitude of the eyebrow raises. That is, eyebrow raises that are a by-product of the speech production process might be smaller in magnitude, since they do not need to be perceived by an interlocutor, whereas brow raises that are directed as a signal to the listener would be larger and more easily perceived. However, there is no reason to believe that big physical gestures are intended for others and

small ones are not, because intention is not directly connected to the physical display. In fact intention is a sticky issue, because it is not possible to determine intention by looking at behaviour without manipulating it. In this thesis, intention was not studied and the investigation focused only on the produced behaviour. Some manipulations could be used in future experiments to address the issue of intention, for instance by manipulating the visual access between participants. Chovil (1989) found that motor mimicry facial displays in the listener were more frequent in face-to-face interaction than when participants could not see each other. Experiments could be done to test whether eyebrow raising on the part of the speaker would also decrease when there is no visual availability between participants. If it did not decrease we might infer that eyebrow raises were not intended signals. However it may be the case that the behaviours on the verbal and visual channels are so strongly connected in the speaker that when their interlocutor cannot see them they will continue to use eyebrow raising even if this was primarily an intended signal.

In any case, it is important to mention that the eyebrow raise might be interpreted by the interlocutor, who, due to its correlation with the linguistic signal has learnt to interpret it as a signal even if originally it was not an 'intended' signal. For this reason, I think that brow raises do have some communicative value.

6.4 Methodological issues

One of the contributions of this thesis, as mentioned in Chapter 3, is the presentation of a methodology for the collection and analysis of audiovisual data in human-human interaction, specifically eyebrow raising in dialogue. The difficulty in studying this kind of data puts high demands on the method employed. For instance, when observing facial behaviour informally, it is difficult to identify and isolate an individual behaviour such as eyebrow raising. When we see and hear a human face engaged in conversation we perceive a whole set of behaviours, some involving facial movements such as those of the lips and jaw, head, eyes, and eyebrows. And each of these seems to serve different purposes. But because we are used to perceiving all this as a whole, it is difficult to isolate a single behaviour such as eyebrow raising in order to study its possible functions. Therefore, it is important to use a rigorous method of analysis, which allows us

to observe the behaviours as they occur naturally but also to separate them and qualify them individually before describing them together. The methodology in this thesis proved successful at identifying some relationships, in this way, between the visual and auditory signals. Advantages and disadvantages of this method are discussed below with the aim of informing future research on this area.

As explained in Chapter 3, the Map Task experimental design was chosen to study facial behaviour as it occurs naturally and spontaneously in human-human interaction, while still controlling, to a certain extent, its environment. This experimental setup allowed the study of natural eyebrow raising produced in the large context of a dialogue. The method had advantages over previous studies where very short synthetic stimuli were used in perception experiments. It also had advantages over production studies where speakers simply narrated stories involving a much smaller degree of interaction with an interlocutor and less predictable structures and goals in their utterances. The current design proved to be successful at eliciting naturally the behaviours of interest within a large interactive context that was minimally controlled to contain specific intentions and known goals. In terms of recording conditions, less optimal were perhaps the presence of camera men in the same room as the participants. While it seems unlikely that the resulting changes to participants' natural behaviour were critical, future research could experiment with modern cameras, fixed or remotely controlled, that could allow the recording of constant closeup views of the participants' faces.

Once the data was recorded, the approach was to annotate the auditory and visual channels separately. This was important and marked a change from some earlier studies. Observations of one channel can easily bias the perception of the other, especially if behaviours between them are correlated. And this leads us to another important point that was noticed while doing the annotation, namely the association between behaviours within the visual channel. Informal observation seemed to reveal a correlation between movements such as head movements and eyebrow raising. Sometimes both behaviours appeared together, others a head movement appeared where perhaps an eyebrow raise would be expected or would have seemed just as natural, and viceversa. If head movements and eyebrow raises covaried sometimes, then it would be difficult to identify a

shared function by observing only one of the two. Thus it would seem appropriate to include several kinds of body movement in an analysis searching for correlates between the visual and auditory channels. However, for the current purpose and because the data annotation is a time-consuming process, looking at eyebrow raising alone was considered the best choice.

The method employed to annotate the eyebrow raises was a human observational system, in which one coder manually annotated the electronic record as described in Chapter 3. One of the problems with this method was that, as was just mentioned, the annotation process was very time-consuming and this limited the amount of data that could be analysed. For instance, a very small number of participants were included in the study. The time spent on annotation also made it impossible, due to limited resources, to carry out a large-scale reliability test of the annotations performed. All this has repercussions for the interpretation of the results of the analyses. Obviously, a faster method would have been preferred, such as the automatic measuring of eyebrow raising. At the time when the annotation was done, however, there was no available automatic system. And unfortunately, current systems such as those derived from computer vision techniques are not yet fully developed for the reliable annotation of some of the subtle movements observed as participants moved freely in this study. Future research in this area should be aware of this caveat in a human observational method and plan time and resources accordingly. Also, because of the subjective nature of such methods, tests for reliability of the annotation system should be performed when possible. A large set of data and coders are necessary in order to perform these tests. In preliminary studies like the one here it is necessary to keep a manageable size of analysis in order to explore possible relations of interest that can then be investigated on studies of larger scale. The current methodology was thus very useful to obtain preliminary findings that can point directions and also save time in future larger studies.

In terms of annotation schemes, one of the decisions here was to annotate the start and ends of the events and then establish associations by looking at their start. In the case of eyebrow raises and pitch accented syllables, another point of interest would have been, for instance, the point of maximum rise of the eyebrows and the point of maximum excursion of the F0. The decision was a purely methodological one, partly based on the fact that the start of the events were

much easier to identify and therefore more reliable. Future studies looking at different points in the events should take measures to ensure reliable identification of those points.

One technical limitation in the methodology of this investigation was the lack of a comprehensive tool for the annotation and analysis of multimodal behavioural data. At the time of this investigation, no available system fully met all the needs of this kind of research, where a detailed visual and acoustic analysis, with good import and export of data, are necessary (see Bigbee et al., 2001 for a review of some systems). The lack of a fully developed system was a great disadvantage that caused problems and delays in the current study. It is hoped that research like this will inform and encourage the development of better multimodal annotation and analysis tools for the study of natural behaviour. Proving the benefits that the industry can gain with this kind of research, seems a positive way to encourage the development of such tools.

6.5 Practical applications: Embodied Conversational Agents

Section 2.7 above introduced Embodied Conversational Agents (ECAs) as an area that can benefit from research on human facial behaviour. The findings of this thesis can provide some guidelines to the design of such systems. As discussed in 2.7.2, in the development of ECAs there are serious challenges derived from the fact that we do not have a good model and understanding of human conversational behaviour to allow us to automatically generate this behaviour. A fundamental problem in creating these animations is the lack of information about how to synchronise the facial movements of the agent with the speech signal. This is not a trivial matter, since misalignment of the auditory and visual channels can be at minimum distracting and, in the most severe case, can affect and break down the communication process in human-computer interaction. A parallel of a distracting interference can be made with a badly dubbed movie. It is not only the asynchronous lip movements that become distracting, the facial movements in general can also be disturbing and interfere with comprehension. In the case of ECAs, similar interference would be annoying in, for instance, educational applications, where an agent with poor synchronisation and inadequate movement would fail to engage the user and would hinder rather than help his

or her focus on the materials being taught. But as ECAs become more pervasive and are applied to more areas of communication systems, poor quality in the generated conversational behaviour may have a much bigger impact. For instance, in an emergency situation, if following instructions from an ECA on how to follow a safe evacuation route, the agent's facial movements could be crucial in delivering the message fast and efficiently. Clearly, as the demand grows for more natural and communicative ECAs, more research on facial movements will be needed in order to improve the design of such systems.

Here it is important to explain that the point is not about making the animated agent look completely human in its physical appearance. In fact, this is not desirable because as the agent looks more human to users, their expectations about its conversational capabilities will most likely grow, and their tolerance for system errors will decrease. What should be human-like is the conversational style of the agent, and particularly the alignment of its movements with events in the speech signal. Thus a cartoon face that shows human-like conversational patterns would be more appropriate than a very realistic human face, because it would allow users to maintain a conversation while keeping them aware that they are interacting with a computer.

Another important point that must be explained is why we should use a face at all in such applications. That is, if we cannot yet generate visual conversational behaviour and if unnatural movements can actually interfere with communication, then why not just use speech? In some situations, such as in noisy environments, where information from the face is known to aid comprehension the auditory signal will not be enough. Lip movements, for instance, do not only *add* information to the acoustic signal, they also integrate with it in a way that can affect perception (e.g. McGurk and MacDonald, 1976). Because a relation appears to exist also between non-articulatory movements of the upper face and the speech signal, adding those movements could contribute important information to the delivery of a message and would help disambiguate the acoustic signal by adding an extra (visual) communication channel. But for these movements to be helpful instead of distracting, the relation between channels must be maintained in the artificial behaviour as it exists in natural behaviour.

The question then is, what will make an ECA show natural conversational behaviour in its facial movements? A basic but important point is to show some variation. As Krahmer and Swerts (2004) already pointed out, an agent that shows no other facial movement than lip movements will look unnatural to the user who will find it unpleasant to interact with. Just as variation in speech is necessary, and some of it is functional because it communicates relevant information, Krahmer and Swerts (2004) argue that facial variation is required as well. To achieve this variation, many artificial agents use some kind of *Perlin noise* (Perlin, 1995), that is, small random movements. This can make the agent appear more natural to a certain extent, but as Krahmer and Swerts explain, the resulting variation is small and not functional in a linguistic sense. Therefore in order to make an ECA appear natural and be communicative, we need to know how much of the variation can be random and how much should be linked to linguistic functions.

In the data analysed in this thesis a large amount of the eyebrow raising behaviour could not be accounted for. As we saw in Chapter 4, only around 27% of the variance in the number of brow raises per utterance could be explained by the model proposed. This obviously does not mean that the rest of the variance will be strictly random, but that it cannot be explained by any of the variables evaluated. Within that 27%, the majority of the variance was actually related to simply the length of the utterance, indicating again that brow raises just occurred randomly. Another large portion of the variance indicated that there was variation between the speakers. While this makes the study of eyebrow raising difficult for the investigator, it is an important piece of information for the generation of this behaviour in artificial agents. If variation is observed between real individuals without causing problems for interaction, then different artificial agents should probably also show variation between them to make them more natural. And at the same time, a single agent reproducing an individual style should be accepted and meaningful to a user even if this style does not generalise to the whole population, as long as it is consistent and coherent within this individual. Finally, a smaller but significant portion of the variation in eyebrow raising was linked to linguistic phenomena in the study presented in Chapter 4, and some other correlations were found in Chapter 5. This could serve as guidelines in testings for more natural ECAs as discussed below.

The findings in this thesis could be incorporated in an ECA providing navigation instructions, for instance, which would be similar to the context in which the participants in the current thesis interacted. An agent in such system could use eyebrow raising when starting the description of a new route or of a new section in a route being described. Within these sections, it could also raise its eyebrows when emphasising important words in the delivered instructions, and when doing so, it should start raising the eyebrows on or slightly before an accented syllable. Also, in close prosodic units, such as groups of accents affected by downstep, an eyebrow raise on the first accent should be preferred to one on later accents. The agent will probably appear more natural if showing eyebrow raising in a small percentage of its utterances, as suggested by the current data in which less than 40% of the utterances contained some eyebrow raising, and only 14.7% of the accented syllables were accompanied by a brow raise. Finally, in contrast with previous suggestions in the literature, it does not seem that the agent would improve its naturalness by showing eyebrow raises on questions and on the introduction of new entities in the dialogue.

The guidelines suggested above could also be incorporated and tested in dialogue systems different from navigation systems, in order to evaluate whether they can be extended to other domains. They are obviously very limited guidelines, but in an area like ECAs, where so little information on appropriate facial movements is available at the moment, even small suggestions to test could lead the way to big improvements in the long run. More importantly, they can encourage further research and the use of ECAs as testing beds for psycholinguistic hypotheses.

6.6 Future directions

Some future directions for research have already been suggested in this and earlier chapters. The most important perhaps is the collection of more data including a larger number of participants. This would allow us to test whether the behaviours observed can be confirmed in a larger sample of the population, and also to explore further possible relations between eyebrow raising and the verbal channel. Also, data from different types of dialogues could be collected to test whether the findings here can be extended to other domains.

About the specific findings in this thesis, Chapter 4 proposed the exploration of the interlocutor's behaviour immediately following a brow raise by the main speaker. This might provide some insight into how speakers' eyebrow raises are interpreted by listeners and ultimately into the role of eyebrow raising. For instance, if the interlocutor often produces a reply following a brow raise by the main speaker, then it could be that the brow raise is interpreted as querying or requesting confirmation of understanding, even if the utterance it accompanies is not a question. On a different matter, Chapter 5 suggested it would be interesting to find out whether the end of a brow raise is aligned with the end of a downstep group, just as the brow raise start was aligned with the start of the group. This would support further the marking of phonological groups by means of eyebrow raising. Also, in connection with the finding that brow raising did not mark a contrast between *new/given* information on referring expressions further analysis of first/second accented mentions was suggested to see if syntactic reductions on second mentions could have meant less need for a visual marker to emphasise this distinction (preliminary observations suggest this was not the case). Also, the exploration of other contrasts such as those between words of direction was proposed.

A way of testing the specific findings of this thesis, as mentioned in section 6.5, would be to implement them in an ECA and evaluate its perceived naturalness and communicative power. In fact, a combination of production, implementation, and perception studies would be a good way of obtaining information on natural behaviour, as suggested by Krahmer and Swerts (2004). But ideally the production studies should involve large natural contexts, as in this thesis, and the perception studies should use longer stimuli than those in previous studies, within a larger context as well.

In this thesis the analysis of brow raises did not distinguish them in terms of their length or their magnitude. Future research could look separately at temporally short/long raises and also at minor/larger raises. It is likely that especially the difference short/long would reveal different functions of eyebrow raising associated to them. In terms of the magnitude of the brow raise, dividing them into minor ones and larger ones, and comparing their contexts, was first suggested as a possible way of testing the intentionality of eyebrow raises. However, because intention is not necessarily linked to the physical display, other ways to test this

were suggested such as experiments manipulating visual availability between participants.

The effect of the physical presence of an interlocutor on the speaker's facial behaviour could be studied in the current corpus by comparing the different conditions in the recording sessions: dialogues and monologues. As explained in Chapter 3, in the monologue rehearsal and recording the speaker gave instructions on a map as if she was addressing a potential viewer who would later have to draw the route from her video-recorded instructions. Preliminary observations suggest that monologues have proportionally more brow raises than dialogues. This could be because if there is no interlocutor present, the speaker has no feedback as to whether she is understood, and this may cause her to reinforce her utterances further by using more brow raises. This suggestion would need further investigation.

A more general suggestion for future research is to look at other body movements together with eyebrow raising, such as head movements. While this would require longer time spent on annotation and analysis, it is believed that with adequate resources much can be learned from a more comprehensive approach to body movement. On the observation of eyebrow raises in the recorded data, it was noticed how sometimes different body movements seemed to alternate, as if sharing a function but expressing it at different times. For instance, sometimes head movements seemed to emphasise a word where perhaps an eyebrow raise would have seemed just as natural and vice versa. Different kinds of movements (e.g. head and eyebrows) could be annotated separately and then combined in order to investigate to what extent they can signal the same linguistic phenomena. Finally, it would be interesting to explore how eyebrow raises, and other movements observed, may be incorporated into a model of speech production such as De Ruiter's Sketch Model for the production of gesture and speech (De Ruiter, 2000, 1998).

6.7 Final conclusions

Eyebrow raises are not something we automatically associate to anything linguistic. However, careful observation of people's faces in conversation does give an impression that their eyebrow raises are somehow connected to what they are

saying or to how they say it. Could we consider eyebrow raises a linguistic phenomenon and should we include them in the study of conversational behaviour? This thesis has proved a correlation between linguistic aspects of a dialogue and the eyebrow raises produced by the speakers, but due to the large degree of variation found it is perhaps too early to classify eyebrow raising as a linguistic phenomenon. Obviously, eyebrow raises may have different functions, not all associated with the linguistic message, and so a clear relation with the linguistic channel may be difficult to find. The most important contribution of this thesis is the fact that with a rigorous method some relations with the dialogue structure and the prosodic structure were found, encouraging further research into eyebrow raises as part of conversational behaviour. Two possible general functions were suggested for eyebrow raises in this context, namely structuring and emphasising. Future research may be able to confirm these functions and to use the findings as guidelines in the design of dialogue systems using Embodied Conversational Agents.

APPENDIX A

Maps used in the corpus collection

Included here are the five maps used in the corpus collection. There were four map pairs, each consisting of an *Instruction Giver* map (with a route on it), and an *Instruction Follower* map (without the route). Also, there was an additional single *Instruction Giver* map, the museum scene, which was used in the first recording (rehearsal of monologue) for each participant.

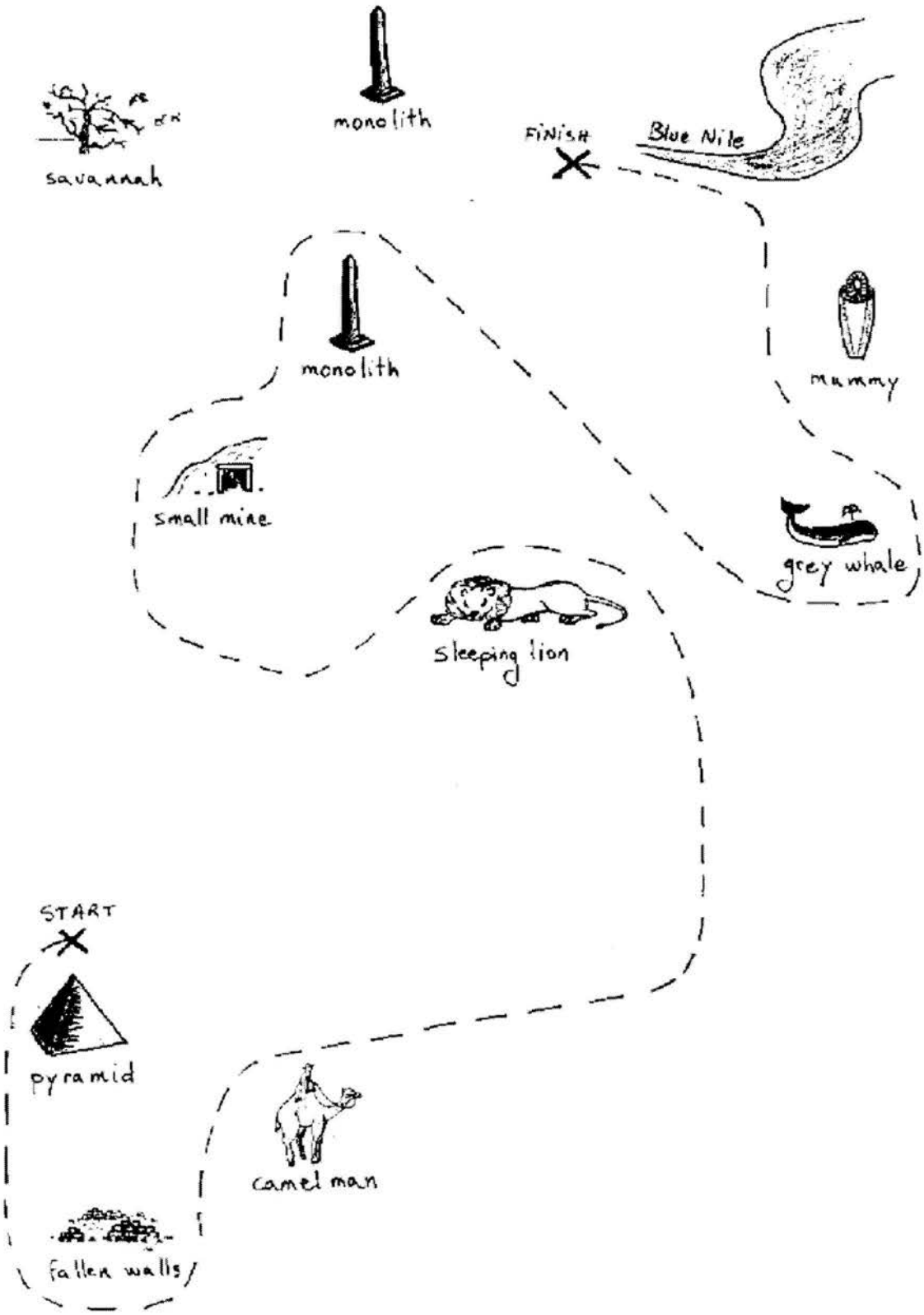

Figure A.1: Desert map (*Instruction Giver*)

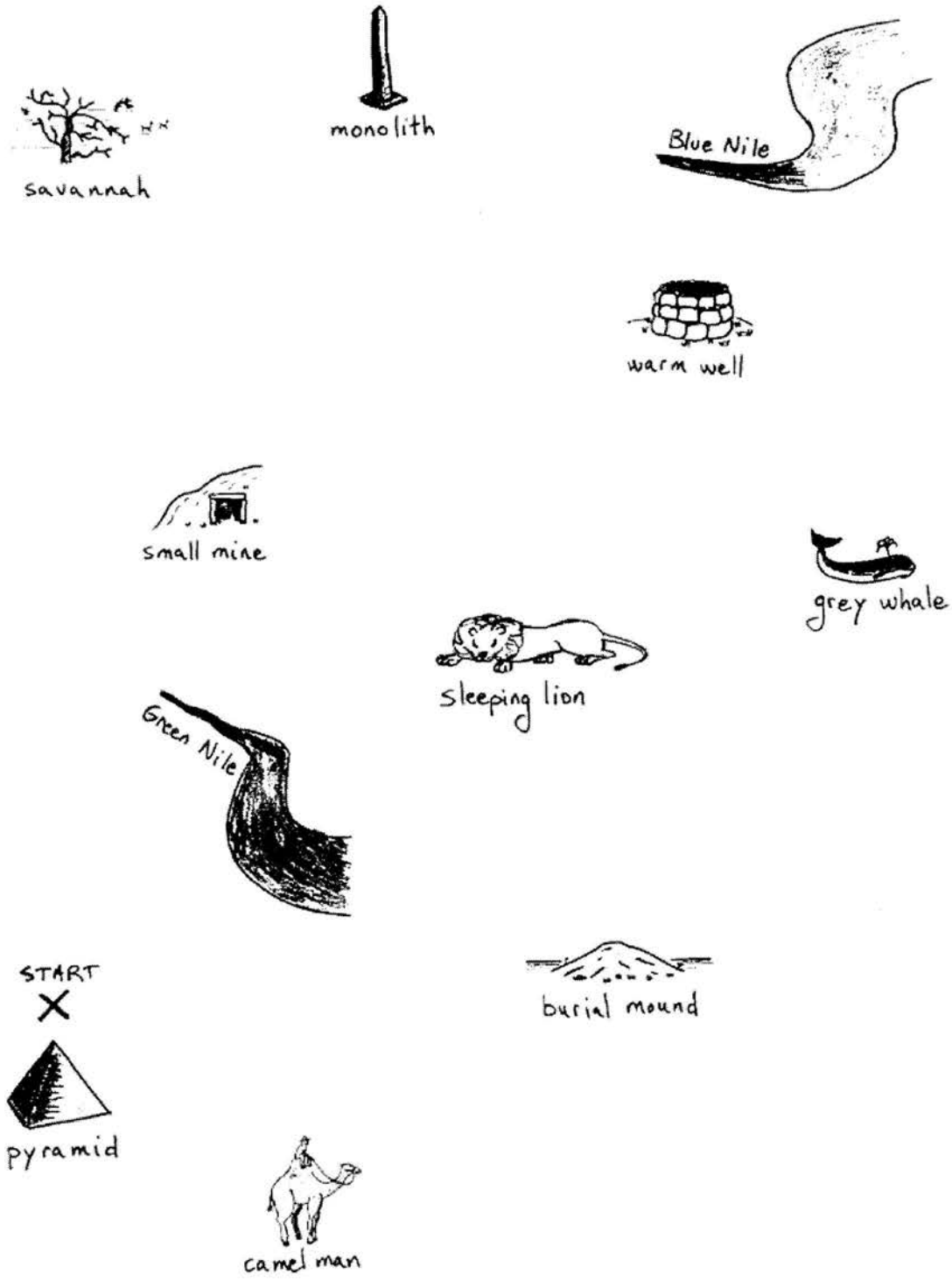

Figure A.2: Desert map (*Instruction Follower*)

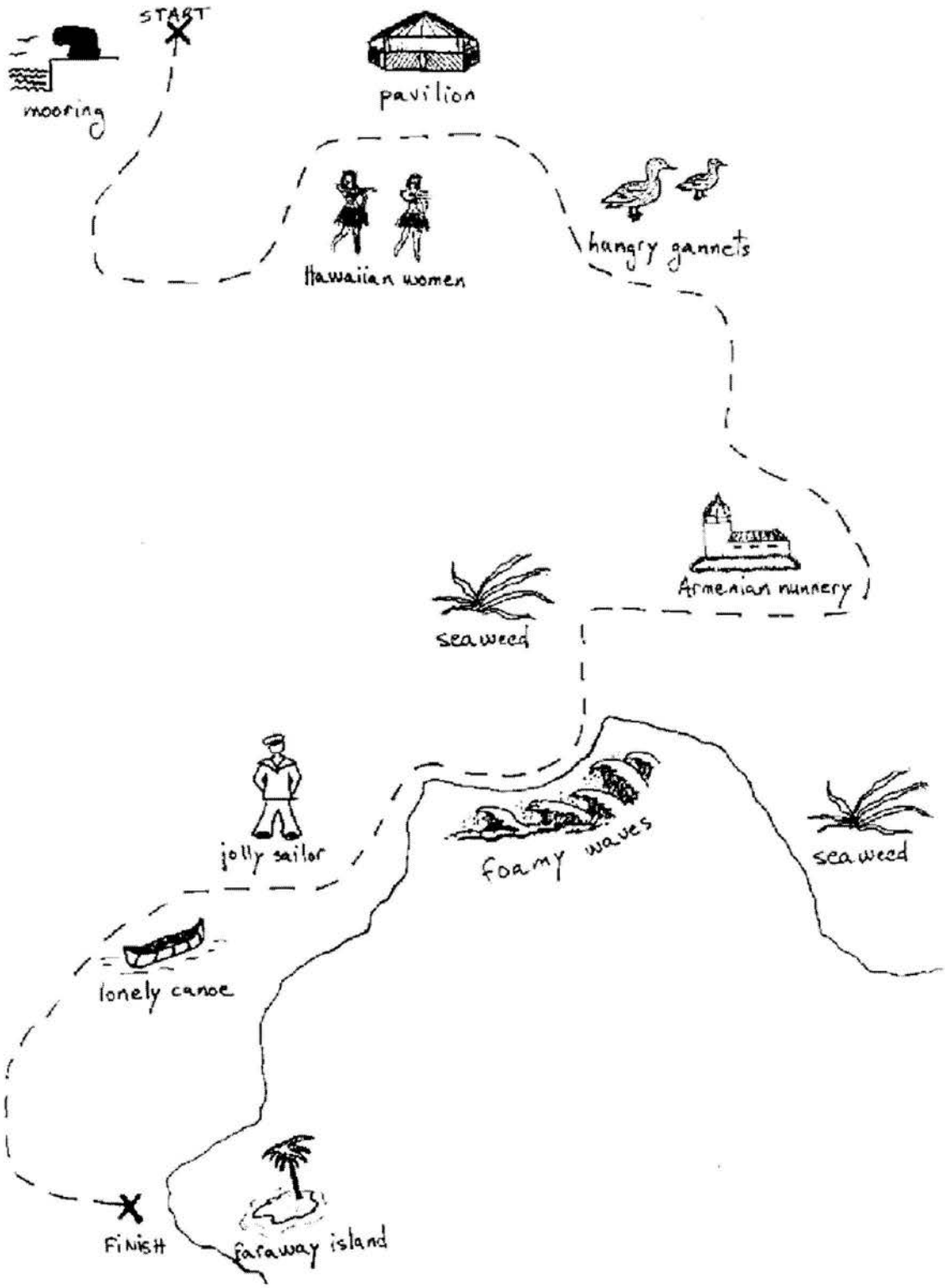

Figure A.3: Sea port map (*Instruction Giver*)

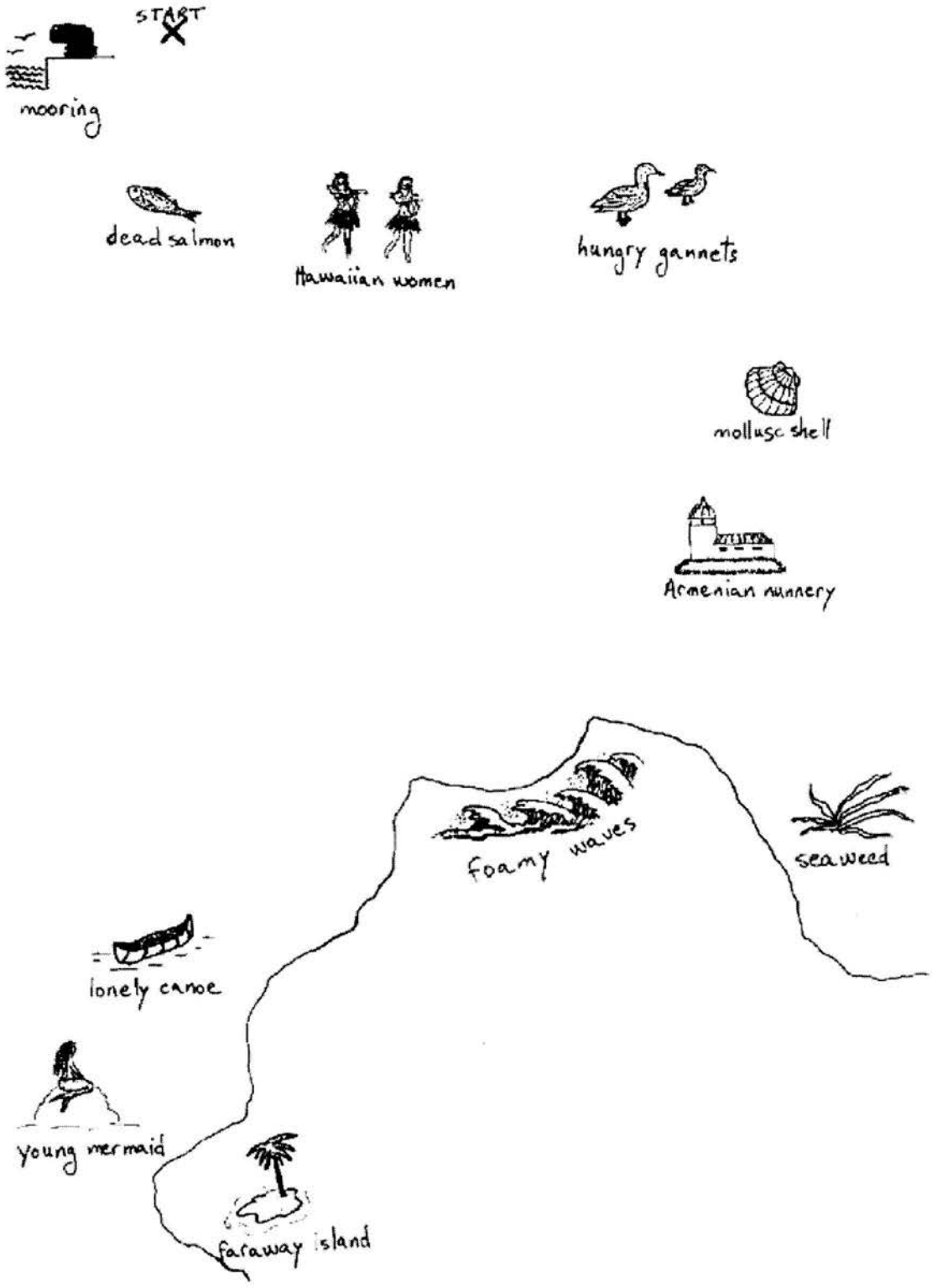

Figure A.4: Sea port map (*Instruction Follower*)

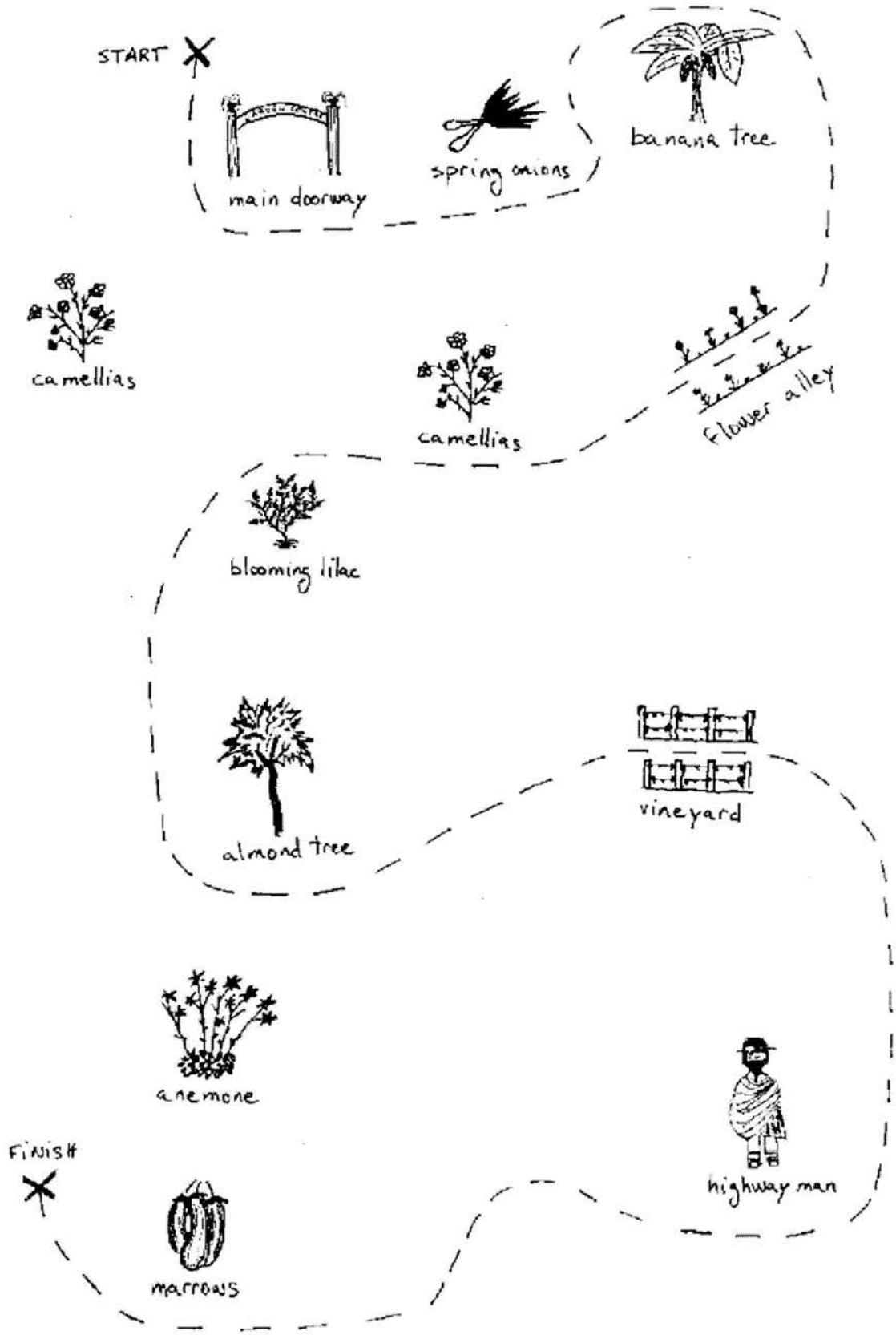

Figure A.5: Garden centre map (*Instruction Giver*)

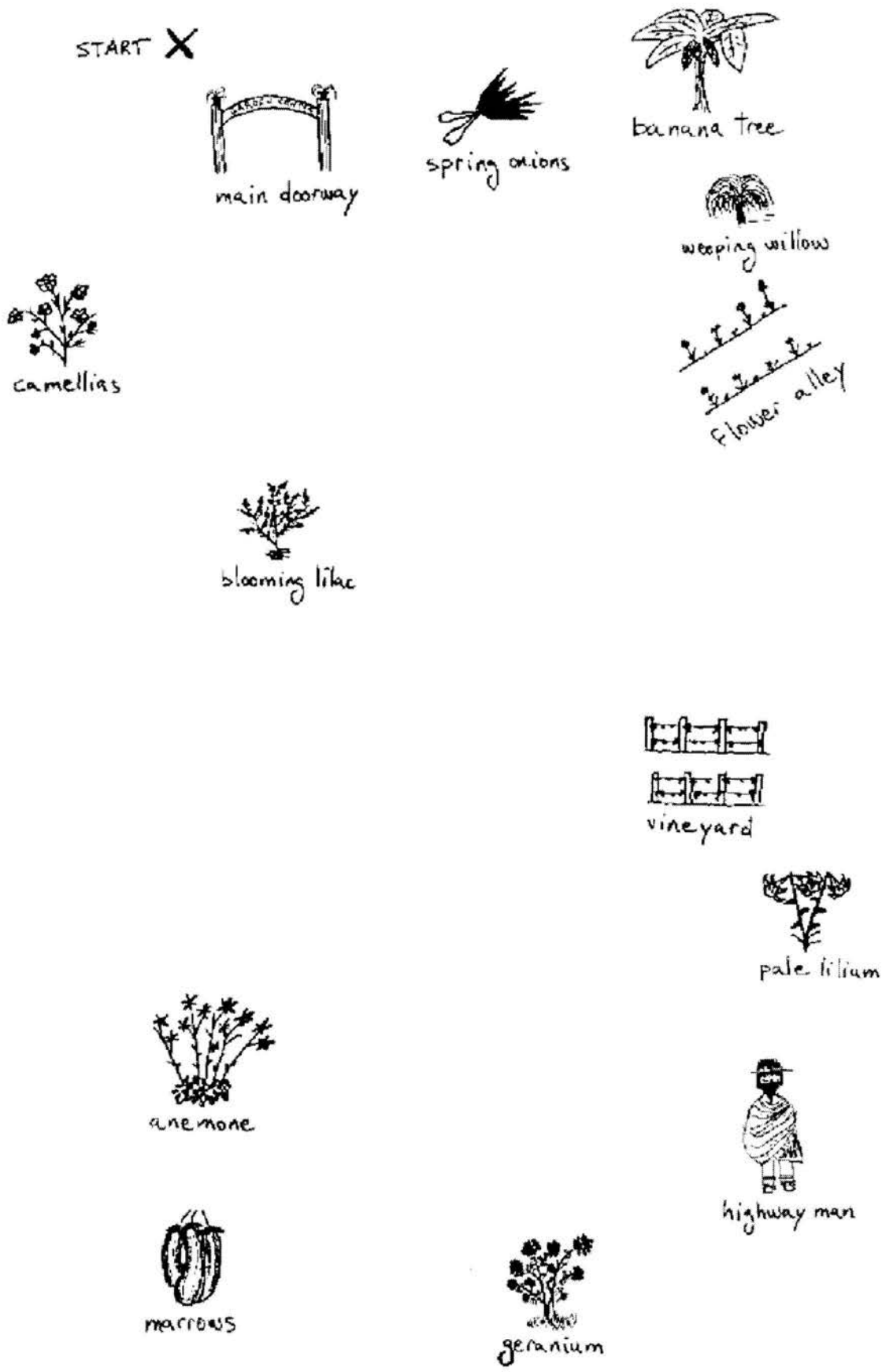

Figure A.6: Garden centre map (*Instruction Follower*)

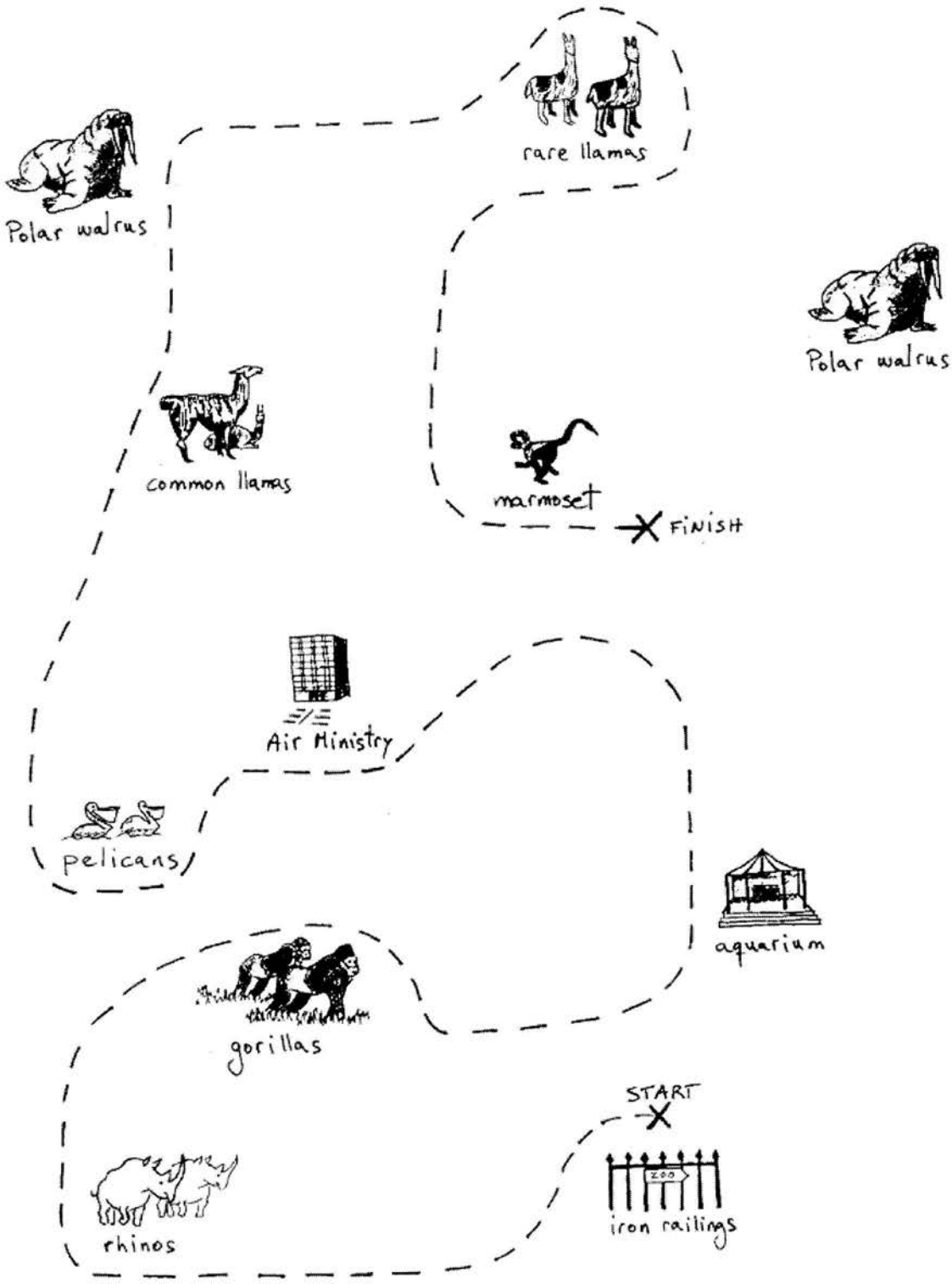

Figure A.7: Zoo map (*Instruction Giver*)

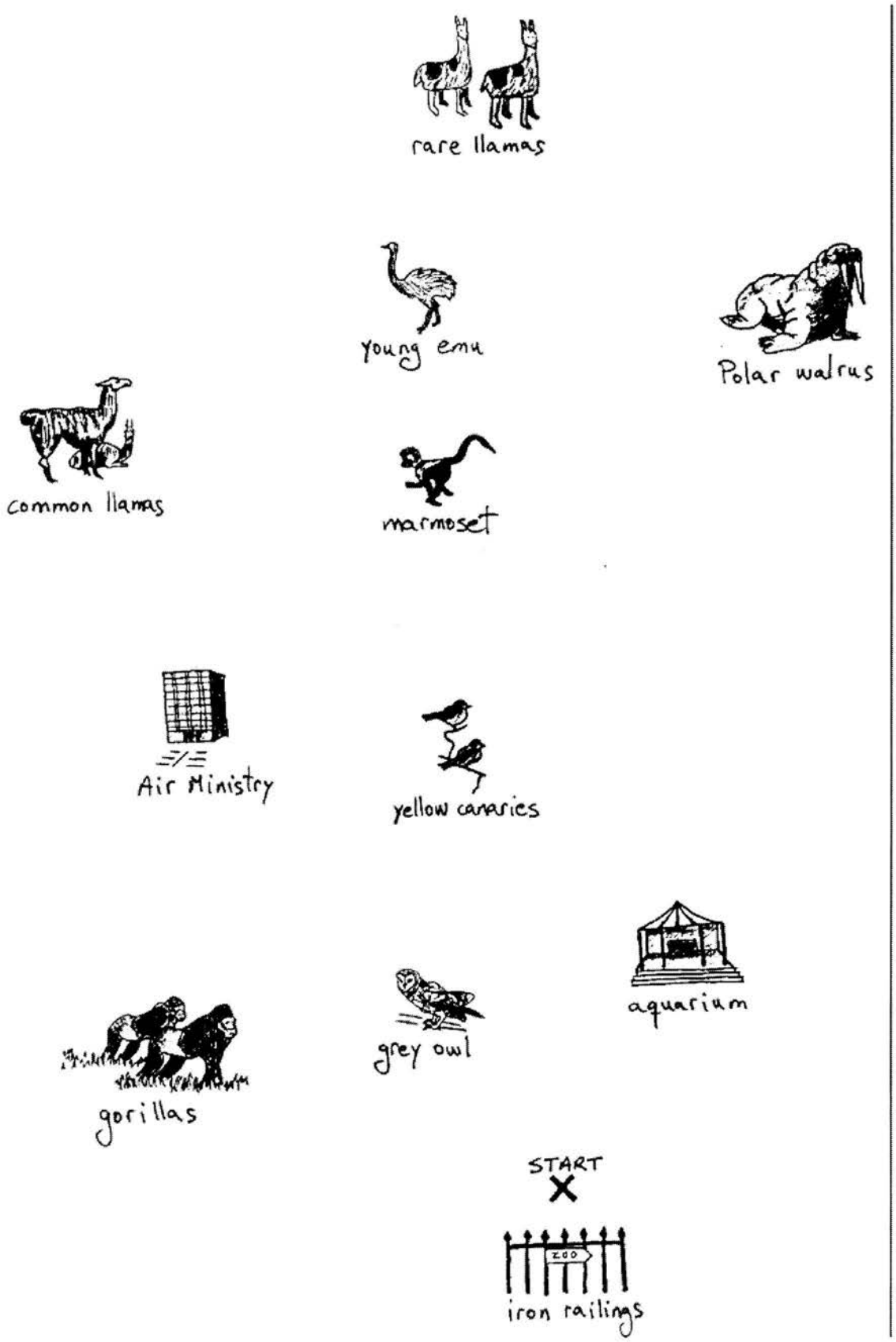

Figure A.8: Zoo map (*Instruction Follower*)

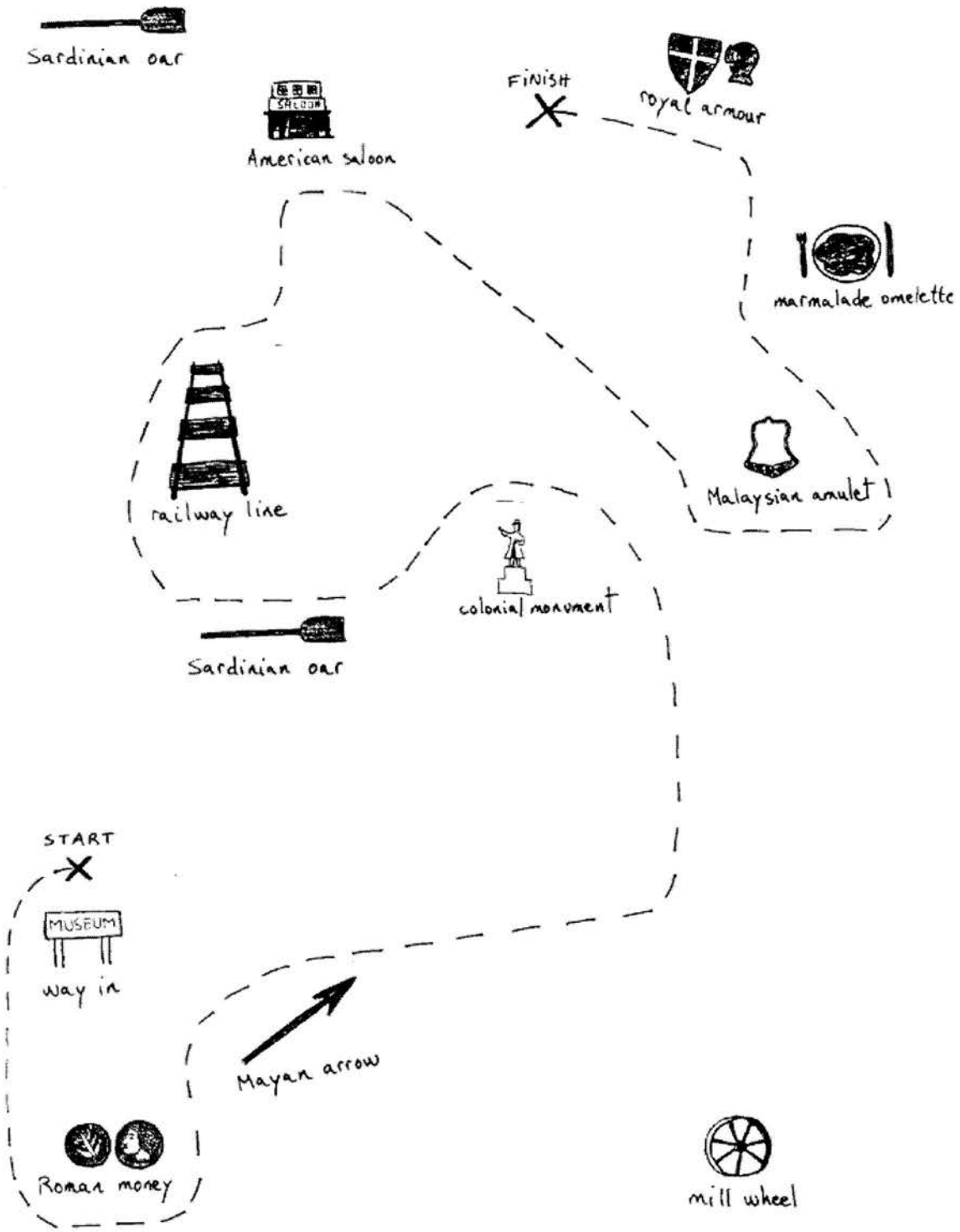

Figure A.9: Museum map (*Instruction Giver*)

APPENDIX B

Instructions to participants

The following written instructions were given to participants (without the headings included in brackets here). The order of the recording sessions was not the same for all participants, and so the instructions were not alwas presented in the order below. In the last recording session participants had to read the list of landmark names. No written instructions were provided for this session.

(General Instructions)

What you have in front of you is a map of a route to a buried treasure. Your task is to describe the route so that someone with another map of the same area, but with no route marked, can draw the route to the treasure.

The person who has to draw the route will have a map drawn by a different explorer, so the two maps may not be identical.

When you instruct the other person, you don't have to get the route right to the millimetre; you just need to avoid the obstacles. You can say whatever you want, whenever you want, but you cannot gesture with your hands.

You will be doing this task several times:

1. The first time is just a rehearsal while we adjust the cameras and the sound recording, and you get your head around the problem.
2. The second time is a real video session (using a different map), in which we will tape your instructions to play back to someone else later on.
3. In the third, your friend will have to follow your instructions.

Right now we just want you to practice and settle in your own mind how you should describe the route so that someone can reproduce it from a video of your instructions. That person, will not be able to ask questions or stop you when they're in trouble, because the video will be shown to them when you're not around.

You might want to keep in mind that when you look at the camera that is filming you, you appear to be looking at the viewer. Remember that you can say anything you want at any speed, but no hand signals!

Do you have any questions?

So now you'll do a rehearsal, giving out the instructions aloud.

(Monologue Recording)

Now we will make the real recording of your instructions, so just imagine someone watching you on video and trying to draw this route on another map of the area.

(Dialogue Recording: *Instruction Giver*)

Now you will describe the route to your colleague. Remember:

What you have in front of you is a map of a route to a buried treasure. Your task is to describe the route to your colleague (who has a map of the same area, but with no route marked), so that she can draw the route to the treasure.

Your colleague has a map drawn by a different explorer, so the two maps may not be identical.

When you instruct her, you don't have to get the route right to the millimetre; you just need to avoid the obstacles. You can say whatever you want, whenever you want, but you cannot gesture with your hands.

(Dialogue Recording: *Instruction Follower*)

You have a map of an area where there is a treasure buried. Your colleague has a map of the same area with a route drawn on it that takes you to that treasure. Your task is to draw that route on your map following her instructions. The maps were drawn by different explorers so they may not be identical.

You don't need to reproduce the route precisely, but you do have to avoid all the obstacles. You both can ask questions and say whatever you want whenever you want, but you cannot look at one another's maps and you cannot gesture with your hands.

Any questions before we start?

APPENDIX C

Instructions to second coder on annotation of brow raises in a subset of the data

The following written instructions were given to the second coder who annotated the number of eyebrow raises in a small subset of the data, as described in Chapter 3.

Instructions for the annotation of number of eyebrow raises

You will be presented with a set of 30 short segments from video recordings of three different participants in several dialogues. Your task is to annotate the number of eyebrow raises that the speaker on the left produces in each segment. For the current research purpose an **eyebrow raise** is defined as:

> Any upward movement, from a baseline neutral position, of at least one eyebrow and observable on the digital video recordings.

There are different ways in which the eyebrows can be raised. In some eyebrow raises only one part of the eyebrows is pulled up. For instance, sometimes only the inner portion of the eyebrows is raised. Other times, one eyebrow is raised more than the other. Also eyebrow raises vary in length and intensity. All these cases count as eyebrow raising without distinction. So basically, you should count one eyebrow raise every time you see at least one part of an eyebrow going up and then down.

You will watch the video segments without sound and hiding the lower part of the participants' face (you will be shown how to do this). Also, because some subtle eyebrow raises cannot be observed at normal speed, you should watch the segments in slow motion first, and as many times as you think necessary, before deciding whether eyebrow raising occurred or not. Horizontal wrinkles forming on the participant's forehead will be a useful indication that the eyebrows are raised.

You have been given a sheet of paper with a numbered list of the segments you will watch. Simply write down, in the space provided, the number of eyebrow raises you observe in each segment.

If you have any doubts, please do not hesitate to ask!

Thank you for your collaboration!

References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C. Thompson, H. S., and Weinert, R. (1991). The HCRC map task corpus. *Language and Speech*, 34:351–366.

Bard, E. and Aylett, M. (1999). The dissociation of deaccenting, givenness, and syntactic role in spontaneous speech. In *Proceedings of ICPhS*, pages 1753–1756.

Bateson, G. and Mead, M. (1942). *Balinese character: a photographic analysis*. New York Academy of Sciences.

Bavelas, J. B. and Chovil, N. (1997). Faces in dialogue. In Russell, J. A. and Fernández-Dols, J. M., editors, *The Psychology of Facial Expression*, pages 334–346. Cambridge University Press.

Bell, C. (1844). *The anatomy and philosophy of expression*. George Bell, third edition.

Bigbee, T., Loehr, D., and Harper, L. (2001). Emerging requirements for multimodal annotation and analysis tools. EUROSPEECH 2001, Aalborg, Denmark.

Birdwhistell, R. (1952). *Introduction to kinesics: An annotation system for analysis of body motion and gesture*. University of Louisville, Louisville.

Birdwhistell, R. (1970). *Kinesics and context*. University of Pennsylvania Press, Philadelphia.

Bolinger, D. (1983). Intonation and gesture. *American Speech*, 58(2):156–174.

Bolinger, D. (1986). *Intonation and its parts: Melody in spoken English*. Stanford University Press, Stanford, CA.

Brown, G., Anderson, A., Yule, G., and Shillcock, R. (1983). *Teaching Talk*. Cambridge University Press.

Brown, G. and Yule, G. (1983). *Discourse Analysis*. Cambridge University Press.

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

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., and Yan, H. (2000a). Human conversation as a system framework: Designing embodied conversational agents. In Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., editors, *Embodied Conversational Agents*, pages 29–63. The MIT Press, Cambridge, MA.

Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., and Rich, C. (2001). Non-verbal cues for discourse structure. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics*, pages 106–115, Toulouse, France.

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (2000b). *Embodied Conversational Agents*. The MIT Press, Cambridge, MA.

Cassell, J., Torres, O., and Prevost, S. (1999). Turn taking vs. discourse structure: how best to model multimodal conversation. In Wilks, Y., editor, *Machine Conversations*, pages 143–154. Kluwer Academic Publishers, The Hague.

Cavé, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., and Espesser, R. (1996). About the relationship between eyebrow movements and F0 variations. In *Proceedings of ICSLP*, volume 4, pages 2175–2178, Philadelphia.

Cavé, C., Guaitella, I., and Santi, S. (1993). Relations entre geste et voix: Le cas des sourcils et de la fréquence fondamentale. In *Images et Langages: Multimodalité et modelisation cognitive*, pages 261–268, Paris. Actes du colloque interdisciplinaire du Comité National de la Recherche Scientifique.

Cavé, C., Guaitella, I., and Santi, S. (2002). Eyebrow movements and voice variations in dialogue situations: an experimental investigation. In *Proceedings of ICSLP*, pages 2353–2356, Denver.

Chovil, N. (1989). *Communicative functions of facial displays in conversation*. PhD thesis, University of Victoria, BC, Canada.

Chovil, N. (1991a). Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction*, 25:163–194. Published as 1991/92.

Chovil, N. (1991b). Social determinants of facial displays. *Journal of Nonverbal Behavior*, 15:141–154.

Chovil, N. (1997). Facing others: a social communicative perspective on facial displays. In Russell, J. A. and Fernández-Dols, J. M., editors, *The Psychology of Facial Expression*, pages 321–333. Cambridge University Press.

Cohn, J., Reed, L. I., Ambadar, Z., Xiao, J., and Moriyama, T. (2004). Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*, The Hague, The Netherlands.

Cohn, J. F. and Ekman, P. (to appear). Measuring facial action by manual coding, facial EMG, and automatic facial image analysis. In Harrigan, J., Rosenthal, R., and K., S., editors, *Handbook of nonverbal behavior research methods in the affective sciences*. NY: Oxford.

Cohn, J. F., Xiao, J., Moriyama, T., Ambadar, Z., and Kanade, T. (2003). Automatic recognition of eye blinking in spontaneously occurring behavior. *Behavior Research Methods, Instruments, and Computers*, 35:420–428.

Condon, W. S. (1970). Method of micro-analysis of sound films of behavior. *Behavior Research Methods and Instrumentation*, 2(2):51–54.

Condon, W. S. (1979). An analysis of behavioral organization. In Weitz, S., editor, *Non-verbal communication*, pages 149–167. Oxford University Press.

Condon, W. S. and Ogston, W. D. (1971). Speech and body motion synchrony of the speaker-hearer. In Kjeldergaard, P. M., Horton, D. L., and Jenkins, J., editors, *Perception of language*. Charles E. Merrill Publishing Company, Columbus, Ohio.

Corina, D. P. (1989). Recognition of affective and noncanonical linguistic facial expressions in hearing and deaf subjects. *Brain and Cognition*, 9:227–237.

Crystal, D. and Davy, D. (1969). *Investigating English style*. Longman, London.

Darves, C. and Oviatt, S. (2004). Talking to digital fish. In Ruttkay, Z. and Pelachaud, C., editors, *From brows to trust: Evaluating embodied conversational agents*, pages 271–292. Kluwer Academic Publishers.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. New York: Philosophical Library. Third edition (1998) with Introduction, Afterword, and Commentary by Paul Ekman: New York: Oxford University Press.

De Ruiter, J. P. (1998). *Gesture and speech production*. PhD thesis, University of Nijmegen.

De Ruiter, J. P. (2000). The production of gesture and speech. In McNeill, D., editor, *Language and gesture: Window into thought and action*, pages 284–311. Cambridge University Press, Cambridge.

Duchenne de Boulogne, G. (1862). *Mécanisme de la physionomie humaine ou analyse électro-physiologique de l'expression des passions applicable à la pratique des arts plastiques*. Renouard, Paris.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292.

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 3(2):161–180.

Efron, D. (1941). *Gesture and Environment*. King's Crown Press, Morningside Heights, NY. Republished 1972 as Gesture, Race, and Culture. The Hague: Mouton.

Eibl-Eibesfeldt, I. (1972). Ritual and ritualization from a biological perspective. In Hinde, R. A., editor, *Non-verbal communication*, pages 297–312. Cambridge University Press, Cambridge.

Eibl-Eibesfeldt, I. (1979). Ritual and ritualization from a biological perspective. In von Cranach, M., Foppa, K., Lepenies, W., and Ploog, D., editors, *Human ethology*, pages 3–55. Cambridge University Press, Cambridge.

Eibl-Eibesfeldt, I. and Hass, H. (1967). Neue wege der humanethologie. *Homo*, 18:13–23.

Ekman, P. (1979). About brows: emotional and conversational signals. In von Cranach, M., Foppa, K., Lepenies, W., and Ploog, D., editors, *Human ethology*, pages 169–249. Cambridge University Press.

Ekman, P. (1980). *The face of man: expressions of universal emotions in a New Guinea village*. Garland STPM Press, New York.

Ekman, P. (1997). Expression or communication about emotion. In Segal, N., Weisfeld, G. E., and Weisfeld, C. C., editors, *Uniting Psychology and Biology: Integrative perspectives on human development*, pages 315–338. APA, Washington, DC.

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17:124–129.

Ekman, P. and Friesen, W. V. (1978). *The Facial Action Coding System*. Consulting Psychologists Press, Palo Alto, CA.

Ekman, P., Friesen, W. V., and Ellsworth, P. (1972). *Emotion on the human face: Guidelines for research and an integration of findings*. Pergamon Press, New York.

Ekman, P., Friesen, W. V., and Hager, J. (2002). Facial Action Coding System. eBook. HTML demonstration version of the manual retrieved from http://face-and-emotion.com/dataface/library/refroom.jsp.

Ekman, P., Friesen, W. V., and Tomkins, S. S. (1971). Facial affect scoring technique: A first validity study. *Semiotica*, 3:37–38.

Ekman, P. and Rosenberg, E. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* Oxford University Press Series in Affective Science, New York: Oxford.

Erickson, D. (1998). The effects of contrastive emphasis on jaw opening. *Phonetica*, 55:147–169.

Erickson, D. (2002). Articulation of extreme formant patterns for emphasised vowels. *Phonetica*, 59:134–149.

Erickson, D., Fujimura, O., and Pardo, B. (1998). Articulatory correlates of prosodic control: Emotion and emphasis. *Language and Speech*, 41(3-4):399–417.

Fasel, B. and Luettin, J. (2003). Automatic facial expression analysis: Survey. *Pattern Recognition*, 36(1):259–275.

Ferrigno, G. and Pedotti, A. (1985). Elite: A digital dedicated hardware system for movement analysis via real-time TV signal processing. In *IEEE Transactions on Biomedical Engineering, BME-32, II*, pages 943–949.

Field, A. (2005). *Discovering Statistics using SPSS.* SAGE Publications, London, second edition.

Foster, M. E. (to appear). *Elegant variation: Non-default choice in generation systems.* PhD thesis, The University of Edinburgh.

Fridlund, A. J. (1997). The new ethology of human facial expressions. In Russell, J. A. and Fernández-Dols, J. M., editors, *The Psychology of Facial Expression*, pages 103–129. Cambridge University Press.

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

Graesser, A., Lu, S., Jackson, G., Mitchell, H., Ventura, M., Olney, A., and Louwerse, M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers*, 36:180–193.

Grammer, K., Schiefenhovel, W., Schleidt, M., Lorenz, B., and Eibl-Eibesfeldt, I. (1988). Patterns of the face: the eyebrow flash in crosscultural comparison. *Ethology*, 77:279–299.

Granström, B., House, D., and Lundeberg, M. (1999). Prosodic cues in multimodal speech perception. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS'99)*, pages 655–658, San Francisco.

Grosz, B. and Hirschberg, J. (1992). Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing*, volume 1, pages 429–432, Banff.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Gustafson, J., Bell, L., Beskow, J., Boye, J., Carlson, R., Edlund, J., Granström, B., House, D., and Wirén, M. (2000). AdApt - a multimodal conversational dialogue system in an apartment domain. In *Proceedings of ICSLP'00*, volume 2, pages 134–137, Beijing.

Halliday, M. (1967a). *Intonation and grammar in British English*. Mouton, The Hague.

Halliday, M. (1967b). Notes on transitivity and theme in English (part II). *Journal of Linguistics*, 3:199–244.

Hewes, G. (1973). Primate communication and the gestural origin of language. *Current Anthropology*, 14:5–24.

Hirschberg, J. and Grosz, B. (1992). Intonational features of local and global discourse structure. In *Proceedings of the Speech and Natural Language Workshop*, pages 441–446, Harriman, New York. DARPA.

Houghton, G. (1986). *The production of language in dialogue: a computational model*. PhD thesis, University of Sussex.

Houghton, G. and Isard, S. (1987). Why to speak, what to say and how to say it: modelling language production in discourse. In Morris, P., editor, *Modelling Cognition*, pages 249–267. Wiley, Chichester.

House, D., Beskow, J., and Granström, B. (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. In *Proceedings of Eurospeech 2001*, pages 387–390.

Izard, C. E. (1971). *The face of emotion*. Appleton-Century-Crofts, New York.

Izard, C. E. (1979). *The maximally discriminative facial movement coding system (MAX)*. University of Delaware, Computer and Network Services, University Media Services, Newark, DE.

Izard, C. E. (1997). Emotions and facial expressions: A perspective from differential emotions theory. In Russell, J. A. and Fernández-Dols, J. M., editors, *The Psychology of Facial Expression*, pages 57–77. Cambridge University Press.

Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., Auer, E., and Berstein, L. (2003). Optical phonetics and visual perception of lexical and

phrasal stress in english. In *Proceedings 16th International Conference of the Phonetic Sciences*, pages 2071–2074, Barcelona, Spain.

Kendon, A. (1972). Some relationships between body motion and speech: An analysis of an example. In Siegman, A. and Pope, B., editors, *Studies in dyadic communication*, pages 177–210. Pergamon, Elmsford, New York.

Kendon, A. (1975). Some functions of the face in a kissing round. *Semiotica*, 15(4):299–334.

Kendon, A. (1980). Gesticulation and speech: two aspects of the process of utterance. In Key, M. R., editor, *The Relationship of Verbal and Nonverbal Communication*, pages 207–227. Mouton.

Kendon, A. (1997). Gesture. *Annual Review of Anthropology*, 26:109–128.

Kopp, S., Tepper, P., and Cassell, J. (2004). Towards integrated microplanning of language and iconic gesture for multimodal output. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, Penn State University, State College, PA.

Kopp, S., Tepper, P., Ferriman, K., and Cassell, J. (to appear). Trading spaces: How humans and humanoids use speech and gesture to give directions. *Spatial Cognition and Computation*.

Krahmer, E., Ruttkay, Z., Swerts, M., and Wesselink, W. (2002a). Perceptual evaluation of audiovisual cues for prominence. In *Proceedings of ICSLP*, pages 1933–1936, Denver.

Krahmer, E., Ruttkay, Z., Swerts, M., and Wesselink, W. (2002b). Pitch, eyebrows and the perception of focus. In *Proceedings of Speech Prosody 2002*, pages 443–446, Aix en Provence, France.

Krahmer, E. and Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication*, 34:391–405.

Krahmer, E. and Swerts, M. (2004). More about brows: A cross-linguistic study via analysis-by-synthesis. In Ruttkay, Z. and Pelachaud, C., editors, *From brows to trust: Evaluating embodied conversational agents*. Kluwer Academic Publishers.

Krauss, R. M., Chen, Y., and Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In Zanna, M., editor, *Advances in experimental social psychology*, pages 389–450. Academic Press, Tampa.

Ladd, D. R. (1996). *Intonational Phonology*. Cambridge University Press.

Ladd, D. R. and Schepman, A. (2003). "Sagging transitions" between high pitch accents in English: experimental evidence. *Journal of Phonetics*, 31:81–112.

Lester, J. C., Stone, B. A., and Stelling, G. D. (1999a). Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments. *International Journal of User Modeling and User-Adapted Interaction*, 9(1-2):1–44.

Lester, J. C., Towns, S. G., and Fitzgerald, P. J. (1999b). Achieving affective impact: Visual emotive communication in lifelike pedagogical agents. *International Journal of Artificial Intelligence in Education*, 10:278–291.

Levy, E. T. and McNeill, D. (1992). Speech, gesture and discourse. *Discourse Processes*, 15:277–301.

Loehr, D. P. (2004). *Gesture and intonation*. PhD thesis, Georgetown University, Washington, DC.

Massaro, D. W. (2002). Multimodal speech perception: A paradigm for speech science. In Granström, B., House, D., and Karlsson, I., editors, *Multimodality in language and speech systems*, pages 45–71. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Massaro, D. W. and Light, J. (2004). Using visible speech for training perception and production of speech for hard of hearing individuals. *Journal of Speech, Language, and Hearing Research*, 47(2):304–320.

McClave, E. (1998). Pitch and manual gestures. *Journal of Psycholinguistic Research*, 27(1):69–89.

McClave, E. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264:746–748.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

McNeill, D., Quek, F., McCullough, K., Duncan, S., Furuyama, N., and Bryll, R. (2001). Catchments, prosody and discourse. *Gesture*, 1:9–33.

Miles, J. and Shevlin, M. (2001). *Applying Regression & Correlation*. SAGE Publications, London.

Morton, H., McBreen, H., and Mervyn, J. (2004). Experimental evaluation of the use of ECAs in eCommerce applications. In Ruttkay, Z. and Pelachaud, C., editors, *From brows to trust: Evaluating embodied conversational agents*, pages 293–321. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Myers, R. H. (1990). *Classical and Modern Regression with Applications*. Duxbury Press, Boston.

Nakatani, C., Hirschberg, J., and Grosz, B. (1995a). Discourse structure in spoken language: Studies on speech corpora. In *Working Notes of the AAAI-95 Spring Symposium on Empirical Methods in Discourse Interpretation*, pages 106–112, Palo Alto, CA.

Nakatani, C. H., Grosz, B. J., Ahn, D. D., and Hirschberg, J. (1995b). Instructions for annotating discourses. Technical Report TR 21–95, Center for Research in Computing Technology, Harvard University, Cambridge, MA.

Neter, J., Wasserman, W., and Kutner, M. (1996). *Applied linear statistical models*. Chicago, London: Irwin, 4th edition.

Pantic, M. and Rothkrantz, M. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445.

Perlin, K. (1995). Real time responsive animation with personality. *IEEE Transactions on Visualization and Computer Graphics*, 1(1):5–15.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. PhD thesis, MIT. Published in 1988 by IULC.

Pitrelli, J., Beckman, M., and Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the International Conference on Spoken Language Processing*, volume 1, pages 123–126, Yokohama, Japan.

Power, R. (1974). *A computer model of conversation*. PhD thesis, University of Edinburgh.

Power, R. (1979). The organization of purposeful dialogues. *Linguistics*, 17:107–152.

Quek, F., Xin-Feng, M., and Bryll, R. (1999). A parallel algorithm for dynamic gesture tracking. In *ICCV'99 International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, pages 64–69, Corfu, Greece.

Rickel, J. and Johnson, W. (2000). Task-oriented collaboration with embodied agents in virtual worlds. In Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., editors, *Embodied conversational agents*, pages 95–122. The MIT Press, Cambridge, MA.

Rosenberg, E. L. (1997). Introduction: The study of spontaneous facial expressions in psychology. In Ekman, P. and Rosenberg, E., editors, *What the face*

reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS), Series in Affective Science, pages 3–17. Oxford University Press.

Russell, J. A. and Fernández-Dols, J. M. (1997). What does a facial expression mean? In Russell, J. A. and Fernández-Dols, J. M., editors, *The Psychology of Facial Expression*, pages 3–30. Cambridge University Press, Cambridge.

Ruttkay, Z. and Pelachaud, C., editors (2004). *From brows to trust: Evaluating embodied conversational agents*, volume 7 of *Human-computer interaction series*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Siegel, S. and Castellan, N. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, second edition.

Silverman, K., Beckman, M. E., Pitrelli, J., Ostendorf, M., Wightman, C., and Price, P., e. a. (1992). Tobi: a standard for labeling english prosody. In *Proceedings, Second International Conference on Spoken Language Processing*, volume 2, pages 867–870, Banff, Canada.

Srinivasan, R, J. and Massaro, D. W. (2003). Perceiving prosody from the face and voice: distinguishing statements from echoic questions in English. *Language and Speech*, 46(1):1–22.

Tian, Y., Cohn, J. F., and Kanade, T. (2003). Facial expression analysis. In Li, S. Z. and Jain, A. K., editors, *Handbook of face recognition*. Springer, New York.

Tomkins, S. S. (1962). *Affect, imagery, consciousness*. Springer, New York.

Tomkins, S. S. and McCarter, R. (1964). What and where are the primary affects? Some evidence for a theory. *Perceptual and Motor Skills*, 18:119–158.

Trager, G. L. and Smith, H. L. (1951). *An outline of English structure*. Battenburgh Press, Norman, OK. Reprinted 1957 by American Council of Learned Societies, Washington.

Wagner, H. L. (1997). Methods for the study of facial behavior. In Russell, J. A. and Fernández-Dols, J. M., editors, *The Psychology of Facial Expression*, pages 31–54. Cambridge University Press, Cambridge.

Wilbur, R. B. (1994). Eyeblinks and ASL phrase structure. *Sign language studies*, 84:221–240.

Wiltshire, A. (1999). *Synchrony of body motion with speech: language embodied*. PhD thesis, University of New Mexico, New Mexico.