L701: Accommodation to Speech Production in Task-Related Dialogue

Helen Robson

A thesis submitted in fulfilment of the requirements for the degree of MSc Speech and Language Processing at the University of Edinburgh

2008

Declaration

This thesis has been composed by and it has not been submitted in any previous application for a degree. The work reported within was executed by myself, unless otherwise stated.

August 2008

Acknowledgements

This project was funded by EU Project JAST (FP6-003747-IP).
I offer my thanks to many people: To Ellen Gurman Bard for her supervision and enthusiasm for this topic. To Robin Hill, for his additional technical support. Also, thanks must go to Jonathan Kilgour and Jean Carletta for their patience and help with the NITE NXT software used within this project, as well as to all of the computing support staff within the Linguistics department, and to Pete Jones for his help in de-bugging many, many bash scripts.
Finally, thanks must go to Paul Foulkes and John Local at the University of York for encouraging my interest in speech production and perception.

Contents

L701: Accommodating to Speech Production in Task-Related Dialogue

Abstract

Using the experimental setting of a joint construction task where visual contact between the two members of dyad is blocked, the communicative channels of eye-track and mouse-track are studied in order to ascertain whether their presence reduces the need for distinct introductory mentions of a referent; a phenomenon which has been found in the visual communication channel, as reported by Anderson et al (1997). A series of multiple regressions on the absolute durations of introductory and second identical mentions, and on the duration differences between corresponding first and second tokens reject this hypothesis, showing that the presence of one extra communicative channel reduces the overall length of tokens, two extra channels increases the overall length of tokens, but showing no effect of communication channel on the difference in duration between first and second tokens. The results are discussed and explanations for these findings are put forward.

1. Introduction

The academic study of speech began in approximately 450 B.C. within Aristotle's school of thought, and since then it has remained a cornerstone in linguistic study. This study deals with the concept of successful communication, and the method in which two people in conversation are able to have faith in the fact that they are talking about, and referring to the same things. As Brown and Dell (1986) explain, two individuals in conversation are likely to share the same processes of speech comprehension and speech production, and therefore if something is easy for a person to understand, it is easy for another person to produce.

This investigation uses the framework of a joint construction task (Carletta, Nicol, Taylor, Hill, de Ruiter, & Bard, under revision) in order to study introductory and second referring expressions used to designate referents within the task, and their relative durations and accessibility. The task itself takes the form of joint construction of a tangram by dyads using networked computers.

Firstly, an overview of the theory of accessibility, the notion of givenness, and the referential process are presented, along with examples of studies which look at phonetic influences on the intelligibility of speech and the influence of communicative context. The method of the current study is then outlined, along with five experimental hypotheses and observations on

the raw data collected. This investigation concludes with detailed results, and discussion of the outcome within the current framework of speech production theory.

2. Background

2.1 Accessibility theory and the notion of givenness

In any written or spoken discourse, the speaker must use the appropriate referring expressions to enable the listener to identify each entity mentioned within. This simple idea is the basis of accessibility theory (Ariel, 1990; 2001). Accessibility theory can be loosely described as the idea that each referring expression within a discourse should allow the listener to access a piece of information which is already stored within his or her memory. This familiar information can be described as a piece of *given* information. The speaker indicates how accessible this piece of given information is to the listener during discourse, and this level of accessibility can be determined from the context of the discourse. Ariel (1990; 2001) proposes a hierarchy of accessibility markers for noun phrases (NPs), with full names and modifiers; long definite descriptions; and distal demonstratives and modifiers classed as lying at the low accessibility end of the scale, and unstressed pronouns, verbal person inflections and zero markers classed as being high accessibility markers. (A full hierarchy can be found in Ariel, 2001:31.) Ariel stresses that the level of accessibility of the entities referred to within the discourse do not necessarily depend on their *physical* salience, but instead on their salience within the discourse. The more a referent is referred to within a discourse, the higher its accessibility. The criteria involved in determining the level of accessibility of a referring expression are *informativity* (the amount of lexical information it contains), *rigidity* (the level of ease with which a unique referent can be assigned to the referring expression), and *attenuation* (the phonological size of the referring expression). These criteria are not necessarily discrete, and can overlap.

Once the referent has been established, its level of accessibility can decrease throughout a discourse. If there is a large relative distance between the previous and current mention of the referent, its level of accessibility decreases. This distance does not necessarily depend on the number of words in-between the two referring expressions, nor is it necessarily temporal; it has been shown that episode boundaries can create a distance (see Ariel, 1990 for evidence).
Terken & Hirschberg elaborate further on this theory, discussing the phenomenon of *deaccentuation*. Terken & Hirschberg define deaccentuation as “the absence of intonational prominence on a referring expression” (1994:125). In English, accent is characterised by a change of pitch in or near to the lexically stressed syllable of the word. (Pierrehumbert, 1980).

Terken & Hirschberg state that deaccentuation is a direct consequence of the level of givenness of the entity being referred to, and that therefore words and phrases which do not

convey any new information to the listener are deaccented. These words and phrases may be repetitions of their earlier presence within a discourse, or it may be possible to merely infer their referent from the information previously gathered during the discourse.

However, this simple explanation of the phenomenon of deaccentuation does not explain cases where given information is accented. In order to explain this exception to the general rule, Terken & Hirschberg study the relative contributions of grammatical role and surface position of the referring expression to the phenomenon. This area has been studied in detail (see Terken & Hirschberg, 1994 for an overview of previous studies), and the experimental evidence collected suggests that the accessibility of a discourse entity directly affects its assignment of both grammatical role and surface position, in that a highly accessible entity is likely to be realised as the grammatical subject of a sentence, and is often realised in a reduced form, such as a pronoun. Terken & Hirschberg conducted an experiment to ascertain to what extent givenness, surface position, and grammatical role affect the accent status of a referring expression. The results of their experiment show that although givenness is not a sufficient explanation for deaccentuation, its interaction with both grammatical role and surface position does provide an explanation for the phenomenon. Therefore, if the grammatical role and surface position of a referring expression is identical to its antecedent, it is likely to be accented. The study also notes that speaker variability is evident in the data, and suggests that different speakers may employ different strategies for deaccentuation.

2.2 The referring process

Conversation is an inherently collaborative process (Grice, 1957). Clark & Wilkes-Gibbs (1986) propose a model to prove this theory, showing that both speakers and listeners work collaboratively in order to establish an agreed definite referent. In their model, the process is always initiated by the speaker, who either invites or introduces a referring expression in the form of a noun phrase. Once the noun phrase has been introduced, both the participants in the discourse can repair, expand upon, or replace it in an iterative process, until the referent has been established. This model therefore comprises of two stages: *presentation* and *mutual acceptance*. The listener can either actively declare their acceptance of the referring expression with affirmative phrases and/or paralinguistic gestures, or the speaker can presuppose the listener's acceptance and immediately continue. In order to test this model, Clark & Wilkes-Gibbs conducted a matching experiment, hypothesising that initial mentions of the objects to be matched (in this case, tangrams) should involve a relatively lengthy presentation and mutual acceptance process, whereas later mentions should be shorter and

immediately mutually accepted. The experimental results support these hypotheses, as the average number of words used to refer to each tangram decreased as the number of trials increased, and the average number of speaking turns per tangram also decreased as trials progressed.

The model proposed by Clark & Wilkes-Gibbs has to be modified by including the principle of least effort (originally suggested by Zipf, 1935). Due to this, the model is adapted to presuppose that speakers and listeners will try to minimise their collaborative effort, whilst still attempting to establish the mutually correct referent. This addition to the model allows it to explain the replacement of noun expressions, as time-pressure, complexity and ignorance create a trade-off situation between these issues and the effort of creating the initial noun phrase. The principle of least effort also adapts the model to allow for the fact that the speaker and listener are not striving for perfect understanding of every utterance, but instead they are only aiming for understanding "to a criterion sufficient for current purposes" (Clark & Wilkes-Gibbs, 1986:35).

However, this view of the referring process has been challenged. Gundel, Hedberg, & Zacharski (1993) present a different theory; proposing the idea that different determiners and referents signify different cognitive statuses, i.e. the location of the referent in memory, therefore allowing the listener to restrict the number of possible referents. Gundel et al. suggest the following Givenness hierarchy shown in figure (i) below:

Figure (i): The Givenness Hierarchy proposed by Gundel et al. (1993:275)

in focus	>	activated	>	familiar	>	uniquely identifiable	>	referential	>	type identifiable
{*it*}		{*that* *this* *this* N}		{*that* N}		{*the* N}		{indefinite *this* N}		{*a* N}

The significant property of this hierarchy is that each cognitive status in the hierarchy necessarily entails all the lower level statuses to the right of it. This property establishes a crucial difference between the Givenness Hierarchy and the Familiarity Scale suggested by Prince (1981b), in which each level is mutually exclusive.

2.3 Referential form and word duration

This model of the process of establishing referents in a discourse can also be extended to include phonetic phenomena, such as the ways in which speakers can signal "new" and "old" word in their speech. Bolinger (1981) asserts that speakers tend to lengthen words which are either unusual in context (and therefore have a low level of accessibility) or which are in an uninformative context. Fowler & Housum (1987) extend these observations to suggest that speakers aim to produce acoustic signals for words which are informative enough that the listener is able to recognise the words. Therefore, if the word is likely to be found in its context, the acoustic signal produced by the speaker may be a reduced and less-informative version. It is suggested that speakers produce these attenuated versions of words whenever possible; specifically when it will not affect the ease of identification of the word. This is coupled with the idea that speakers can only attenuate words if their identity can be determined at least partly by other information provided by the context. Fowler (1988) uses these findings to suggest a link between this phenomenon of attenuation and the fact that reductions and elisions are found more often in casual speech styles as opposed to more formal speech styles.

Fowler & Housum tested this hypothesis by conducting a series of experiments looking at the duration and intelligibility of first and second mentions of words. The results of this experiment showed that second mentions of words were significantly shorter than first mentions of words, and when presented to participants in isolation, were more difficult to identify. However, when the words were presented in context, no such loss of intelligibility was found. Fowler & Housum then went on to test whether participants were able to identify words as "old" or "new" when presented with a series of words out of context. The results of this experiment showed that listeners were able to identify the difference between first and second mentions, and were able to use this information to facilitate the retrieval of the word's prior context.

Fowler (1988) extended this study by conducting a series of experiments looking at which conditions facilitate durational shortening in repeated tokens of words. Fowler mentions the fact that repetition effects may affect the findings of previous experiments, as motor effects may play a part in the production of reduced forms of words, as afferent feedback could aid the movement of the articulators used in creating each sound (Perkell, Guenther, Lane, Mathies, Perrier, Vick, Wilhelms-Triarico, & Zandipour, 2000).

The series of experiments presented the following findings: words do not undergo durational shortening when they are produced in a list, although shortened forms of the same words are produced when they are uttered in the context of meaningful prose. Another finding was that words do not undergo any shortening when they are preceded by a homophone. This rejects the hypothesis that repetition tasks may play a part, as identical articulatory trajectories are employed to produce the homophones. This finding is backed up by the research carried out by Bard, Brew & Cooper (1991) whose experimental findings indicate that durational shortening only occurs on the repetition of a word as long as both words refer to the same referent.

In regards to speaker design, Bard, Sotillo, Anderson, Doherty-Sneddon, & Newlands (1995) adapted the idea of speakers designing their utterances specifically towards the needs of the listener, by presenting experimental evidence suggesting that the speaker's design is principally egocentric, as speakers tended to produce significantly less intelligible versions of words which are repeated from earlier in a discourse, even if the listener was not at hand to hear the first production of the word.

Fowler, Levy & Brown (1997) continued to study this phenomenon, by looking at the role of episode boundaries with regards to both durational shortening of repeated tokens, and the shortening of words at the lexical level. The study involved participants watching an episode of a television series, before recounting the storyline of the episode to a participant who had not viewed the film. These narrations were then segmented into episodes, signified by mention of a scene change or similar. Repeated tokens (specifically the names of characters within the film) were then extracted from the narrations; half with the repeated token occurring within an episode, and half with the repeated token occurring in adjacent episodes. The results of this study show that the appearance of episode boundaries between repeated references to an entity tend to restrict the use of shorter referring expressions, although the referent should still be highly accessible to the listener, and the same general pattern is also found with regards to durational shortening of repeated tokens. In a second experiment, half of the speakers narrated the film to a listener for half the number of total trials, before continuing to narrate the film to a new listener. Fowler et al. reported the same pattern of results with regards to durational shortening, irrespective of whether the new listener had heard the first mention of a word or not.

These results are compatible with the results presented by Bard et al. (1995) which are themselves expanded upon in the study of intelligibility of referring expressions by Bard, Anderson, Sotillo, Aylett, Doherty-Sneddon & Newlands, (2000) and Bard & Aylett, (2005).

In these studies, Bard et al. conducted a series of experiments to determine whether adjustments in intelligibility in spontaneous speech are based on a model of the listener's knowledge. However, the results of the experiments showed that the listener's knowledge had no significant effect on intelligibility, and that only the speaker's knowledge showed an effect. Bard et al. (2000) explained these results in terms of a Dual Process Model of speech production, where the main process of word priming is solely based on the knowledge of the speaker; with optional slower inferences being drawn from the listener's model, explaining the fact that speakers seem to be insensitive to subtle aspects of the listener's knowledge, such as whether the listener can or cannot see the referent of a referring expression (Bard & Aylett, 2005).

This model of speech production also asserts that priming is triggered by the givenness of a word. This claim is backed up by previous findings of a study conducted by Bard & Anderson (1994) which showed evidence that introductory mentions in a dialogue of an object which is physically present, tend to articulated more quickly than introductory mentions of new entities. This finding is also used to explain the phenomenon of second mentions of words within a discourse being attenuated, regardless of whether both mentions were uttered by the same speaker or not.

2.4 Phonetic influences on the intelligibility of speech

It is a well-documented fact that the intelligibility of a word is directly influenced by its phonetic structure and the level of accent bestowed upon it (Huttenlocher & Zue, 1984, Stevens, 2002). However, Hawkins & Warren (1994) suggest that these factors are the main influence of the intelligibility of speech, rather than the givenness of a word in discourse, or repeated mentions of a word. Hawkins & Warren also go on to suggest that segmental differences in the phonetic structure of a word do not affect the intelligibility of all words equally, and point out that the relative influence of each phonetic segment is dependent on the structure of the word and its potential competitor words: if the phonetic segment is located after the word's uniqueness point, i.e. after the point where the word is distinguishable from every other word in the language, then the relative contribution of that phonetic segment to the intelligibility of the word is marginal (Pisoni & Goldfinger, 1990). The variability of an individual's speech is also noted as a potential influence on a word's intelligibility. Hawkins & Warren support these claims with a series of identification tasks using both words and CV segments from words taken from conversational speech. The results of these experiments show that word repetition alone does not have a significant effect on intelligibility, but that

local phonetic features of accent and place/manner of articulation are the factors which contribute most to the intelligibility of CV segments and entire words. Hawkins & Warren conclude that these phonetic features are the most important factor in intelligibility, stating that accented words are more intelligible, irrespective of their status of being first or second mentions of a word.

2.5 The influence of communicative context on speech

The intelligibility of speech is also affected by communicative context. Anderson, Bard, Sotillo, Newlands, & Doherty-Sneddon (1997) tested the experimental hypothesis that speakers produce more casual speech when they are in face-to-face communicative context, by comparing the intelligibility of introductory mentions of new referents produced by dyads who were able to be in visual contact, which those produced by dyads who were unable to see each other. The results of their initial experiment support this hypothesis, as the speech produced in the added-visual contact condition was significantly less intelligible. In a further post-hoc study, Anderson et al. tested the hypothesis that visual cues were used by the listener in order to aid word recognition, therefore explaining the previous results. However, this study showed that speakers did not reduce intelligibility of these introductory mentions when looking at their interlocutor, but instead increased intelligibility in these conditions. Anderson et al. hypothesised that the reasons for this increased intelligibility are that the added visual channel is used by the speaker as a means of checking the comprehension of the listener: speakers increase their clarity when comprehension issues are apparent, and decrease intelligibility when the added visual channel is available to them, as it is an efficient means of gauging understanding.

This theory is supported by further studies by Doherty-Sneddon, Anderson, O'Malley, Langton, Garrod & Bruce (1997), and Monk & Gale (2002). Doherty-Sneddon et al. show that speakers used the visual channel in order to check the state of the interaction, as participants in their experiment who were only able to use the audio channel elicited significantly more auditory feedback from their partner.

Video-mediated dialogues are becoming more and more commonplace in the modern world, and these contexts provide a new environment for a type of face-to-face communication. Anderson & Howarth (2002) studied this phenomenon with respect to word duration, the referential form used, and also the response to cognitive load. Sweller (1988) describes cognitive load as the amount of "mental energy" needed in order to process a certain piece of

information. In order to control for this in an experimental setting, Anderson & Howarth manipulate the time allowed to complete a map task, therefore increasing cognitive pressure. Previous studies by Horton & Keysar (1996) and Rosnagel (2000) have shown that referential forms are often reduced when cognitive load is increased in an experimental setting.

Anderson & Howarth found that speakers articulated words more slowly overall when participating in a video-mediated dialogue, but still continued to articulate the second mentions of words more quickly than the first mentions. Another interesting finding from this study was a greater number of shorter referring expressions such as pronouns were used in the video-mediated dialogue when cognitive load was increased. Anderson & Howarth explain these results within the framework of the Dual Process Model of speech production outlined above (Bard et al., 2000), and suggest that the slower articulation of words in a video-mediated context could be due to a greater communicative distance between speakers, resulting in hyperarticulation.

Brennan (2005) examined other communicative contexts by presenting an experiment for pairs of participants in which they complete spatial matching tasks on networked computers, in the form of map tasks. In half of the trials, the dyads were able to speak to each other, and the director of the dyad could also see his partner's mouse cursors on his screen, and in the other half of the trials the dyads had to rely on dialogue alone. Brennan predicted that in the trials with the additional visual evidence of the mouse cursors, the collaborative referring process as suggested by Clark & Wilkes-Gibbs (1986) should be significantly faster.

The results of this experiment support this hypothesis, as the most efficient trials were those where additional visual evidence was available to the director of the dyad, and that the dyads used less than half the number of average words used in the verbal-only trials.

Brennan then goes on to discuss the relative strengths of tracking eye-gaze movements and mouse movements in referential communication tasks. Eye-tracking has been used in many studies of spoken language due to the "eye-mind" assumption, which Just & Carpenter (1980:331) express as the idea that "that there is no appreciable lag between what is being fixated [upon by the eye] and what is being processed". An advantage of eye-tracking is that it is much more temporally precise than mouse-tracks, as saccades are executed much quicker than movements of the mouse. However, irrelevant saccades are often made, and Brennan suggests that in complex map tasks, saccades do not necessarily denote potential referents during the referential process, but instead may merely indicate that the participant is gathering information about the task as whole. This is in contrast to mouse movements, which although

slower, are more directly linked to the participant's intentions and hypotheses about potential references.

Bard, Hill, & Foster (2008) use both the eye-track and mouse-track modalities of communication in a recent study which looks specifically at the level of accessibility of introductory mentions to shapes in a joint construction task to create tangrams, performed by dyads on networked computers. Half of the dyads were assigned Manager-Assistant roles, and half were assigned no role. The study examines these differing communication modalities as well as any action involving the referent which is being performed by the either member of the dyad. The results of this experiment showed that only 16% of introductory expressions were indefinite noun phrases, whilst 84% were phrases of higher accessibility. Further analysis proved that the communication modalities available to the dyad affected the results: if the referent was visibly being moved on the dyad's screens, then the number of deictic referring expressions increased, whilst the number of indefinite noun referring expressions decreased. The number of deictic referring expressions also increased when a mouse cursor was hovering over the referent. The analysis also found that actions which were invisible to the listener of the dyad were also significant, as there was a shift from the use of indefinite referring expressions towards definite referring expressions when a mouse cursor was hovering over the referent, even when this action was invisible to the other member of the dyad. The role assignment also proved to be significant, but only in the case of invisible mouse gestures: those dyads assigned Manager-Assistant roles tended to use definite noun phrases and deictics, as opposed to indefinite noun phrases. These results allow Bard et al. to hypothesise that it is the speaker's knowledge which is crucial, as opposed to that of the listener. No movements of the constituent shapes carried out by the listener were significant, even when these movements were visible to the speaker; a further finding which backs up this claim.

However, Bard et al. acknowledge some limitations in the analysis of these results. Firstly, the results attributed to the presence/absence of roles in the dyad are based on the speech of both the manager and assistant combined, as the amount of data for manager and assistant was not sufficient to be compared. The results also show that the results do not necessarily correspond with an accessibility *hierarchy*, and that perhaps there are other factors which are playing a part in the relative accessibility of a referring expression. The design of the experiment may also confound results, because as the experiment proceeds in the different conditions, the introductory expressions to shapes are not completely new as they have already been used in a previous condition of the experiment. Therefore the referents become somewhat predictable from the context of the experiment (Prince, 1981b).

2.6 Research hypotheses

Four specific hypotheses have been developed for testing on this data, which directly follow from the literature in Section 2. These predictions will be dealt with individually, and concern the durations of introductory and second referring expressions in general; whether the added communication channels of eye-gaze and mouse-track will affect the durations of introductory and second referring expressions; and whether the order of speech and non-speech parts of the trial will affect the durations of introductory and second referring expressions.

The starting point for this investigation was the study by Anderson et al. (1997) outlined above, which looked at the relationship between the intelligibility of introductory mentions of a referent, and the absence/presence of the visual channel of communication. In this task, the relationship between durations of introductory and second mentions of a referent and the added channels of communication of mouse-track and eye-track is explored.

The main hypothesis of this experiment is as follows: in a joint construction task where visual contact between the two members of a dyad is blocked, participants will use any added communication channel (in this case, the mouse-track and/or eye-track of their partner) in the same way as the visual channel in face-to-face dialogue in their introductory mentions of a referent; therefore reducing the difference in duration between first and second mentions of a referent.

As well as this main hypothesis, there are the following secondary hypotheses:

The duration of introductory mentions of referents will be significantly shorter than second mentions of referents in all initial conditions of the experiment, following the findings of Fowler & Housum (1987).

The greater the number of communication channels available to each dyad, the shorter the durations of both the introductory and the second mentions of the referent.

If the dyad has encountered the non-speech part of trial before the speech part of the trial, the lesser the need for distinct first tokens, therefore the shorter the duration of both the introductory and second tokens.

3. Method

3.1 Participants

72 students at the University of Edinburgh were paid to participate in this experiment. These students were then assigned a partner of the same sex, creating 36 same-sex dyads. The members of each dyad did not know each other prior to this experiment. 4 of the dyads had to be discarded due to technical failures, leaving 32 same-sex dyads taking part in the experiment.

3.2 Apparatus

Each member of each dyad in this experiment sat approximately 40cm away from a CRT computer monitor in a sound-attenuated room. These monitors were networked together and were set up so that the participants faced each other, although direct eye-contact between the two was impossible due to the position of the monitors. SR-Research EyeLink II head-mounted eye-trackers were used to eye-track each participant monocularly. Connected to these eye-trackers were head-mounted microphones which recorded each participant's speech on a separate channel. Video recordings were made of all movements of constituent shapes, partially constructed tangrams and cursors on each participant's screen. These audio and video channels were also combined, creating a set of composite videos.

3.3 Materials

16 different target tangrams were created, none of which resembled a nameable entity. Each tangram was made up of 11 constituent shapes. The different constituent shapes are listed in Table 1 below:

Table 1: The constituent shapes used to create tangrams in the JCT

Size	Colour	Shape	Number Available
Small	Olive	Right-angled isoceles triangle	2
Small	Sand	Right-angled isoceles triangle	2
Medium	Red	Right-angled isoceles triangle	2
Large	Orchid	Right-angled isoceles triangle	2
Large	Cyan	Right-angled isoceles triangle	2
	Yellow	Parallelogram	1
	Magenta	Square	2

For each tangram, the set of 13 constituent shapes at the right of each participant's screen consisted of two small olive triangles, two small sand triangles, two medium red triangles, two large orchid triangles, two large cyan triangles, two magenta squares and a single yellow parallelogram. As each tangram consisted of 11 of these shapes, there were always two spare constituent shapes, which differed for every tangram.

3.4 Experiment design

Each of the 32 same-sex dyads built 2 tangrams in each of 8 experimental conditions, or 16 different tangrams in total. The experimental conditions were created by the factorial manipulation of following three channels of communication: speech, gaze (the current eye-track of each participant cross-projected onto the other participant's screen), and mouse (the current mouse-track of each participant cross-projected onto the other participant's screen). The speech and non-speech conditions were counter-balanced, and the gaze and mouse modalities within these conditions were pseudo-randomised using a latin square. Half of the dyads were assigned roles of manager and assistant in the task. Each manager was instructed to oversee the accuracy, speed and cost of each task, and also to decide when each task was complete, whilst the assistant was told to follow the instructions of the manager. The other 16 dyads were not given any specific roles.

As this project is concerned with the participants' dialogues, only the speech conditions of this experiment will be analysed. It should also be noted that each participant could always see their own mouse-track, but when the other participant's gaze and mouse were hidden, the only communication between the participants was through speech alone.

3.4 Task

The data used in this experiment was collected and coded as part of the Joint Construction Task (Carletta et al., under revision), henceforth referred to as JCT. This data was also used by Bard et al. (2008). The task takes the form of a collaborative game undertaken by 2 players on networked computers. The purpose of the game is to construct a target tangram out of its constituent geometric shapes. An example screen of the tangram task is shown in figure (i) below:

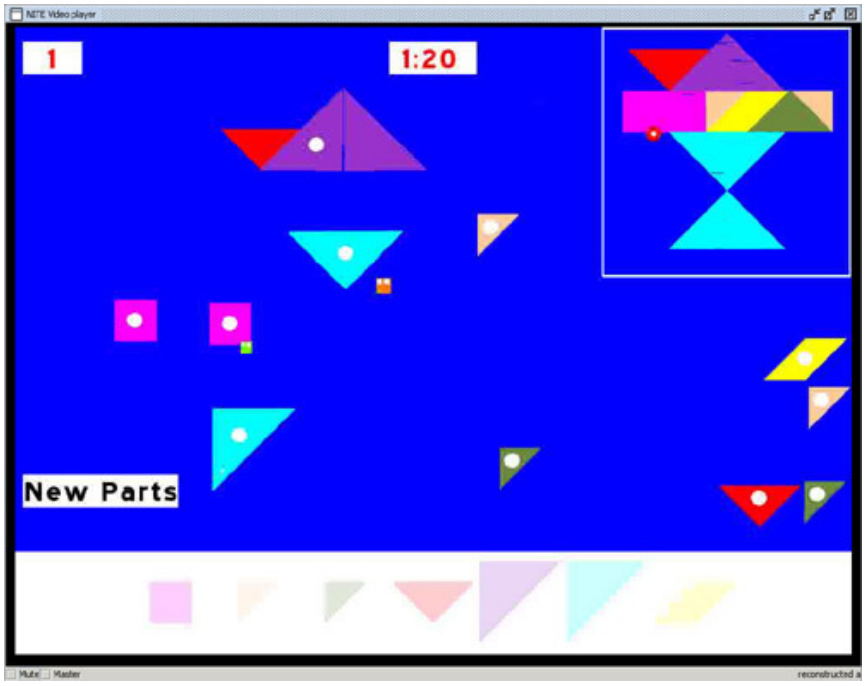

Figure (ii): an example of the JCT as seen on the participants' computer screens

As figure (ii) shows, each player's computer screen consists of the following elements: a target tangram (top right), a set of its constituent geometric shapes to be used in reproducing the target tangram (centre right), a set of replacement shapes to be used in case of breakages (bottom), a counter recording the number of breakages, a clock measuring the time elapsed during the task (top centre), and a clear work space (centre).

The aim of the JCT is to construct a target tangram as quickly and as accurately as possible in collaboration with a partner. The participants have 8 minutes in which to complete each tangram. In order to construct each tangram, constituent shapes must be moved into position and joined together. In order to do this, participants must move each shape by left-clicking on it with their mouse, and dragging it into position. Each shape can be rotated by right-clicking on it. However, if both participants simultaneously click on the same shape or partially-constructed tangram, it 'breaks' and disappears from the screen, and a replacement must be found in the spare parts area at the bottom of the screen. Breakages also occur when participants move one shape across another on the screen. When a mouse is holding a shape, it changes colour on the screen, allowing the players to differentiate between a mouse hovering over an object and one holding an object. Each participant's mouse cursor appears in a distinct colour on the computer screen.

Two shapes can be joined only if held by different players. As soon as two parts of the shapes meet, they are permanently joined together, therefore necessitating accurate collaborative

movements. If the participants decide that a join is not sufficiently accurate, the partially created tangram can be purposely broken. This rule creates a trade-off between speed of construction and accuracy of construction. Figure (ii) below shows a screenshot of a target tangram completed by a pair of participants:

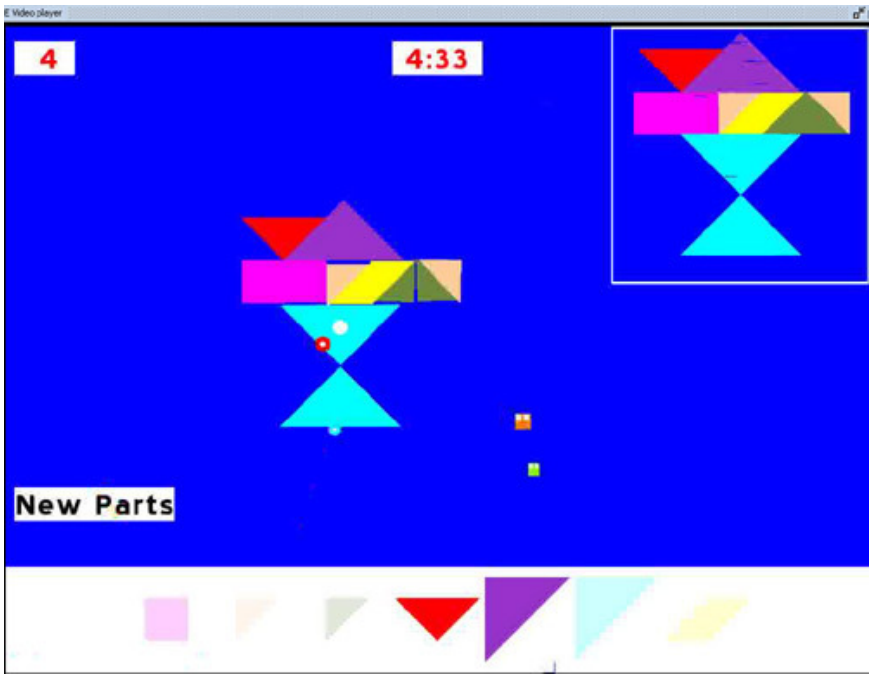

Figure (ii): Example of a replica (centre) of the target tangram (top right) constructed by participants

Each tangram is designated as complete when one participant presses the spacebar on the computer keyboard, and the 2nd participant confirms its completion by pressing their spacebar. The accuracy score for the replicated tangram is then displayed to both players, along with the total number of breakages per trial in the top left-hand corner of the screen.

4. Results

4.1 Analysis: All data

4.1.1 Finding repeated mentions

Each dialogue was transcribed orthographically by a team of coders, and each referring expression within this dialogue was then time-stamped for its start and endpoint. Using the NITE NXT software package (Carletta et al., 2005), each referring expression which related to an onscreen object was then coded for, and linked to a referent. The coders were free to use the video and audio files in order to ascertain which object each referring expression was denoted. If the object of reference was not certain, it was not linked to a specific referent, and was instead coded as 'uncertain'. A sample of the coded material as seen in NXT is shown in figure (iii) below:

Figure (iii): a sample dialogue coded using NXT

This project is concerned only with the first and second mentions of the constituent shapes used to create each tangram in each condition by each dyad.

The NXT software package, with local additions, made it possible to watch composite videos of dyads completing the JCT, together with audio and coded transcriptions. A sample screen of NXT during this process is shown in figure (iv) below:

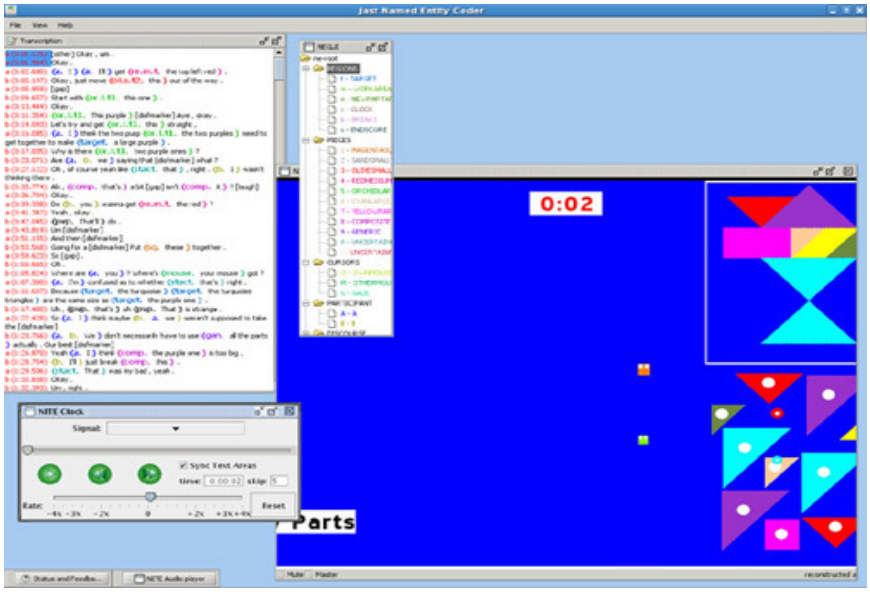

Figure (iv): a sample screen of the NXT software used during the data analysis

Firstly, I extracted all first and second referring expressions used to refer to each constituent shape used to make up the tangram, with the assumption that the initial mention of each constituent piece of the tangram by each dyad would state its colour, shape, and optionally also its position. However, upon analysis of the data, this pattern was not evident in the dialogue of any dyad. Instead, the initial referring expressions used to indicate each constituent shape were often pronominal forms such as "it", "this" and "mine".

In order to illustrate further the types of referring expressions utilised by each dyad, example (a) below is an extract of the dialogue of a dyad, who were able to see each other's eye-track on their respective screens, and the first and second mentions of different constituent shapes have been highlighted:

Example (a): Excerpt from the dialogue in the first condition encountered by dyad 02:

B: Okay yeah and I'll just always go straight up here, do you see where **mine**'s moving now? I'll just move **it** round there. And then we'll need to work together to sit **them** together.
A: Yeah, 'kay. This is hard. I don't know how to even-
B: It is weird.
B: Okay do you want me to- I'm gonna move **this yellow one**, okay. And then should we stick these **two blue ones** together?
B: Right, cool. We've got **those two** together. And shall I move **this**- I'll get **the yellow one**, okay?

As the above example shows, the referring expression used for the first mention of constituent shapes varies, with the pronominal forms "it" and "mine" used for small olive triangle and the fuller referring expression of "this yellow one" used to refer to the yellow parallelogram. On further inspection of the dialogue of each dyad, it is evident that the colour of each constituent shape is the significant piece of information used to differentiate between pieces, with the majority of dyads using the referring expressions "the [COLOUR] one" and "the [COLOUR]" in order to make a distinction between each piece.

The colour terms used by each dyad to refer to the constituent shapes are also a point of interest. There are eleven basic colour terms in English (Berlin & Kay, 1969), and six out of the seven constituent shapes used in the JCT can be referred to by using one of these terms. However, the constituent shape classed in the JCT as the Small Sand Triangle is not covered by any of these basic terms. The colour terms used by the participants to refer to each constituent shape in the JCT are shown in Table 3 below:

Table 2: The colour terms used by participants in the JCT to refer to each constituent shape. Colour terms outwith the eleven basic colour terms in English are shown in italics.

Constituent Shape	**Terms used in dialogues**
Small Olive Triangle	green
Yellow Parallelogram	yellow
Medium Red Triangle	red
Large Orchid Triangle	purple, blue
Large Cyan Triangle	blue, light blue, green
Magenta Square	pink, *fuchsia*
Small Sand Triangle	*beige*, *peach, tan, salmon, orange, flesh-coloured, pinkish*

As Table 2 shows, the dyads in this experiment tend to use basic colour terms, where possible. Basic colour terms are by definition monolexemic; therefore light blue is not strictly a basic colour term, but a subordinate hue of the basic colour term "blue". I observed that the three dyads which used the term "light blue" to refer to the Large Cyan Triangle used this more specific colour term in the first stages of the task, before reducing the referring expression to "blue" once the referent object had been identified correctly. One dyad used the basic colour terms "green" *and* "blue" to refer to this constituent shape, whilst also using the term "green" to refer to the Small Olive Triangle. It was noted that this ambiguity adversely affected this dyad's performance, as they struggled to make clear which specific constituent shape they were referring to.

The Magenta Square constituent shape was also referred to as “fuchsia”, which is classed as a descriptive colour term as opposed to a basic colour term, as its name is derived from an object in the real world. The Small Sand Triangle, which is not covered by any of the eleven basic colour terms in English, is referred to in a variety of ways by the dyads; the most common terms used are “beige”, “peach” and “salmon”. “Peach” and “salmon” are also classed as descriptive colour terms, whereas “beige” is classed as an abstract colour term, but too rare to be classed as a basic colour term. It was noted that the initial referring expressions used to refer to the Sand Small Triangle often tended to include qualifying expressions; two examples of this include “the kind of pink-ish thing” and “the peach-y thingy”, which is in direct contrast to the initial referring expressions used for consistent shapes covered by one of the eleven basic colour terms: for example, initial referring expressions for the Medium Red Triangle constituent shape tended to be more specific, such as “that red one” and “the red”.

Another issue found whilst analysing the raw data was that dyads tended to stop referring to each constituent shape as the experiment progressed, and as they became used to the experimental setting, they developed a successful technique for replicating the tangrams without the need for as much dialogue. The dyads which encountered the non-speech part of the experiment first also tended to not refer to each constituent shape, as they had already developed a successful method of constructing the tangrams without the need to use dialogue as an aid. The same issue was also found when dyads encountered the task with both communication channels of eye-track and mouse-track available to them. These extra communication channels aided the dyads to such an extent that they tended to no longer refer to the constituent shapes at all, or to immediately use reduced forms of referring expressions such as “it” or “this”. As Bard et al. (2008) showed, dyads to indicate the referent by either moving the referent, or hovering their mouse over it (irrespective of whether the listener was able to see the speaker’s mouse-track).

Due to these issues, the remaining data were not sufficient to fill each cell in a within-subjects analysis, as there weren’t a sufficient number of directly comparable referring expressions used by each dyad in each experimental condition. Therefore the analysis was adapted in order for it to fit the data, and this final experimental design is outlined in the following section.

4.1.2 Selection

In order to fill each cell of the design evenly, all mentions of each constituent shape per dyad were drawn from coded transcriptions. Out of the 32 dyads who participated in this experiment, 5 had to be discounted as no repeated mentions of referring expressions used to designate the constituent shapes were found within their dialogue in each experimental condition. All tokens containing speech disfluencies, plural forms, and all tokens consisting of pronominal forms and non-lexical items such as "this one" were also discounted from the analysis, in order to control the data as much as possible. Disfluencies such as "the red- uh blue" triangle" were omitted as the repair phase of the utterance is often found to be accented (Howell & Young, 1991), and this feature could affect the duration and level of intelligibility of the token.

The first and second tokens of referring expressions used to designate each constituent shape were not always directly comparable, as they often used the colour of the shape as an adjective and also as a noun, as shown in example (c) below:

Example (c): Excerpt from the dialogue in the second condition encountered by dyad 17:

A: Uh sure, I'll grab **the purple.**
B: Then for **the purple one** on top?

This is to be expected, as this experiment is analysing natural speech, as opposed to speech elicited under laboratory conditions. In order to overcome this issue, the first and second repeated referring expressions which did not include any speech disfluencies, non-lexical items, or pronominal forms were extracted. The number of other expressions which were used in between the two repeated mentions was noted, so that it could be controlled for in the analysis. Tokens were only extracted from the first two conditions encountered by each dyad, due to the lack of data for later conditions, as explained above.

The distribution of these tokens is shown in figure (v) below:

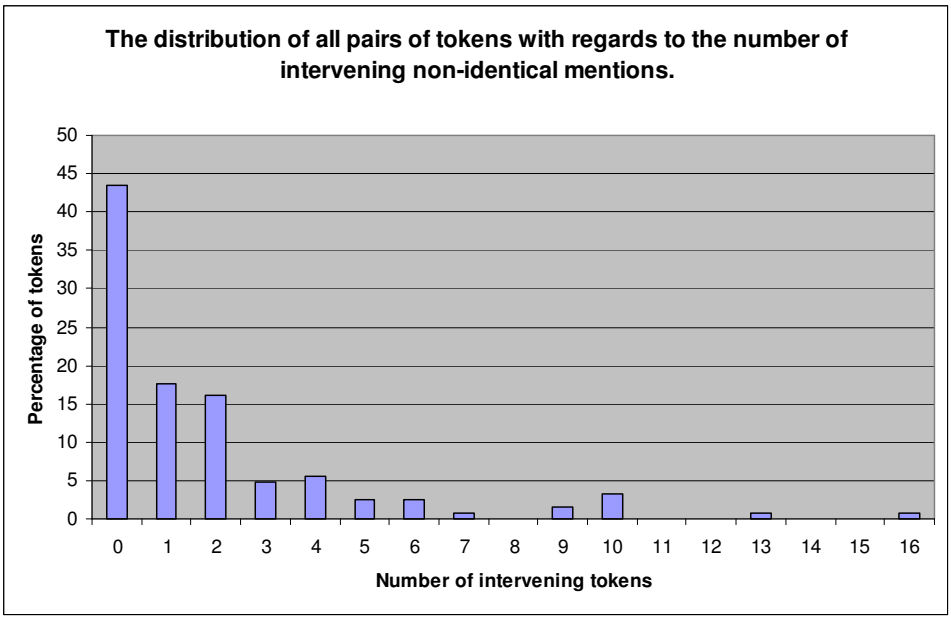

Figure (v): The distribution of all pairs of tokens with regards to the number of intervening non-identical mentions.

As this graph shows, 44% of the extracted repeated mentions are adjacent, but over 18% of the repeated mentions are separated by 1 mention, and 17% of the repeated mentions are separated by 2 mentions.

In total, 304 tokens were extracted, comprising of 152 first mentions and 152 second mentions. Within these 304 tokens, there were 25 different lexical items: 12 colours used as adjectives, 9 colours used as nouns, 2 shapes used as nouns, and 2 adjectives designating position on the screen.

4.1.3 Segmentation

These tokens were then segmented from the audio recordings using Praat speech software[1], following the guidelines drawn up by Turk, Nakai & Sugahara (2006). As Turk et al. state, the segmentation of speech is somewhat artificial, as the articulatory gestures used to create speech sounds overlap substantially when they are produced in succession. This is shown in figure (iv) below:

[1] http://www.praat.org

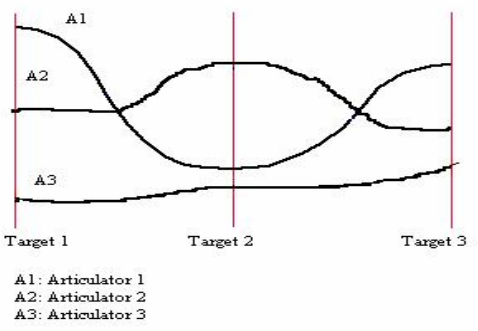

Figure (iv): an idealised diagram showing three tracks of separate articulators moving through three phoneme targets, following Browman & Goldstein (1992).

Turk et al. propose that stops, fricatives and affricates can be segmented by following the criterion of onset and release of oral constrictions, and that segmentation based on the criterion of voicing onset and offset, which is not necessarily interchangeable, is not as useful a criterion as it cannot be applied to as wide a number of sound classes. However, they point out in their paper, some classes of sound are harder to segment than others. Turk et al. suggest that oral stops, sibilant fricatives and affricates can be reliably segmented in all contexts; that nasal stops and weak voiceless fricatives can be reliably segmented in certain contexts; and that approximants and weak voiced fricatives should be avoided, if possible. However, as the data collected within this task is naturally occurring speech, as opposed to speech elicited in an experimental context, the segments designated as being difficult to segment reliably were impossible to avoid. The overall distribution of the tokens with regards to the classes of sounds found as initial and final segments of each token is shown in figure (vii) below:

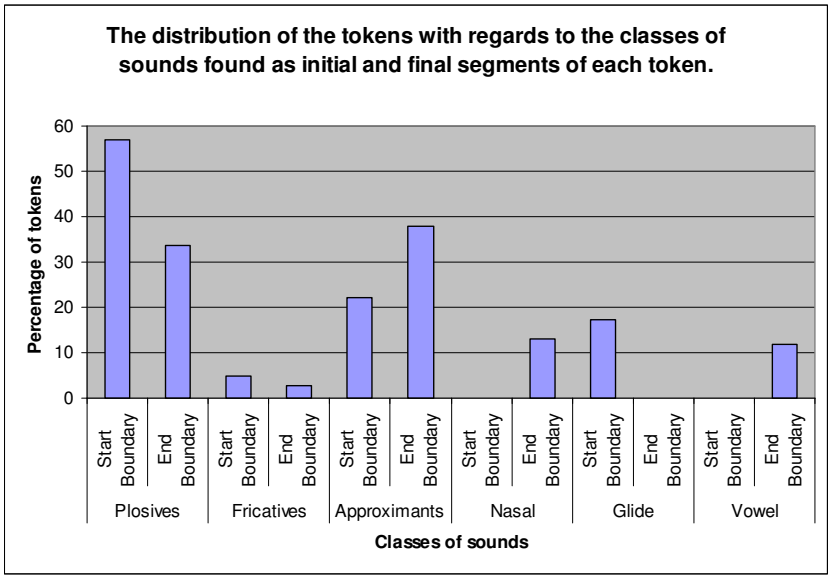

Figure (vii): The overall distribution of the tokens with regards to the classes of sounds found as initial and final segments of each token

As the graph in figure (vii) shows, 57% of the initial boundary segments of the tokens were plosives, which are relatively easy to segment reliably. However, 21% of the initial boundary segments were approximants, which are more difficult to segment. The final boundary segments of the tokens were also spread between 5 classes of sounds; the most frequent class of sounds found in this position was approximants (38%), a class which Turk et al. recommend are to be avoided if at all possible.

An example of a difficult token to segment is the word "yellow" in the phrase "the yellow one". This is an especially difficult case as the word begins with an approximant following a vowel, and ends with an approximant preceding a vowel. In these cases, I followed the suggestion of Turk et al., who emphasise that consistency is paramount, and I used the same criteria for segmenting each token I found in this context: I segmented the token at the midpoint in the transitional glide from the preceding and proceeding vowels. I am aware that these midpoints are hard to define in a uniform manner each time; however I was left with little choice. An example of the waveform of one of these tokens is shown in figure (viii) below:

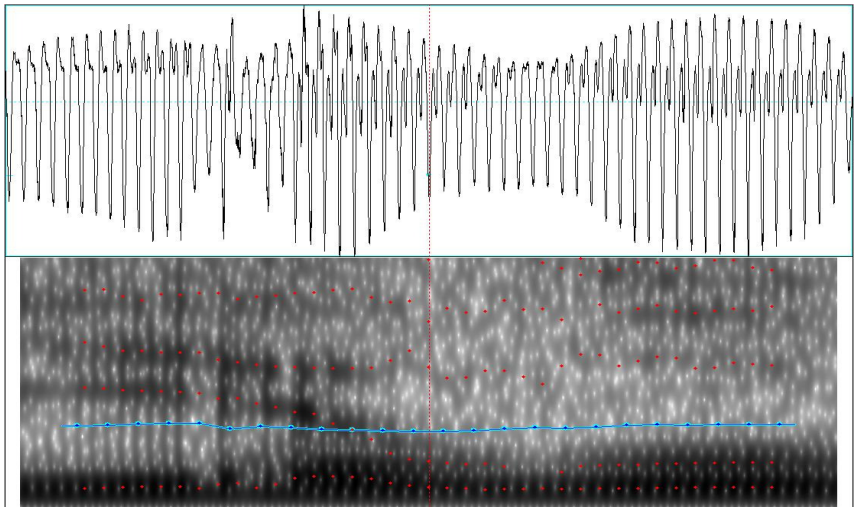

Figure (viii): a waveform of the segmented token "yellow" extracted from the context "the yellow one"

A second problem was the variability in quality of the speech recordings. As the recordings were made over a period of up to 60 minutes using a head-mounted microphone, the quality varies, partly due to the head movements of the participants throughout the experiments, and partly due to some participants speaking very quietly. The background noise can be clearly seen in figure (viii) above. This created extra issues during the segmentation process, as the boundaries between segments were less obvious, and the unavoidable background noise also

exacerbated this. An example of the speech waveform a token uttered by a particularly quiet speaker is shown in figure (ix) below:

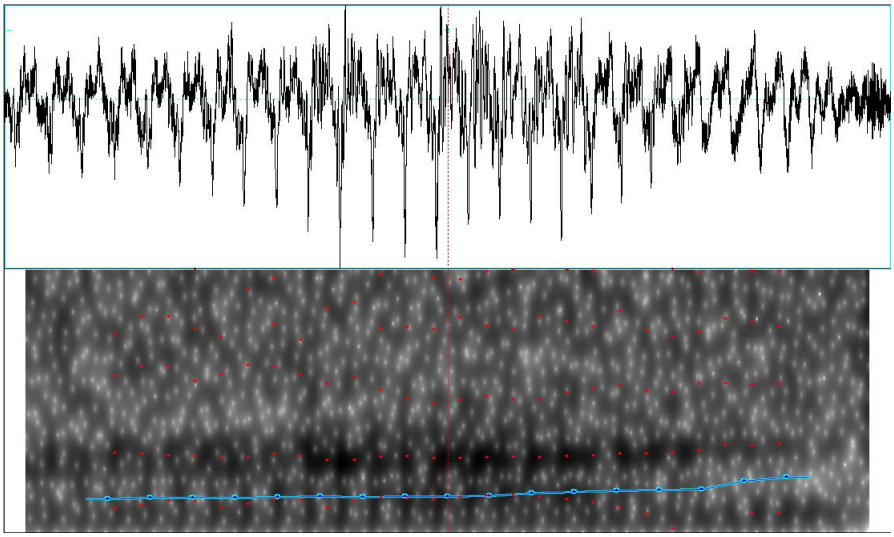

Figure (ix): a waveform of the segmented token “red” extracted from the context “the red one” in poor recording conditions.

As figure (ix) shows, the background noise is so loud in relation to the speaker’s voice that the boundaries of the first and last segments of each token are difficult to accurately define.

In numerous cases, when both members of the dyad spoke simultaneously, both speakers were picked up by both microphones, thus distorting the waveforms and obscuring segment boundaries. In these cases outlined above, I followed the segmentation criteria set out above as closely as possible, but had to discard 5 pairs of tokens due to concern about their levels of accuracy in segmentation. Some 294 tokens, in 147 pairs, remained for the final analysis.

4.1.4 Coding of variables

Each token, having been segmented, was then coded for the following variables shown in Table 3 below:

Table 3: the variables coded for each token used in the analysis

Variable	**Description**
Role	Whether the dyad was assigned roles
Condition Order	Whether the tokens were produced in the 1st or second experimental condition encountered by the dyad
Gaze	Whether the dyad had visual access to the eye-track of their partner on screen
Mouse	Whether the dyad had visual access to the mouse-track of their partner on screen
Combined Gaze Mouse	Whether the dyad had visual access to both the eye-track and mouse-track of their partner on screen
Any Added Channel	Whether the dyad had visual access to either the eye-track or mouse-track of their partner on the screen
Speech Order	Whether the dyad encountered the speech condition first or second
Stimulus Order	Whether the word was produced as token 1 or token 2
Total order of mention	How many times the constituent shape had been referenced in the experiment up until this point
Total order of mention in condition	How many times the constituent shape had been referenced in the experimental condition up until this point
Word	Which word the dyad produced
Shape being referred to	Which constituent shape the dyad was referring to
Word frequency	The frequency of the word (obtained from the British National Corpus)
Number of syllables	The number of syllables in the word
Nuimber of segments	The number of segments in the word
Dyad ID	Which dyad produced the word
Noun	Whether the word produced was a noun
Lexical contrast	Whether the word produced was being used contrastively
Duration	The length of the word measured in seconds

The variables described in Table 3 were included in the analysis for the following reasons:

Role, *Condition Order*, *Gaze*, *Mouse*, and *Speech Order* were included in the analysis as these were conditions of the experiment. The variable *Combined Gaze Mouse* was also added to the analysis in order to check whether having visual access to both the eye-track and mouse-track on screen has a greater effect than having visual access to only one of the extra communication channels, therefore creating a cumulative effect of communication channels. Similarly, the variable *Any Added Channel* was included to check whether the specific added

communication channel is irrelevant; the only effect occurring due to the dyad's access to an added communication channel of any kind.

Each token was also coded for *Stimulus Order*, allowing the analysis to differentiate between introductory mentions of a referent and the second repeated mention. In order to control for the number of intervening non-identical tokens in-between each introductory mention and the following repeated mention, the variables *Total Order of Mention* and *Total Order of Mention in Condition* were added to the analysis. *Total Order of Mention* coded each token for how many times the referent of the referring expression had previously been referred to during the experiment, and *Total Order of Mention in Condition* coded each token for how many times the referent of the referring expression had previously been referred to during the specific experimental condition being encountered by the dyad.

In order to control for the differing word frequencies of the tokens extracted for the analysis, each token's relative *Word Frequency* was obtained from the British National Corpus[2]. I was careful to obtain the frequency of each word according to its grammatical class; for instance the word "red" used as a noun has a relative frequency of 27 per million words, whereas the word "red" used as an adjective has a relative frequency of 126 per million words.

Each token was also coded for the number of *Syllables* and the number of *Segments* it was comprised of. This variable was included in the analysis in order to control for the findings of Zipf (1949), that the frequency of words in natural language is inversely proportional to their length. Both syllables and segments were included as syllables alone may not account fully for token duration, as syllables can be comprised of one or more segments. Segments give a more accurate measurement of a word's length, and therefore the inclusion of this variable allows the analysis to control for the fact that segmental reduction is more likely to occur in tokens comprised of a large number of segments (see Kohler, 1990 for a description of this phenomenon in German).

The variable *Noun* was included in order to control for the phenomenon of phrase-final lengthening in English (Oller, 1973). As the tokens in the analysis included adjectives used as the contrastive item in the NP, it was necessary to code for whether the tokens were nouns or adjectives, as nouns are found phrase-finally in NPs in English.

[2] http://www.natcorp.ox.ac.uk/

The variable *Lexical Contrast* was also added to the analysis. It should be noted that the coding for lexical contrast in this instance is not based on lexical stress placed on the word by the speaker, but instead based on whether the words were used in a lexically contrastive context. Some examples of this are shown below in example (d) below:

Example (d): Examples of phrases used to refer to constituent shapes, extracted from the dialogue of dyad 12. Words coded for lexical contrast are shown in italics:

B: Okay I'll get **the red-** oh no **the *purple* triangle**, yeah?
[…]
A: Oh god. Okay this time. Okay do you wanna get **the *peach* one**? Yeah and I'll get **the *purple* one**.
[…]
B: Okay. You grab **a *green***.

As example (d) shows, the italicised words above were coded for showing lexical contrast. In the case of phrases such as "the purple triangle", both "purple" and "triangle" were both extracted and coded for, with "purple" being coded for providing lexical contrast. In phrases such as "the peach one", only the word "peach" was extracted and segmented, and it was coded for lexical contrast. In phrases where the colour term was used as a noun, it was coded for both lexical contrast and use as a noun.

In order to study the duration differences between specific pairs of repeated mentions, extra variables were also added to the analysis. These variables are shown in Table 4 below:

Table 4: Extra variables added in order to answer the second research question

Variable	**Description**
Difference in duration	The difference in duration between token 1 and token 2, measured in seconds
Difference in the order of mention	The relative distance between token 1 and token 2 in the dialogue

The relative distance between token 1 and token 2 in each dialogue was calculated by coding each token for how many times the constituent shape had already been referred to in some way during the experiment. The difference between was then calculated by subtracting the number of mentions coded for token 1 from the number of tokens coded for token 2. This variable was included in order to test the principles of accessibility theory, as both distance and episode boundaries are factors within this analysis.

The full data lists can be found in the Appendix.

4.1.5 Regression

Regressions were performed on the data in two ways. Firstly, multiple linear regressions were carried out with stepwise entering of dependent variables on the entire data set (N=294). However when the significant predictors were explored further, it was obvious that outlying data points had affected the regression. Therefore the data set was then examined for outliers by calculating the mean and standard deviation statistics for the absolute durations of each introductory mention of a referent (henceforth referred to as *the first token*) and each second identical mention of a referent (henceforth referred to as *the second token*). These statistics are presented in Table 5 below:

Table 5: The means and standard deviations for the absolute durations of token 1 and token 2 of the entire data set.

	N	Minimum	Maximum	Mean	Std. Deviation
Duration of token 1	147	.108202	.736300	.31055320	.113125213
Duration of token 2	147	.130124	1.183784	.28411203	.136098616
Valid N (listwise)	147				

Boxplots were then created to identify the outliers by manipulating the data to look at the absolute durations of all first and second tokens, categorised by the presence and absence of each of the added communication channels. Examples of these boxplots are shown in figures (x) and (xi) below:

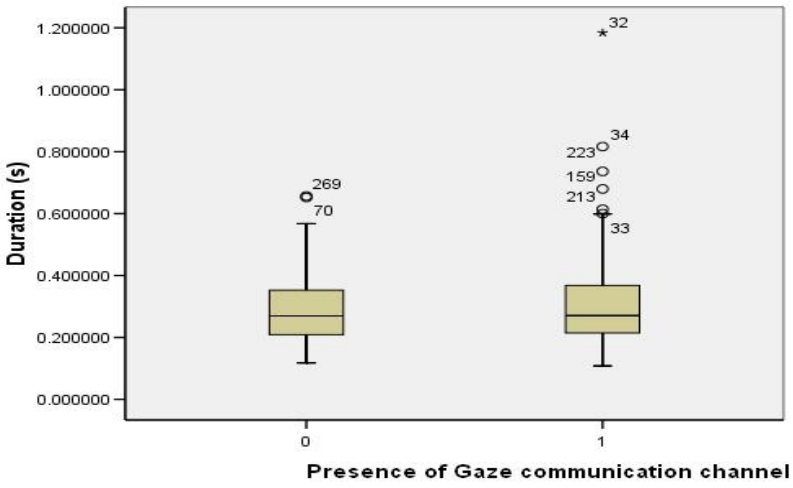

Figure (x): Boxplot showing outliers of the data when categorised by the presence of the Gaze communication channel

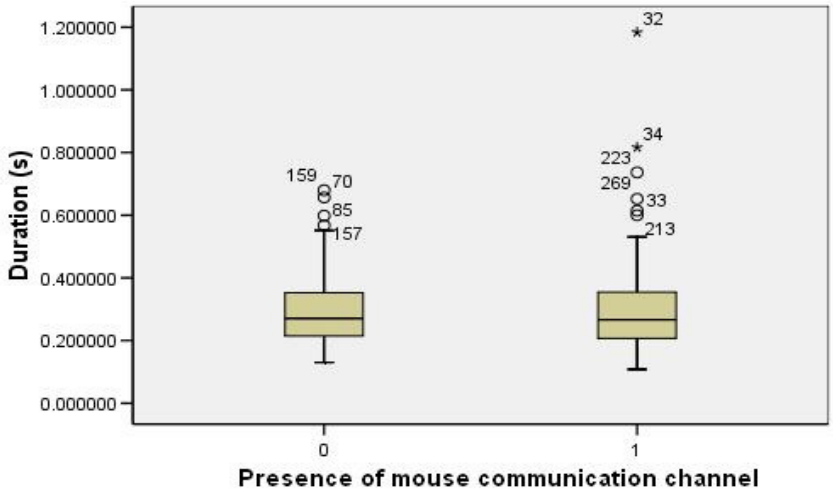

Figure (x): Boxplot showing outliers of the data when categorised by the presence of the Mouse communication channel

The box part of the graph represents the interquartile range of the data, and the whiskers represent the bottom and top quartile of the data set. As the boxplots in figures (x) and (xi) show, there were outliers present in the data. These outliers are represented by the circles and stars on the graphs. The data points represented by circles have values of between 1.5 and 3 box-lengths from the 75th percentile or 25th percentile. The data points represented by stars are extreme values with values more than 3 box-lengths from the 75th percentile or 25th percentile.

In order to clean up the data, each of the data points indicated as outliers by this boxplot analysis were omitted from the final regression analysis, in order to ascertain that they were not responsible for any of the significant results. On further investigation of the outliers, it was obvious that these outliers tended to be tokens segmented from the dialogue of specific dyads. In 2 cases, every token from a dyad had to be removed from the analysis. In order to ascertain the reason for these tokens having especially long durations, the tokens were studied in context. This investigation showed that one particular dyad tended to coordinate the length of a referring expression with the movement of the referent from one position to another; similar to the phenomenon of co-ordination of two-handed movements from disparate starting points (Kelso, Southard & Goodman, 1979). Further investigation of this phenomenon could be fruitful, although it is outside the scope of this current paper.

In total, 23 data points were identified as outliers, and these measurements, together with their paired data points were deleted, leaving 124 pairs of tokens from 26 dyads in the final analysis. The mean and standard deviations of the durations of the first and second tokens in the revised data set are presented in Table 6 below, and a bar chart illustrating the information is shown in figure (xii):

Table 6: The revised means and standard deviations for the absolute durations of first and second tokens of the revised data set.

	N	Minimum	Maximum	Mean	Std. Deviation
Duration of token 1	124	.108202	.510053	.29466227	.086371012
Duration of token 2	124	.130124	.535174	.26021097	.091236208
Valid N (listwise)	124				

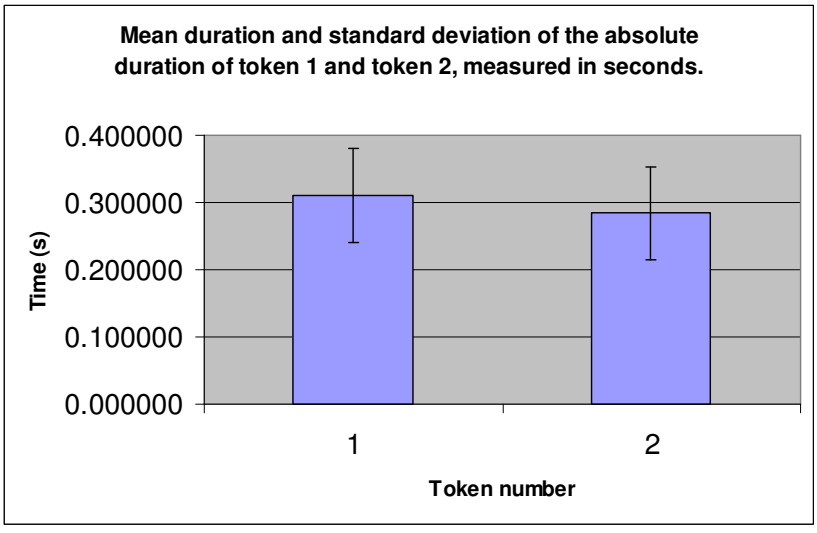

Figure (xii): the mean duration and standard deviation of the absolute duration of all first and second tokens in the revised data set.

Multiple linear regressions were then carried out on the revised data sets. The absolute durations of first and second tokens were regressed on the following 13 predictors: role, condition order, presence or absence of an added communication channel, the number of extra communication channels available, order of speech within the experiment, whether the token was the introductory or the second mention of the referent, the number of times the referent had previously been referred to in the experiment, the number of times the referent had previously been referred to the current experimental condition, the word frequency, the number of syllables in the word, the number of segments in the word, whether the word is a noun, and whether the word was being used in a lexically contrastive manner. It should be noted that predictors coding for eye-track and mouse-track were originally added to the analysis, but as they failed to reach significance, these predictors were omitted and the predictors specifying presence or absence of an added communication channel and the number of extra communication channels available were included in their place.

The 13 predictors listed above accounted for over a third of the variance in absolute duration of the tokens (R^2 = .358, N=248), which was highly significant, $F(13,247) = 11.6$, p=.000.

The presence of an added communication channel (β=-.283, p=.012), the cumulative number of communication channels available (β=.245, p=.032), the order of mention of the referent β=-.216, p=.000) and the number of segments in the word (β=.392, p=.000) all demonstrated significant effects on the absolute duration of tokens, and a summary of these results are presented in Table 7 below:

Table 7: The predictors which exhibit significant effects on the absolute duration of tokens.

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Presence of Added Channel	-.053	.021	-.283	-2.522	.012
Cumulative Number of Added Channels	.032	.015	.245	2.155	.032
Stimulus Order	-.039	.010	-.216	-3.953	.000
Number of Syllables	*.023*	*.014*	*.146*	*1.700*	*.090*
Number of Segments	.029	.007	.392	4.337	.000
Noun?	*.021*	*.012*	*.112*	*1.740*	*.083*

The number of syllables in the word (β=.146, p=.090) and whether the word was a noun or not (β=.112, p=.083) also approached statistical significance and have therefore been included

in the summary regression table. These results show that second mentions of referents are significantly shorter than introductory mentions. The significant effect of the number of segments on the absolute duration of tokens is to be expected, as the greater the number of segments in a word, the longer the word duration.

The results also show that the presence of an added communication channel reduced the absolute duration of tokens, but that the greater the number of added communication channels available to the dyads, the longer the absolute duration of tokens. This result seemed somewhat contradictory, so it was explored further. A bar chart showing the mean duration of tokens according to the number of extra communication channels available is shown in figure (xii) below:

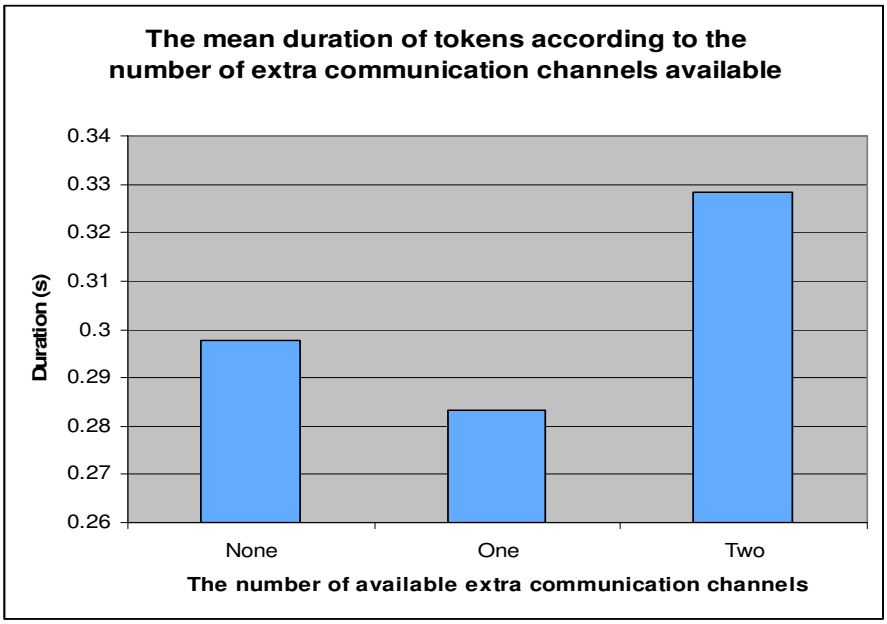

Figure (xii): the mean duration of tokens according to the number of extra communication channels available

As this graph shows, the addition of one extra communication channel had a significant effect on the duration of tokens, shortening them significantly, whereas two communication channels lengthened the duration of tokens significantly. This finding can be studied in further depth by specifying the communication channels which were available to each dyad. A graph presenting these mean durations can be seen in figure (xiii) below:

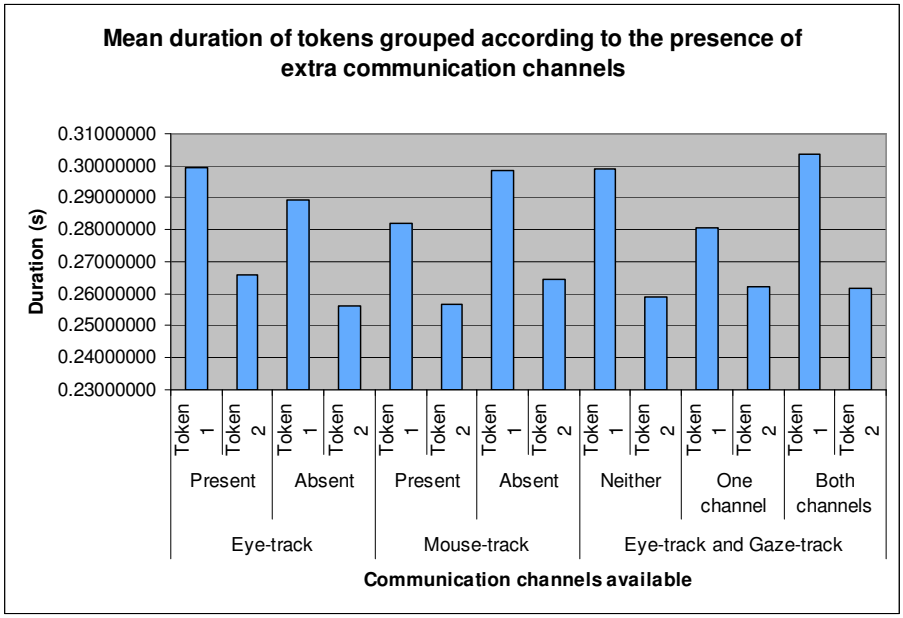

Figure (xiii): the mean duration of tokens according to which communication channels were available to the dyad

As this graph shows, the presence of the Gaze communication channel actually increases the durations of first and second tokens whereas the presence of the Mouse communication channel reduces the durations of first and second tokens. The U-shaped distribution of durations seen in figure (xii) is only partially found in figure (xiii) when the first and second tokens are separated. The introductory mentions follow the pattern found in figure (xii), but the second mentions increase in duration if extra communication channels are available to the dyad.

In order to find out the percentage of variance in duration accounted for by each predictor, the regression outlined above was repeated 13 times; each time omitting one predictor from the analysis and recording the difference in R^2. The results of these regressions can be seen in graphical form in the pie chart in figure (xiv):

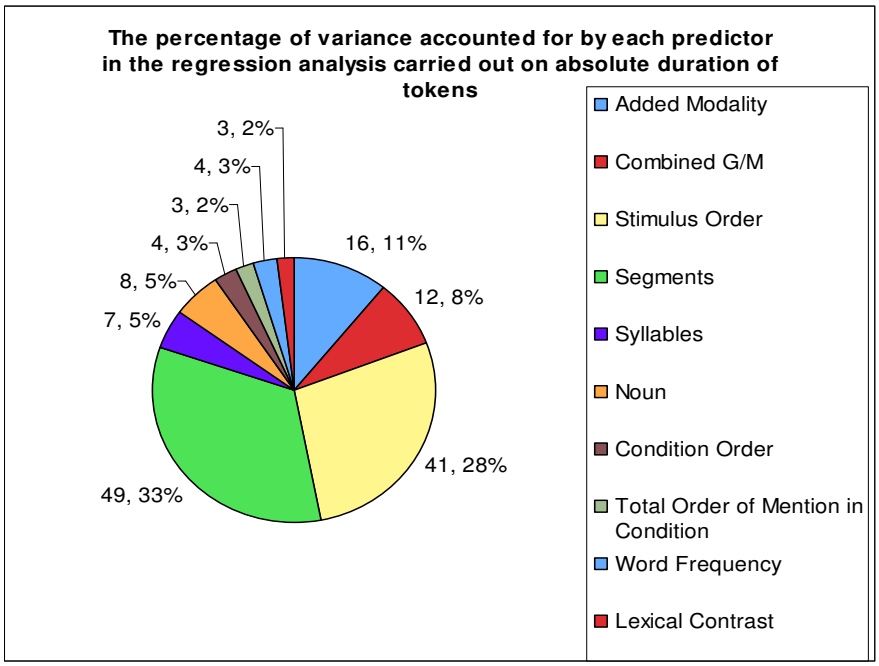

Figure (xiv): the percentage of variance accounted for by each predictor in the regression analysis carried out on the absolute duration of tokens

This graph shows that the greatest amount of variance in the absolute duration of tokens is accounted for by the number of segments, and by whether it was the introductory mention of the referent, or the second mention. In order to see if this effect remained when the duration difference between corresponding first and second mentions was measured, this difference of duration was regressed on the following predictors: role, condition order, presence or absence of an added communication channel, the number of extra communication channels available, order of speech within the experiment, the word frequency, the number of syllables in the word, the number of segments in the word, whether the word is a noun, whether the word was being used in a lexically contrastive manner, and the distance between the first and second token (measured by the total number of intervening mentions of the referent).

The 11 predictors listed above accounted for barely 0.02% of the variance in duration differences between the pairs of tokens (R^2 = .0.24, N=124), which was not significant, $F(11,123) = 1.274$, p=.249.

None of the predictors demonstrated a significant effect on the duration differences between tokens, although the order of speech in the experiment (β=-.172, p=.070), and the word frequency (β=-.215, p=.084) were approaching significance. These results showed that dyads who encountered the speech condition of this experiment second, tended to exhibit reduced

duration differences between introductory and second mentions of a referent. Word frequency would be expected to have the observed effect: the more frequent the word, the smaller the durational difference between first and second mentions. As these results did not reach significance, they must be interpreted with caution.

Due to the U-shaped distribution curve found in the absolute durations mean durations, regressions were then carried out on subsets of the data. This is due to linear regression only finding linear patterns in data; therefore U-shaped distributions would not be found. The data was split according to communication channel, creating the following subsets of data:

- No extra communication channel and Gaze only
- No extra communication channel and Mouse only
- No extra communication channel and both Mouse and Gaze
- Gaze only and both Mouse and Gaze
- Mouse and both Mouse and Gaze
- One communication channel (Gaze or Mouse) and both Mouse and Gaze
- No extra communication channel vs any added communication channel

The results for absolute durations of tokens were as follows. The predictors used in the absolute duration regressions remained constant throughout the analyses, with the exception of the presence of any extra communication channel, which was excluded due to its collinearity with the cumulative number of available channels predictor.

For the subset of data with no extra communication channel and Gaze only, the 12 predictors accounted for over a third of the variance in absolute duration of the tokens (R^2 = .0.381, N=176), which was highly significant, $F(12,175) = 9.977$, p=.000.

The order of mention of the referent (β=-.215, p=.001) and the number of segments in the word (β=.397, p=.001) both demonstrate significant effects on the absolute duration of tokens. A summary of these results are presented in Table 8 below:

Table 8: The predictors which exhibit significant effects on the absolute duration of tokens in the subset of Gaze only and no added extra communication channel.

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Stimulus Order	-.040	.012	-.215	-3.388	.001
Number of Segments	.027	.008	.397	3.473	.001

These results show that second mentions of referents are significantly shorter than introductory mentions. The significant effect of the number of segments on the absolute duration of tokens is to be expected, as the greater the number of segments and segments in a word, the longer the word duration.

For the subset of data with no extra communication channel and Mouse only, the 13 predictors accounted for over a third of the variance in absolute duration of the tokens (R^2 = .0.322, N=128), which was highly significant, $F(12,127) = 6.031$, p=.000.

The order of mention of the referent β=-.216, p=.006) and the number of segments in the word (β=.329, p=.005) both once again demonstrate significant effects on the absolute duration of tokens. A summary of these results are presented in Table 9 below:

Table 9: The predictors which exhibit significant effects on the absolute duration of tokens in the subset of Mouse only and no added extra communication channel.

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Stimulus Order	-.039	.014	-.216	-2.808	.006
Number of Syllables	*.034*	*.020*	*.199*	*1.692*	*.093*
Number of Segments	.032	.011	.329	2.884	.005

As these results show, the effect of the number of syllables in a word is approaching significance. This is to be expected, as the number of syllables is another method of measuring word length, although it is not as accurate a means as counting the number of segments.

For the subset of data with no extra communication channel and both extra communication channels only, the 12 predictors accounted for over a quarter of the variance in absolute duration of the tokens (R^2 = .0.266, N=128), which was highly significant, $F(12,127) = 4.287$, p=.000.

The order of mention of the referent (β=-.288, p=.000) and the number of segments in the word (β=.330, p=.005) both once again demonstrate highly significant effects on the absolute duration of tokens. Additionally, the presence of roles ((β=-.199, p=.028) in the experimental context has a significant effect on the absolute durations of tokens; reducing their duration. The grammatical class of the word (β=.211, p=.040) also shows a significant effect, as nouns tended to be of longer duration. A summary of these results are presented in Table 10 below:

Table 10: The predictors which exhibit significant effects on the absolute duration of tokens in the subset of both Gaze and Mouse and no added extra communication channel.

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Role	-.038	.017	-.199	-2.220	.028
Order of Conditions	*-.032*	*.019*	*-.177*	*-1.711*	*.090*
Stimulus Order	-.053	.015	-.288	-3.592	.000
Number of Segments	.033	.011	.330	2.888	.005
Noun?	.041	.019	.211	2.081	.040

The effect of the order of conditions (β=-.177, p=.090) encountered by the dyad also approaches significance in this data view, showing that the duration of tokens decreased as the experiment progressed.

For the subset of data with Gaze only and both extra communication channels, the 12 predictors accounted for over a third of the variance in absolute duration of the tokens (R^2 = .0.376, N=120), which was highly significant, $F(12,119) = 6.974$, p=.000.

The order of mention of the referent (β=-.195, p=.024) and the number of segments in the word (β=.530, p=.005) both once again demonstrate highly significant effects on the absolute duration of tokens. Additionally, the grammatical class of the word (β=.248, p=.041) also shows a significant effect, as nouns tended to be of longer duration. A summary of these results is presented in Table 11 below:

Table 11: The predictors which exhibit significant effects on the absolute duration of tokens in the subset of both Gaze and Mouse and the Gaze only communication channels.

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Stimulus Order	-.035	.015	-.195	-2.286	.024
Number of Segments	.033	.012	.530	2.884	.005
Noun?	.047	.023	.248	2.070	.041

For the subset of data with Mouse only and both extra communication channels, 11 predictors accounted for over a third of the variance in absolute duration of the tokens (R^2 = .0.376, N=120), which was highly significant, $F(11,71) = 2.997$, p=.003. The lexical contrast predictor had to be omitted from this analysis as it was a constant in this subset of data.

The cumulative number of added communication channel available to the dyad (β=.242, p=.039) showed the only significant effect in this subset of data: the addition of Gaze to Mouse as a second communication channel increased the absolute duration of tokens. The effect of order of mention of the referent (β=-.227, p=.063) approached significance once again, showing that the second tokens of each token pair are shorter than the introductory mentions. Additionally, the grammatical class of the word (β=.298, p=.081) also shows a significant effect, as nouns tended to be of longer duration. These results are summarised in Table 12 below:

Table 12: The predictors which exhibit significant effects on the absolute duration of tokens in the subset of both Gaze and Mouse and Mouse only.

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Cumulative Number of Added Channels	.040	.019	.242	2.109	.039
Stimulus Order	*-.038*	*.020*	*-.227*	*-1.897*	*.063*
Noun?	*.050*	*.028*	*.298*	*1.774*	*.081*

The final subset of data to be analysed in this way is the subset with either Mouse only or Gaze only and both extra communication channels. In this data set, 12 predictors accounted

for over a third of the variance in absolute duration of the tokens (R^2 = .0.392, N=156), which was highly significant, $F(12,155) = 9.316$, p=.000.

Once again, the number of segments (β=.564), p=.000) and the order of mention of the referent (β=-.175, p=.017) showed a significant effect on the absolute durations of tokens. However, the cumulative number of added communication channel available to the dyad (β=.131, p=.069), the grammatical class of the word (β=.163, p=.065), and whether the word is being used contrastively or not (β=.206, p=.062) also approached statistical significance. Words used contrastively tend to be longer in duration in this subset of data. A summary of these results is shown in Table 13:

Table 13: The predictors which exhibit significant effects on the absolute duration of tokens in the subset consisting of both Gaze and Mouse and either Mouse only or Gaze only tokens.

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Cumulative Number of Added Channels	*.027*	*.015*	*.131*	*1.836*	*.069*
Stimulus Order	-.031	.013	-.175	-2.416	.017
Number of Segments	.038	.010	.564	3.919	.000
Noun?	*.029*	*.016*	*.163*	*1.859*	*.065*
Lexical Contrast	*.094*	*.050*	*.206*	*1.878*	*.062*

The same subsets of data as outlined above were also used to study the difference in duration between corresponding first and second tokens. In order to study this effect, the following ten predictors were used in the regression analysis: role, condition order, presence or absence of an added communication channel, the number of extra communication channels available, order of speech within the experiment, the word frequency, the number of syllables in the word, the number of segments in the word, whether the word is a noun, whether the word was being used in a lexically contrastive manner, and the distance between the first and second token (measured by the total number of intervening mentions of the referent).

Only one of these subsets exhibited any significant effects: in the subset of data with Mouse only and no extra communication channels, 10 predictors accounted for barely 0.3% of the variance in duration differences between the pairs of tokens (R^2 = .0.037, N=64), which was not significant, $F(10,63) = 1.244$, p=.286.

The number of segments (β=-0.243, p=.037) showed significant effects in this data set, but this is to be expected, as explained previously. The cumulative number of added channels

(β=-0.243, p=.080) is approaching significance for this data set, and shows that the greater the number of added channels available to the dyad, the smaller the durational differences between first and second mentions of a referent.

Within the subsets of Gaze only and both added communication channels, and either Mouse only or Gaze only and both added communication channels, predictors approached statistical significance, but no predictors showed a significant effect. These results have been added into a summary of regressions on these data sets, and can be found in Table 14 below:

Table 14: The predictors within specified subsets which exhibit significant effects on the duration differences between pairs of token.

Subset of data	Predictor	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
No extra communication channel and Mouse only	*Cumulative Number of Added Channels*	*-0.052*	*0.029*	*-0.243*	*-1.786*	*0.080*
	Number of Segments	-0.043	0.020	-0.418	-2.145	0.037
Gaze only and both added communication channels R^2=0.77, N=60, F(10,59) =1.493, p= .170	*Speech Order*	*-0.042*	*0.021*	*-0.266*	*-2.002*	*0.051*
	Frequency	*-0.001*	*0.000*	*-0.349*	*-1.860*	*0.069*
	Noun	*-0.054*	*0.027*	*-0.355*	*-1.953*	*0.056*
Gaze only or Mouse only and both added communication channels R^2=0.002, N=78, F(10,77) =1.018, p= .438	*Frequency*	*0.000*	*0.000*	*-0.296*	*-1.729*	*0.088*

These regressions were then all repeated by calculating the percentage reduction of a token from its introductory mention to its second mention, and substituting this value for the duration difference calculated in seconds. This change in regression technique was included to control for the fact that it is easier to shorten a longer word than a shorter word. By calculating the percentage reduction, this controls for the varying lengths of the words extracted for analysis. However, this extra analysis did not affect the results of the original regression: no other predictors were found to be having a significant effect on token reduction. These results have therefore been omitted from the results section.

4.2 Analysis: Adjacent Pairs

4.2.1 Selection

The analysis was then reduced in order to study the effects of these predictors on adjacent pairs of tokens, in order to see whether restricting the data set further would provide any more significant predictors. As shown previously in figure (v), 44% of the tokens extracted for the original analysis were adjacent pairs; therefore the analysis was restricted to only include these tokens. This left a total of 108 tokens to be included in the analysis.

4.2.2 Coding of variables

The same variables were used in this analysis as were used for the analysis outlined in section 4.1. The full list of variables and their reasons for inclusion can be found in Table 3. The variables *Total Order of Mention* and *Total Order of Mention in Condition* were omitted from the analysis, as there were no intervening tokens between the pairs of adjacent tokens.

4.2.3 Regression

Multiple linear regressions were then carried out on the revised data sets. The absolute durations of first and second tokens were regressed on the following 11 predictors: role, condition order, presence or absence of an added communication channel, the number of extra communication channels available, order of speech within the experiment, whether the token was the introductory or the second mention of the referent, the word frequency, the number of syllables in the word, the number of segments in the word, whether the word is a noun, and whether the word was being used in a lexically contrastive manner. It should be noted that once again predictors coding for eye-track and mouse-track were originally added to the analysis, but as they failed to reach significance, these predictors were omitted and the predictors specifying presence or absence of an added communication channel and the number of extra communication channels available were included in their place.

The 11 predictors listed above accounted for over a third of the variance in absolute duration of the tokens (R^2 = .379, N=108), which was highly significant, $F(11,107) = 6.941$, p=.000.

Although these predictors account for a greater amount of the variance in the absolute duration of the tokens than the full data set, fewer predictors reach significance. In this data set, only the number of segments in the word (β=.334, p=.032) showed a significant effect, and the order of tokens tended towards significance (β=-.148, p=.055). These results show that second mentions of referents are significantly shorter than introductory mentions. Again, the significant effect of the number of segments on the absolute duration of tokens is to be expected, as the greater the number of segments in a word, the longer the word duration.
Interestingly, the presence of added communication channels, and the number of added communication channels fail to reach significance within this subset of data, which is in direct contrast to the results found on the full data set, summarized in section 4.1.4. A summary of the regression on the adjacent tokens data set is presented in Table 15 below:

Table 15: The predictors which exhibit significant effects on the absolute duration of adjacent tokens.

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Stimulus Order	*-.028*	*.014*	*-.148*	*-1.945*	*.055*
Number of Segments	.027	.012	.334	2.176	.032

In order to see whether the revision of the data set to only include adjacent pairs of tokens would show the significant effects of any predictors, the duration difference between corresponding first and second mentions was measured, and this difference of duration was regressed on the following predictors: role, condition order, presence or absence of an added communication channel, the number of extra communication channels available, order of speech within the experiment, the word frequency, the number of syllables in the word, the number of segments in the word, whether the word is a noun, and whether the word was being used in a lexically contrastive manner.

Once again, The 10 predictors listed above accounted for barely 0.03% of the variance in duration differences between the pairs of tokens (R^2 = -.0.35, N=54), which was not significant, $F(10,67) = .822$, p=.609.

None of the predictors demonstrate a significant effect on the duration differences between adjacent tokens, although the order of conditions experienced by each dyad in the experiment

(β=-.285, p=.086), approaches significance. This result shows that the difference in duration between adjacent tokens decreased as the experiment progressed.

Unfortunately, due to the smaller number of tokens used in this analysis, it was not possible to further subdivide the data set according to communication channel, as the data sets were too small to be reliable for regression analysis.

5. Discussion

In the work reported in this paper, my main aim was to discover whether any added communication channel (in this case, the mouse-track and/or eye-track) between two members of a dyad in a joint construction task would work in the same way as the visual channel in face-to-face dialogue in their introductory mentions of a referent; therefore reducing the difference in duration between first and second mentions of a referent, following the research of Anderson et al. (1997).

The initial regression analysis showed that in general, the presence of an added channel of communication reduced the duration of tokens, but that the presence of both extra channels of communication increased their duration. This did not fully support the hypothesis of the greater the number of communication channels available to each dyad, the shorter the durations of both the introductory and the second mentions of the referent. The results from the addition of one communication channel support the hypothesis, but the addition of a second extra communication channel do not follow this precedent. I suggest that the reasons for this could be that one communication channel aids the dyad and increases their level of common knowledge, but that the addition of a second communication channel adversely affects a player's capability of following the instructions of the game as well as the mouse- and eye-movements of their partner, as the cognitive load is too great. This explanation follows the theory of Anderson et al. (1997), who suggest that speakers increase their clarity when comprehension problems are apparent within a dialogue. The finding that nouns tended to also be of longer duration can be explained by the phenomenon of phrase-final lengthening as described by Oller (1973).

However, when the analysis was restricted to comparing corresponding first and second mentions of tokens, there was no significant effect of additional communication channels to the duration differences between corresponding tokens. When the data was further divided into subsets according to the specific communication channels open to the dyads, the cumulative number of extra communication channels was only significant in the subset of data of Mouse only and both Gaze and Mouse communication channels, where it increased the overall durations of first and second tokens. This finding can again be explained by the theory put forward by Anderson et al. (1997).

The lack of statistical support for the hypothesis that the greater the number of communication channels available to each dyad, the shorter the durations of both the introductory and the second mentions of the referent, must be explained in another way.

Anderson et al. (1997) showed experimentally that the visual channel of communication is used as a tool by the speaker to check on the state of an interaction, and that this added channel allowed first and second mentions of a referent to be significantly less intelligible (and therefore shorter in duration). It therefore could be that the visual channel of communication, i.e. the times when speakers are in face-to-face contact whilst communicating, works in a different way to that of seeing a speaker's eye-track or their movements with a mouse. The eye-track may not be a reliable added channel of communication, because saccades of the eye can be initiated in as little as 120ms at an extremely high velocity (Kirschner & Thorpe, 2006); therefore these movements may be too difficult to track successfully by each member of a dyad for purposes of comprehension.

The hypothesis that the duration of introductory mentions of referents will be significantly longer than second mentions of referents in all conditions of the experiment, following the findings of Fowler & Housum (1987) is supported by the results in this analysis. The mean duration of the introductory mentions of a referent were significantly shorter than second mentions, and this can be clearly seen in figures (xii) and (xiii). Regression analyses show that the order of mention of each token has a significant effect on the duration of a token, as shown in Table 7. This finding is supported by the regressions carried out on the subsets of data, which show that the order of mention of tokens is significant in every subset of data apart from the subset of data containing the added communication channels of Mouse only, and both Mouse and Gaze. However, even in this data set, the result tended towards significance (p=.06). I present these results as evidence for the theory proposed by Fowler & Housum (1987) that the second mentions of referents are significantly shorter than introductory mentions.

The final hypothesis submitted in this experiment was that if the dyad has encountered the non-speech part of trial before the speech part of the trial, the lesser the need will be for distinct first tokens; therefore the duration of both the introductory and second tokens will be shorter. However, this hypothesis is not supported by the analyses outlined above: speech order did not reach statistical significance at any point in the analysis, although this predictor tended towards significance in the subset of data of Gaze only and both added communication channels (p=.051), showing that the dyads which encountered the speech part of the experiment second tended to reduce the difference in duration between first and second mentions of a referent. I suggest that the reasons for this are that these dyads had established a successful method of constructing tangrams without the need for dialogue; therefore introductory mentions of referents in the speech part of the experiment did not need to be as

intelligible. This result cannot be thought of as support for the original hypothesis, and therefore this hypothesis is rejected.

In summary, this study has shown that the visual channel of communication exhibits different properties to other added communication channels, such as eye-track and mouse-track. This finding should be explored further, by creating an experimental setting which directly tests the relative properties of these three communication channels simulataneously. This study also shows the relative advantages and disadvantages of using 'real' speech as opposed to speech elicited under laboratory conditions. Although it may be simpler to collect a sufficient number of tokens under laboratory conditions, the importance of studying speech in context can never be underestimated.

Bibliography

Anderson, A., Bard, E.G, Sotillo, C., Newlands, A., & Doherty-Sneddon. (1997). Limited visual control of the intelligibility of speech in face-to-face dialogue. *Perception & Psychophysics,* 59(4), 580-592.

Anderson, A. & Howarth, B. Referential form and word duration in video-mediated and face-to-face dialogues. In Bos, Foster, & Matheson (eds): *Proceedings of the sixth workshop on the semantics and pragmatics of dialogue (EDILOG 2002),* Edinburgh, UK, September 2002, 1-20.

Ariel, M. (1990). *Accessing Noun Phrase antecedents.* London: Routledge.

Ariel, M. (2001). Accessibility theory: An overview. In Sanders, Schliperoord and Spooren eds. *Text representation*. (pp. 29-87). Amsterdam: John Benjamins (Human cognitive processing series).

Bard, E. G. & Anderson, A. H. (1994). The unintelligibility of speech to children: Effects of referent availability. *Journal of Child Language,* 42(1), 1-22.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M. Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language,* 42 (1), 1-22.

Bard, E.G, & Aylett, M. (2005). Referential form, word duration and modelling the listener in spoken dialogue. In J. Trueswell & M. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-action traditions*. (pp. 173-191). Cambridge, MA: MIT Press.

Bard, E.G., Brew, C. & Cooper, L. (1991). *Psycholinguistic studies on increment recognition of speech: an introduction to the messy and sticky.* (No. BR1375). DYANA, Esprit Basic Research Action.

Bard, E.G., Hill, R., & Foster, M.E. (2008). What tunes accessibility of referring expressions in task-related dialogue? In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci 2008)*, Washington, DC, July 2008.

Bard, E.G., Sotillo, C., Anderson, A.H., Doherty-Sneddon, G., & Newlands, A. (1995). The control of duration in running speech. In K. Elenius & P. Branderud (eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences* (Vol. 4). Stockholm: Stockholm University.

Berlin, B., & Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution.*Berkeley & Los Angeles: University of California Press.

Bolinger, D. (1981). *Two kinds of vowels, two kinds of rhythm.* Bloomington, IN: Indiana University Linguistics Club.

Brennan, S. E. (2005). How conversation is shaped by visual and spoken evidence. In J. Trueswell & M. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-action traditions*. (pp. 95-129.) Cambridge, MA: MIT Press.

Browman, C.P., & Goldstein, L. (1992). Articulatory Phonology: An Overview. *Phonetica,* 40, 155-180.

Brown, P.M., & Dell, G.S. (1986). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology,* 19, 441-472.

Carletta, J., Evert, S., Heid, U., Kilgour, J. (2005). The NITE XML Toolkit: data model and query. *Language Resources and Evaluation Journal* 39(4): 313-334.

Carletta, J., Nicol, C., Taylor, T., Hill, R., de Ruiter, J. P., & Bard, E.G. (under revision). Eyetracking for two-person tasks with manipulation of a virtual world. *Behavior Research Methods, Instruments, and Computers.*

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

Dohery-Sneddon, G., Anderson, A.H., O'Malley, C., Langton, S., Garrod, S., & Bruce, V. (1997). Face-to-face and video-mediated communication: a comparison of dialogue structure and task performance. *Journal of Experimental Psychology: Applied.*, 3(2), 105-125.

Fowler, C. (1988). Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech,* 31(4), 307-319.

Fowler, C. & Housum, J. (1987). Talkers' signalling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language,* 26, 489-504.

Fowler, C., Levy, E. T., & Brown, J.M. (1997). Reductions of spoken words in certain discourse contexts. *Journal of Memory and Language,* 37, 24-40.

Grice, H.P. (1957). Meaning. *Philosophical Review,* 66, 377-388.

Gundel, J.K., Hedberg, N., & Zacharski, R. (11993). Cognitive Status and the form of referring expressions in discourse. *Language* 69(2), 274-307.

t'Hart, J., Collier, R., & Cohen, A. (1990). *A Perceptual Study of Intonation.* Cambridge: CUP.

Hawkins, S. & Warren, P. (1994). Phonetic influences on the intelligibility of conversational speech. *Journal of Phonetics,* 22, 493-511.

Horton, W.S. &Keysar, B. (1996). When do speakers take common ground into account? *Cognition,* 59, 91-117.

Howell, P. and Young, K., 1991. The use of prosody in highlighting alterations in repairs from unrestricted speech. *The Quarterly Journal of Experimental Psychology* 43A(3), 733–758.

Huttenlocher, D. P., & Zue, W. (1984). A model of lexical access based on partial phonetic inofrmation. *Proceedings of the ICASSP 84,* 26(4), 1-4.

Just, M. A. & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension', *Psychological Review,* 87(4), 329-354.

Kelso, J.A.S., Southard, D.L., & Goodman, D. (1979). On the coordination of two-handed movements. *Journal of Experimental Psychology: Human Perception and Performance,* 5, 229-238.

Kirschner, H. & Thorpe, S.J. (2006). Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vision Res*., 46(11), 1762-1776.

Kohler, K.J. (1990). Segmental reduction in connected speech in German: phonological facts and phonetic explanations. In Hardcastle, Marchal (eds*.) Speech production and speech modelling*. Kluwer: Dordrecht.

Monk, A. & Gale, C. (2002). A look is worth a thousand words: full gaze awareness in video-mediated conversation. *Discourse Processes,* 33(3), 257-278.

Oller , D. K. (1973). The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America,* 54*,* 1235-1247.

Perkell, J.S., Guenther, F.H., Matthies, M.L., Perrier, P., Vick, J., Wilhelms-Triarico, W., & Zandipour, M. (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics,* 28, 233-272.

Prince, E. F. (1981b). Toward a taxonomy of given-new information. In P. Cole (eds.) Radical Pragmatics. (pp. 223-256.) New York: Academic Press.

Rosnagel, C. (2000). Cognitive load and perspective-tasking: applying the automatic controlled distinction to verbal communication. *European Journal of Social Psychology,* 30, 429-445.

Stevens, K. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America,* 111, 1872-1891.

Sweller, J. (1988). Cognitive load during problem-solving: Effects on learning. *Cognitive Science,* 12, 257-283.

Terken, J. & Hirschberg, J. (1994). Deaccentuation of words representing 'given' information: effects of persistence of grammatical function and surface position. *Language and Speech,* 37(2), 125-145.

Turk, A., Nakai, S. & Sugahara, M. (2006). Acoustic segment durations in prosodic research: a practical guide. In Sudhoff, Lenertová, Meyer, Pappert, Augurzky, Mleinek, Richter &

Schließer (eds): *Methods in Empirical Prosody Research*. (pp. 1-28). Berlin, New York: De Gruyter.

Tyler, L.K., & Wessels, J. (1985). Is gating an online task? Evidence from naming latency data. *Perception and Psychophysics,* 38, 217-222.

Zipf, G.K. (1935). *The psychobiology of language.* New York: Houghton-Mills.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley.