# A Collection of Self-repairs from the Map Task Corpus


Jean Carletta, Richard Caley, and Stephen Isard
Human Communication Research Centre
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW
Scotland
Email: {J.Carletta,R.Caley,S.Isard}@edinburgh.ac.uk
Contact phone: +44-31-650-4438
Contact fax: +44-31-650-4587



**Abstract**

As part of a study of message generation in spontaneous speech, we have collected information about self-repairs in ten dialogues from the HCRC Map Task Corpus. This document describes the coding scheme which was used, which (1) identifies where the repairs occur on the transcripts, (2) describes what material interrupts the message during editing and how the editing phase relates to the utterance prosodically, (3) specifies how the major components of the repair are related to each syntactically and semantically, and (4) suggests what sort of speaker difficulty is likely to have caused the repair. We include evidence that the judgment of whether or not a repair has occurred between two particular words or during a particular move is reliable. We also include a rough summary of the distribution and rate of repairs in the dialogues analysed.

# Contents

# 1 Introduction

One way in which spontaneous speech differs from read speech and from written language is that it contains self-repairs. Self-repairs (or, "repairs", for short) are places in speech where the flow of speech is interrupted because some sort of trouble has been detected (Levelt [10]). For instance, in the utterance "Go from left again to /uh .../ from pink again to blue", the flow of speech is interrupted when the speaker realises that the word "left" is incorrect. Repairs can also occur because the speaker thinks of a better way of conveying his message, or even has just run out of words which he is ready to say. Repairs can be used to investigate the cognitive processes underlying language production because they are a visible trace of where those processes break down, showing us the limits of those processes and something of how they interact with each other. There are several existing collections of repairs in various domains. Levelt [10, 11] collected repairs in Dutch monologues which describe connected patterns of coloured dots aligned in a two-dimensional grid. Blackmer and Mitton [2] collected repairs from listener contributions to a call-in radio show featuring current events. In addition, Hieke [4] has collected information specifically about a class of repairs which we term hesitations for native English and German speakers and for native Germans speaking English. The domains for these collections are specifically designed to illuminate different aspects of spontaneous language production such as formulation and self-monitoring.[1]

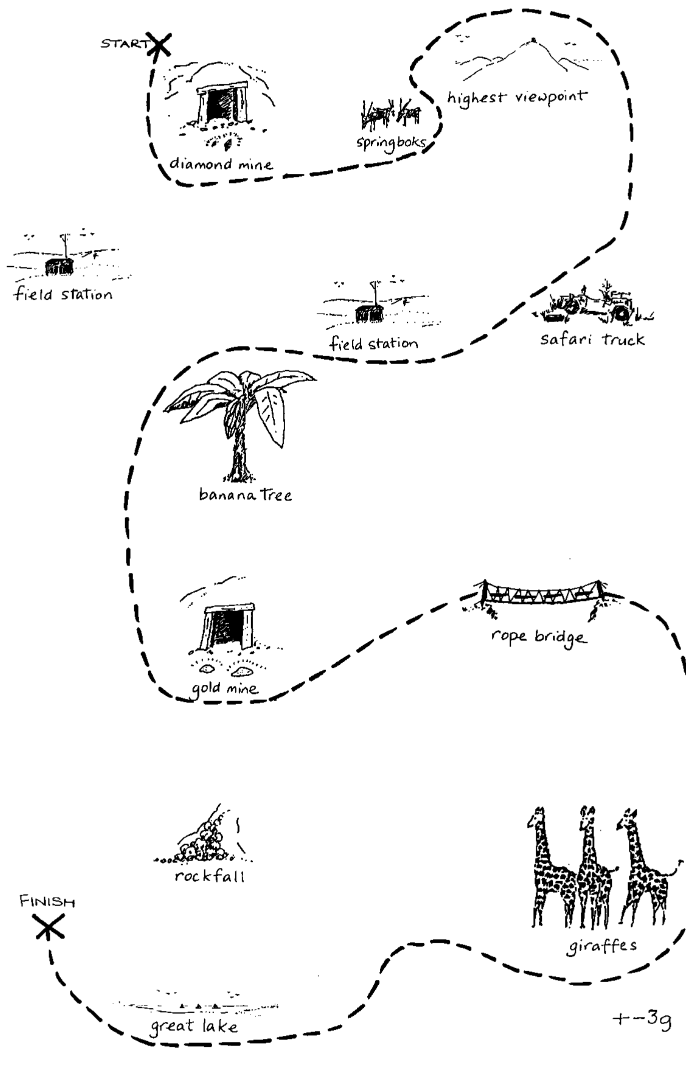
In this document, we describe a collection of repairs from the HCRC Map Task Corpus [1] which is designed specifically to investigate the semantic planning of utterances (or, more precisely, *message generation*, the process of getting from a communicative intention to a language independent "conceptual structure" specifying the meaning of a single utterance (Levelt [11])). The Map Task Corpus is a collection of 128 task-oriented dialogues involving approximately fifteen hours of spontaneous speech. In the dialogues, two participants are given somewhat different versions of a simple map with approximately fourteen landmarks on it. The participants can not see each other's maps. The landmarks differ somewhat between the two maps. In addition, one participant has a route drawn on the map (see figure 1). The task is for the second participant to duplicate the route. The Map Task Corpus is generally useful for a study of repair because the design of the corpus varies conditions which may be related to repair, such as whether or not the participants have eye contact or are familiar with each other. Moreover, it is also a particularly good place to look for repairs which tell us something about message generation. Since the dialogues are task-oriented, the messages produced are constrained and more predictable than in, say, Blackmer and Mitton's dialogues. The maps give us complete information about the participants' knowledge of the domain, making it easier to decipher a speaker's communicative intentions than in other task-oriented domains. Finally, because the task involves dialogue (unlike in Levelt's work), participants must produce utterances relatively quickly to keep the partner from speaking, but also have less idea what to say next, because it may depend on the partner's contributions. These characteristics of the domain, along with previous analysis of the Map Task Corpus, make our collection useful for investigating aspects of language production which previous collections were not intended to address.

# 2 The coding scheme

Because we have collected repairs for a novel purpose, we have developed our own way of coding the repairs. Our coding relies on Levelt's terminology. Levelt [10, p. 44] defines each repair in terms of the *moment of interruption (MOI)*, or "point at which the flow of speech is interrupted for 'editing' of some sort." The *original utterance (OU)* is the material which precedes the MOI; it is followed by an *editing phase* consisting of a period of hesitation,

[1] There is also work on repairs for the ATIS corpus [13], which contains dialogues between travel agents and their customers, collected specifically to aid speech recognition for that domain.

Figure 1: An Example Route Giver Map



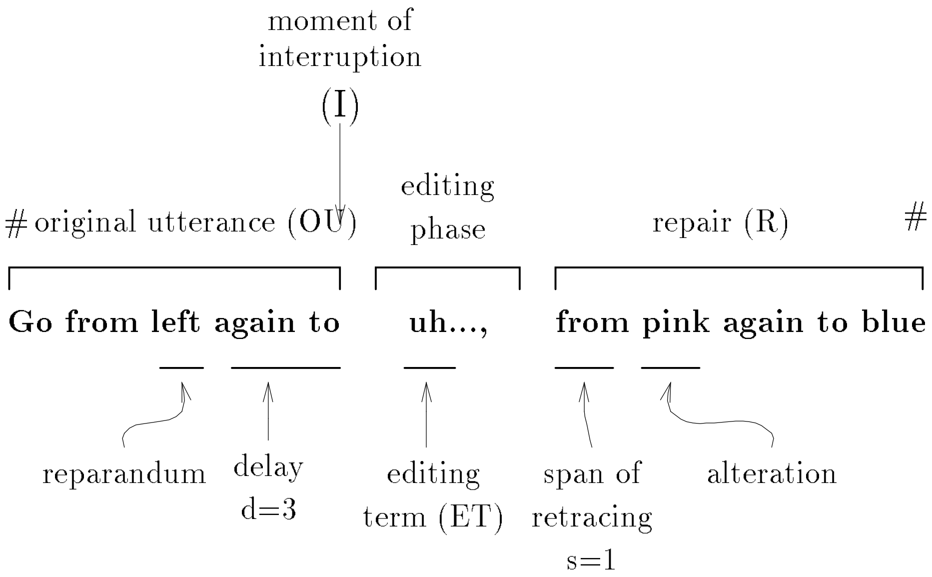



Figure 2: The structure of repair. Some technical notions.

possibly with an editing term, and the *repair (R)*. In both Levelt's work and our own, a repair does not necessarily correct the OU but may just add to it. In addition, we define the *target utterance* to be some fluent utterance which conveys the meaning which the speaker intends as the meaning of the disfluent utterance; for most (but not all) repairs, the coder's choice of target is uncontroversial. Levelt's *reparandum* is then the portion of the OU which is not present in the target (if any), and the *alteration* is the portion of the repair which replaces the reparandum in the target (if any). The MOI may be delayed by some number of syllables after the place in the surface speech associated with the trouble, and the repair typically retraces some part of the OU in order to help the listener determine how the OU and repair fit together to form the target utterance. Figure 2, taken from Levelt [10, p. 45], applies Levelt's definitions to an example. We define the target utterance for the example to be "Go from pink again to blue."

The coded information for each repair uniquely identifies it and locates it in the dialogue corpus, categorises it in terms of how the repair relates to the OU, describes the main components of the repair, and suggests the type of trouble which the speaker was experiencing when the repair occurred. This text describes the coding abstractly; information about the form of the data files is in appendix A.

## 2.1 Identifying and locating the repair

We coded information to identify each repair and locate it in the transcript. The data files are indexed by the turn number (as given on the *T line of the transcript), and the number which sequentially indexes the repairs in a single turn, beginning counting at 1. In addition, we give the position of the MOI in terms of the number of "words" into the turn where it occurs, and an index into the *M line of the transcript giving the dialogue move in which the repair occurred, where that coding exists.

## 2.2 Classifying the repairs

Our repairs divide into four major categories which are closely related to Blackmer and Mitton's [2]: filled hesitations, unfilled hesitations, replacements, and insertions.

*Hesitations* are repairs where (1) the OU and repair concatenate to form the target, allowing for the deletion of some words from the beginning of the repair which retrace or repeat the end of the OU, and (2) it appears that the speakers knew the utterances were incomplete when they reached the MOI. The first condition excludes cases such as "Go to the left of /uh/ the **right** of the swamp", where ordinarily the target would be "Go to the right of the swamp" rather than "Go to the left of the right of the swamp". In some cases, intention was clear from the prosody of the utterance; in others, it was judged by looking at the speaker's map. The second condition differentiates hesitations from seamless insertions, as defined below. *Unfilled hesitations* are marked simply by a gap in the flow of speech. To eliminate some types of pauses which are not related to planning, we counted only those pauses which were longer than we imagined would occur if the speaker were reading the utterance rather than producing it spontaneously. *Filled hesitations* are indicated by something more substantial: one of the standard editing terms ("uh", "erm"), repetition of words, retracing, grunts, and some (but not all) uses of "sort of", and "waffling" behaviours such as "hold on" which appeared to be used primarily to stall for time. We separate unfilled hesitations from filled ones simply because the identification of unfilled hesitations is controversial and because previous work (see, for example, Hieke [4]) suggests that the use of particular editing material, or any editing material at all, might be related to the hesitation's cause.

*Replacements* are repairs in which (1) the OU and repair do not concatenate to form the target, even allowing for retracing or repeated words, and (2) some parts of the OU must be deleted in order to reach the target. Levelt's main example ("Go from left again to/ uh.../ from pink again to blue") is a prototypical replacement stemming from using the wrong word. However, the reparandum of a replacement can be anything from a single phoneme ("Have you got cornfu-/ fields marked?") to the entire OU ("The start is/ Do you have the extinct volcano?"). Although some replacements are marked by editing material (typically "uh" or "sorry" — see James [6, 7]) and others are not, we have not divided them into separate categories on this basis, because what counts as a replacement is relatively uncontroversial even without explicit marking. We additionally code replacements for whether, in Levelt's terms [10], they are appropriateness or error repairs. In error repairs, the reparandum is incorrect, whereas in appropriateness repairs, the reparandum is merely less appropriate in context than the alteration.

*Insertions* are a special case of repair whose surface forms can look similar to either hesitations or replacements. In an insertion, the repair adds further detail to some part of the utterance which has already been formulated. In some insertions, part of the OU is retraced in order to add the new material (e.g., "Well go straight down to the bottom of your page / almost to the bottom of the page", or, from Levelt's [10] data, "We beginnen in het midden met / in het midden van het papier met een blauw rondje (*we start in the middle with / in the middle of the paper with a blue disc*)"). These cases are not replacements because they do not delete any of the OU. Other insertions are attached to the OU fairly seamlessly (allowing for an editing phase), typically as added clauses ("It's three inches above it / if you go south"), added adverbials ("Go to your right / horizontally"), nouns of apposition ("you're next to banana tree / the word") or continuations from topical noun phrases ("Rockfall / I've got").[2] Blackmer and Mitton [2] do not count these as insertions, and they look like hesitations when they are written down. However, hearing them made us believe that the speakers had not conceptualised the added material until after they would have preferred to include it in the utterance and that the speakers would have performed seamed insertions if they had not found ways to add the new material after the OU. Our impression is that the intonation contour of an insertion's OU makes it appear to be a

[2]Following Miller [12], we consider "NP + clause" constructions to be grammatical in spoken English.

complete utterance, unlike that of a hesitation's OU, which appears to signal that the speaker will continue. Since the seamed and seamless insertions arise in much the same way, we have included them in a single class. Note that it is possible for the repair to merely insert words into the OU without the repair being an insertion, if the alteration does not just modify the existing structure of the OU. For instance, the repair "Right, you've got / I don't suppose you've got the baboons," would be classified as an appropriateness replacement, because "I don't suppose" wouldn't ordinarily be said to modify some part of the OU. Since no part of the OU is replaced in an insertion, technically the reparandum is empty; for convenience, we define the reparandum of an insertion to be that part of the OU which the alteration modifies.

## 2.3 Coding the editing phase

All repairs were coded for what sort of material occurred during the editing phase, including (1) periods of silence, (2) "uh" or "erm" sounds, (3) "sort of" or "kind of" (in cases where they did not appear to be included to make the following description less precise), (4) prospective repetition of material from the end of the OU or from the partner's previous utterance (see Hieke [4]), (5) "waffle" words which did not contribute to the meaning of the utterance except possibly as discourse markers (such as "hold on" and some instances of "well" and "right"), and (6) "other" (infrequent editing terms and behaviours such as grunts which could not easily be classified). Editing phases were also coded for whether they were projected normally or *sotto voce*. We did not maintain a strict division between editing terms and filler. Editing terms signal that an error has occurred and that a repair will commence, and filler is other material which serves to stall for time. Although we agree some behaviours divide well along these lines (for instance, "oh" and "sorry" as editing terms (James [7]) and "erm", waffle, and prospective repetition as filler), "uh" sounds, traditionally editing terms, proved impossible to classify.

In addition to noting information about the editing phase itself, we also noted information about how it related to the material surrounding it. We noted when the end of the word or part word preceding the MOI seemed lengthened or clipped, and also roughly whether we heard the editing phase as interrupting the intonation contour of the utterance or breaking up the rhythm of the utterance. Lengthening of continuants and nasals, if it goes on for long enough, gains the speaker time in much the same way as a period of silence. Clipping (by which we mean breaking off the speech stream prematurely) and breaks in the prosody sometimes appear to signal a disfluency, although usually in the presence of other cues. These behaviours were coded by ear rather than by a more rigorous method, simply to note the reasons why we thought a disfluency had occurred.[3]

Our characterisation of editing material is rather more liberal than that of previous collectors, which leads us to count more occurrences of both unfilled and filled hesitations than we would under their definitions. We allow unfilled hesitations to be signaled by excessive lengthening even if no period of actual silence occurred because some instances of lengthening appear to stall for time in much the same way as editing material. One-sixth of the unfilled hesitations (31 of 188) are coded as being signaled in this way. Paradoxically, we have five instances of unfilled hesitations for which the only coded signal is clipping. Allowing "waffle" and "sort of" filler cases inflates our filled hesitation count slightly compared to other filler definitions. For instance, we consider "*Ten inches? Get away... I don't know if I'd say that...* Do you have the abandoned truck" to be a filled hesitation, partly because the italicised material is spoken *sotto voce*. 14 filled hesitations out of 249 were coded as involving only these kinds of filler; one of these would be recoded as an unfilled hesitation if these fillers were disallowed.

[3]For a more thorough approach to the relationship between prosody and disfluency, see Nakatani and Hirschberg [13].

## 2.4 Coding the Repair's Main Components

Marking an editing phase in a transcript automatically marks the end of an OU and the beginning of the accompanying repair. We did not explicitly mark the beginning of the OU or the end of the repair, although Levelt defines them as the previous and following "sentence" boundary. For replacements, we marked part of the OU as a reparandum and part of the repair as an alteration; in cases where the entire OU was abandoned, the reparandum was the entire OU and the alteration was the entire repair. For insertions, we defined the reparandum to be the element of the OU which the insertion modified and the alteration to be the modification. For instance, for the insertion "Well go straight down to the bottom of your page/ almost to the bottom of the page", the reparandum is "to" and the alteration is "almost". Hesitations do not contain reparandums and alterations by definition. For some repairs, we thought we knew roughly what the OU would have been if the speaker had not noticed the trouble or if the delay before detection had been longer. For instance, in the repair "If you go round in a s-/ kind of semicircle", it seemed clear that the word at the end of the OU would have been "semicircle". In these cases, we indicated what we thought the completion would have been because this makes the classifications easier to justify.

Retracing is indicated only indirectly by overlap between the material preceding the OU (or, in the case of hesitations, the end of the OU) and the beginning of the repair. The amount of delay before detection of trouble can be determined from the number of syllables of the OU which follow the reparandum. However, we did specifically code whether the MOI immediately followed trouble or whether there was a delay. In addition, since Levelt [10, 11] relates whether or not the word at the end of the OU is completed to whether or not it belongs in the target utterance, we specifically coded whether we heard the MOI as occurring between words or within a word.[4]

In addition, some attempt was made to code the syntax of the OU, repair, reparandum, and alteration. Using a set list of syntactic categories, we coded the types of the largest complete constituents at the end of the OU, at the beginning of the repair, and in the reparandum and alteration. For instance, in the unfilled hesitation "but stay quite close to/ the granite quarry", the end of the OU is coded as a preposition, and the beginning of the repair as a noun phrase; there is no reparandum or alteration. The syntactic categories are given in appendix B. The intention was that this information could be used to investigate both which constituents were most often associated with trouble and how much retracing took place in a repair (see Levelt [10, 11], van Wijk and Kempen [18], Langer [9], Pizey [15]). However, the codings which currently appear in the files are not very accurate and are not complete. Errors in the coding are usually caused by choosing too small a component (e.g., in the example, coding the repair syntax as a determiner) or by difficulties in applying the syntactic categories to specific examples.

## 2.5 Coding the source of the trouble

Because we are interested specifically in how repairs arise out of pressure on the speaker, especially during message generation, we have included information about what characteristics of the message the speaker had already determined when he produced the OU, and which, if any, of these the repair changed. For instance, in the repair "Have you got a fallen/ you've got a fallen cairn, haven't you?", the speaker seems to have originally decided to ask a question about the partner's knowledge without prejudicing the answer, and in the repair, changed the query to a form expecting a "yes" answer.[5] Although there are many different decisions which are potentially interesting, we only divide them into rough categories. In the map task, referring expressions for landmarks do not often require conceptualisation

[4]This can be a difficult distinction to make, especially in cases where a partial word includes any part of the final syllable, and thus any work which uses this coding should make some comment on its reliability. It is not clear how Levelt overcame this difficulty.

[5]This sort of information is potentially interesting because it may suggest the order in which the speaker made decisions about the message or the kinds of decisions which are most likely to cause trouble.

because landmarks have labels which are written on the map. This makes conceptualisation quite simple for most utterances, at least once the function of the utterance has been decided. The exception is utterances which describe a section of the route or the position of a landmark. Accordingly, we use a richer set of distinctions for these utterances.

For each repair, we classified decisions in each category as

**(1)** not made before the MOI,

**(2)** made before the MOI and retained in the repair,

**(3)** made before the MOI and changed by the repair (i.e., a decision which is a source of the trouble),

**(4)** made before the MOI and made obsolete because of some other trouble, or

**(5)** not applicable in the current utterance.

Note that these classifications are often not obvious; for instance, in the repair "If you move/ uh... / left to the edge of your page...", the speaker could just be having difficulty naming the direction "left", or he could be having difficulty deciding how to describe where to move at all. Even in cases where the coder had a reasonable idea of why the repair occurred, it was was easy to confuse, for instance, categories (3), (4), and (5). Moreover, in dialogues which were coded early in the study, categories (3) and (4) are not distinguished. Thus any work which used this part of the coding would have to be careful about the claims which were made for it. We stopped coding this information for unfilled hesitations part of the way through the data collection specifically because they provide the least information about the source of the trouble.

Our coding uses the following categories:

**Move Type** This category records decisions about what basic communicative goal the message has. Utterances in dialogue are characterised by the communicative goal or intention (for instance, to ask a question or to give some information to the partner), often called the "dialogue move" (see, for example, Power [16], Houghton and Isard [5], and Clark and Schaefer [3]). Dialogue move analyses can be given at varying levels of detail; for these purposes, we use the twelve move types which Kowtko et al. [8] have set up specifically for the map task and which distinguish, for instance, instructions, spontaneous explanations, yes/no questions, wh-questions, and acknowledgements of something the partner has said. Intuitively, move type is often the first decision made about a message, since there are typical patterns of moves (often called "games") which occur because they fulfill the goals most effectively (e.g., questions, which request information, are usually followed by answers, which give the information requested). However, it is possible for speakers to make other decisions first; for instance, if one is unsure of a fact, one may know that one wants to express it but not whether to tell it to the partner or to ask the partner if it is true. Speakers also quite often change their minds about which dialogue move to make, usually because they have discovered some other goal which must come prior to the one which they chose first. In the repair "The start is/ do you have the extinct volcano?", the OU, if completed, probably would have been something like "The start is to the left of the extinct volcano", but the speaker realised that it would be better to find out whether or not the partner had the extinct volcano first. (Of course, the speaker will ordinarily return to the original goal once the goal of the repair is out of the way, and so these repairs often only temporarily interrupt the structure of the dialogue.) Thus even a very basic decision such as the dialogue move is open to indecision and change by repair.

**Speech act** Even once a basic dialogue move type has been chosen, there are still many possible surface forms; for instance, in asking a yes/no question one can choose an untagged, positively tagged, or negatively tagged form. Similarly, in giving instructions there is a choice between, for instance, imperative and subjunctive forms. Choosing a speech act can be seen as fine-tuning the choice of intention for secondary purposes such as showing an expectation for a particular response or being polite. This category records decisions about which surface level speech act to use.

**Overall semantics** This category records decisions about the overall meaning of an utterance. For utterances which describe a section of the route or the position of a landmark, this category is used to record whether or not the speaker had decided what section of the route (i.e., how large a part of the path) or which landmark to describe by the time of the MOI. For non-descriptive utterances, it is used to record whether or not the speaker had an overall conceptual structure in place for the utterance. For most non-descriptive utterances, this will imply that all conceptualisation decisions have been made; however, in utterances with complex referential decisions, further information is recorded in the tactics category.

**Description strategy** In the map task, speakers have a choice between absolute descriptions in terms of directions and distances (such as "go up three centimetres") and relative descriptions in terms of landmarks or other prominent features (such as "go between the palm beach and the swamp" or "go to the middle of the page"). Speakers may also choose some hybrid between the two. Intuitively, since map task participants appear to have trouble both constructing and understanding absolute descriptions, the choice largely depends on whether or not there are suitable landmarks on the map, and so if a relative description is chosen, the speaker will already know which landmarks to use. This category is only used for descriptive utterances, and records whether or not the speaker appears to have decided between an absolute or relative description. For other utterances, it should be marked as n/a.

**Tactics** Whether a speaker is constructing an absolute or relative description, there are still many other choices necessary to fill out the conceptual structure of the utterance. For absolute descriptions, the speaker must choose the way of expressing directions (e.g., up/down/left/right vs. N/S/E/W), the measure (i.e., centimetres or inches), and the distance. For relative descriptions, the speaker must decide what relationship expresses the one on the map (e.g., "around", "to", "near"). This category records whether or not speakers appear to have made these decisions by the MOI, and if they have, whether or not the repair maintains them; if they have, then the conceptual structure for part of the utterance being realised should be complete. It is used primarily for descriptive utterances, although it has also been used to record decisions about reference choices in cases of non-descriptive utterances with complex reference.

**Basic syntactic structure** In incremental sentence production, the formulator begins working on concepts as soon as the conceptualiser suggests them, choosing words to realise the concepts and gluing them together by ordering them and adding functor words in order to create a series of motor commands which match the emerging conceptual structure. Syntactic structure largely depends on which concepts are ready first and how quickly the speaker must perform. We expect that repairs which change the syntactic structure of an utterance usually arise when articulation begins before conceptualisation is complete, forcing the speaker into either a dead-end or a highly dispreferred construction. However, some may arise out of errors in the formulator itself. This category records whether the repair maintains or changes the syntactic structure of the OU.

**Realisation** This category records any other formulation and articulation decisions; where it is marked as having changed, it usually involves changing a bad lexical choice or correcting a problem with articulation such as using the wrong phoneme.

**Other** This category is used to record any changed decisions which are identifiable but which don't appear to fit into any of the defined categories. It is marked as n/a for most repairs.

In addition, the coding identifies the "trigger" category for the repair, which we think of as the kind of decision with which the repair is "most closely associated". For replacements, if we were to think of decisions as being made by one process for each of the rough types we have, the trigger would be the process by which we discover that a repair is necessary. For example, in the repair "The start is/ do you have the extinct volcano", every decision for the OU including the move type is thrown away, but what probably causes the repair is the decision to use the extinct volcano to describe the location of the start, since at that point the speaker will realise that it is better to establish whether or not the partner has the extinct volcano first. Of course, in most cases it is not the message generation processes at all which discover the need for a repair, but the speaker's self-monitor. In these cases we mark the trigger as the category with whose decisions the self-monitor has found fault; for instance, if the problem is incorrect lexical choice, then the trigger is coded as "realisation".

# 3 The Collection

Two coders were involved in the data collection. We developed our coding scheme on one dialogue and then each coder independently analysed a subsequent dialogue, so that we could establish the reliability of the coding. Our corpus of repairs includes a corrected joint analysis of the first dialogue and a corrected version of the primary coder's analysis of the second dialogue, plus seven further dialogues from the primary coder and one further dialogue from the secondary coder. In turns with speaker overlap, we only completed the coding up until the point where the overlap occurred, since it was difficult to hear and interpret repairs after that point. We included map task trials in both of the eye contact and familiarity conditions. We found 723 repairs in the ten dialogues; an overall summary of their distribution is given in figure 3. For each dialogue, it gives whether or not the participants (a) had eye contact and (b) were familiar with each other; the number of coded (1) filled hesitations (2) unfilled hesitations, (3) replacements, (4) insertions, and (5) words in the entire dialogue; (6) the number of repairs per hundred words, excluding unfilled hesitations and counting only those which occur in non-overlapped turns; (7) the percentage of turns which contained such repairs; and (8) the percentage of turns which were not wholly analysed because of speaker overlap. Columns (6) to (8) are meant to give an idea of rate of repair in the dialogues; we exclude unfilled hesitations because it is difficult to collect them reliably.

# 4 Reliability of the Coding Scheme

To date, we have only considered the reliability of judgments about whether or not a repair occurs during a particular dialogue move or whether or not one occurs between two particular words. In order to establish this reliability, the two coders independently analysed one dialogue (naq8c1). We define "dialogue move" to be those coded for the map task and defined in Kowtko et al. [8]. The map task transcripts separate the speech into words, using a special way of marking items which are not properly words. The type of these items is coded to show, for instance, whether they are part words or grunts. For our purposes, we define "word" to be either a contiguous set of alphabetic characters on the transcript including hyphens, or a marked item which is coded as an abandoned word (e.g., "carpen-") or quasi-phonetic "word" (e.g., "hideup", "jubbely", and "uh-huh"). We exclude marked items which are coded as partial repetitions (e.g., "-boks", "-doned") or quasi-phonetic forms which typically appear in editing phases (e.g. "uh", "oh", "eh", "erm"). Given this

Figure 3: A Summary of the Data

| | (a) eye? | (b) fam? | (1) f.h. | (2) u.h. | (3) rep | (4) ins | (5) words | (6) rate | (7) %turns | (8) %over |
|---|---|---|---|---|---|---|---|---|---|---|
| eaq3c3 | yes | yes | 32 | 23 | 27 | 12 | 1139 | 9.06 | 38.10 | 26.06 |
| eaq3c4 | yes | yes | 25 | 11 | 21 | 7 | 987 | 6.97 | 32.67 | 19.20 |
| eaq4c1 | yes | no | 10 | 25 | 5 | 1 | 437 | 9.90 | 41.07 | 9.68 |
| eaq4c2 | yes | no | 78 | 20 | 35 | 14 | 1438 | 10.91 | 38.73 | 14.97 |
| eaq8c6 | yes | yes | 24 | 42 | 40 | 20 | 1936 | 7.18 | 31.84 | 13.12 |
| naq2c6 | no | no | 21 | 14 | 16 | 4 | 648 | 8.75 | 36.36 | 11.49 |
| naq6c5 | no | yes | 15 | 8 | 26 | 4 | 975 | 4.46 | 17.76 | 32.70 |
| naq8c1 | no | yes | 16 | 13 | 13 | 6 | 1175 | 4.23 | 19.23 | 27.78 |
| naq8c7 | no | no | 9 | 15 | 10 | 4 | 976 | 4.39 | 19.19 | 33.56 |
| naq8c8 | no | no | 19 | 17 | 15 | 6 | 980 | 6.22 | 29.46 | 21.13 |
| All givers | | | 208 | 149 | 151 | 67 | 7172 | 8.57 | 43.88 | 22.07 |
| All followers | | | 41 | 39 | 57 | 11 | 3519 | 4.79 | 16.84 | 20.55 |
| TOTAL | | | 249 | 188 | 208 | 78 | 10691 | 7.34 | 30.31 | 21.32 |

Figure 4: Raw category judgments by word

| BY WORD | second coder's judgment | | | | |
|---|---|---|---|---|---|
| | filled hesitation | unfilled hesitation | replacement | insertion | fluent |
| first coder's judgment | | | | | |
| filled hesitation | **8** | 0 | 0 | 0 | 5 |
| unfilled hesitation | 0 | **8** | 0 | 0 | 1 |
| replacement | 2 | 0 | **6** | 0 | 1 |
| insertion | 0 | 0 | 1 | **4** | 1 |
| fluent | 1 | 7 | 2 | 2 | **765** |

N= 828

definition, our coders rarely placed equivalent repairs between different words, although they did sometimes disagree about whether or not a repair had occurred at all.

Figure 4 summarises the agreement of the two coders on whether or not a repair occurs between two particular words.

In the figure, numbers in cells which are not along the main diagonal correspond to coder mismatches. Most of the large values in these cells were caused by coding difficulties, undiscovered in the dialogue used to develop the coding scheme, which were subsequently corrected. Locations which were coded as filled hesitations by the first coder and fluent by the second coder tended to have fillers which were present on the tape but absent in the transcript. Most locations which were coded as replacements by one coder and not the other were cases where there was disagreement about the meaning of the original utterance. For instance, if a speaker began a turn with "OK" or some other confirmation and subsequently indicated a misunderstanding, one coder might have taken "OK" to simply acknowledge the partner's turn, whereas the other coder might have understood it only as acceptance of that turn's content. In this case, only the first coder would have marked a replacement. These difficulties were corrected by explicitly telling the coders that filled hesitations must be identified by ear because the transcripts could not be expected to be accurate to that level, and by requiring replacements to alter the weakest plausible meaning of the original

Figure 5: Reliability of fluency judgment by move

| BY MOVE | second coder's judgment repaired | fluent |
|---|---|---|
| first coder's judgment | | |
| repaired | 14 | 4 |
| fluent | 3 | 112 |

N = 133
Percent agreement = 94.7%
Less than .1% chance that two random coders making fluent/repaired judgments in the given proportions would reach this level of agreement (Fisher exact probability).

utterance.

The remaining large discrepancy is between locations which the first coder classed as fluent and the second coder marked as unfilled hesitations. We have found no consistent explanation for the difference, and conclude that the method of collecting unfilled hesitation data used for this dialogue is unreliable. Identifying unfilled hesitations is difficult because not all pauses are caused by trouble in deciding what to say. For instance, even in read speech, pauses are often used to mark syntactic boundaries and add dramatic emphasis. In spontaneous speech, pauses sometimes occur because the speaker is monitoring something else which is happening in the environment, thinking about what the partner has said, or letting his actions catch up with his words (as in the map task utterance, "I'm at the right... place... now".) It is especially difficult to decide whether or not short pauses count as hesitations. Accordingly, we collected only gross classification data for unfilled hesitations in the remaining dialogues.

If we discard unfilled hesitations, the judgment of whether or not a repair occurs between two particular words or during a particular dialogue move appears to be sufficiently reliable to use. Reliability is frequently cited as some form of percent agreement (e.g., Passonneau and Litman [14], TOBI [17], Kowtko et al. [8]), such as the percentage of judgments on which two coders agreed. in this context, this is misleading because repairs are coded during only approximately 5% of moves and after only approximately 3% of words. For instance, the distribution of agreement rates for two coders who each randomly mark 5% of moves as containing repairs would centre on 90%. Thus, figures 5 and 6 give both the percent agreement between the two coders and the result of a contingency table test showing a chance of less than < .1% that this rate of agreement would occur for random coders who code the same number of repairs as the real coders.[6]

# 5 Summary

In this phase of our research, we have defined a way of coding many interesting aspects of self-repairs which (1) identifies where the repairs occur on the transcripts, (2) describes what material interrupts the message during editing and how the editing phase relates to the utterance prosodically, (3) specifies how the major components of the repair are related to each syntactically and semantically, and (4) suggests what sort of speaker difficulty is

[6]These figures omit judgments in turns which involved speaker overlap. Although we collected those repairs in overlapped turns for which the MOI clearly occurred before the beginning of the overlap or for which a complete clause had occurred between the end of the overlap and the MOI, the transcripts only locate the overlaps by turn, and so we cannot make judgments about fluency for individual words and moves within overlapped moves with the information that we have. Overlap occurred in 25% of the turns in the test dialogue.

Figure 6: Reliability of fluency judgment by word

| BY WORD | second coder's judgment repaired | fluent |
|---|---|---|
| first coder's judgment | | |
| repaired | 21 | 5 |
| fluent | 7 | 781 |

N= 814
Percent agreement = 98.5%
Less than .1% chance that two random coders making fluent/repaired judgments in the given proportions would reach this level of agreement (Fisher exact probability).

likely to have caused the repair. This coding has been completed for ten map task dialogues, varying the eye contact and familiarity conditions. We have shown that judgments about whether a repair occurs between two particular words or during a particular move are reliable; however, further work must be done to establish the reliability of other aspects of the collection. We expect that the coding will be useful both to people interested in specific aspects of repair (such as, for instance, validating Levelt's main interruption rule [10] or investigating how repairs attach syntactically to the utterances they modify) and to those interested in how spoken and written language differ.

# A The data collection form

The repairs are organised in one directory per dialogue. Each repair is described in a separate data file in terms of a set of field names and values. The files contain one field per line in any order; each line contains the field name, a tab, and the field value. The fields and their permitted values are given below. Fields for which no value has been specified do not contain anything after the tab or do not occur in the data file; fields with null values are specified by empty square brackets ([]). The name of a data file is given in the form "x.y_z.`data`", where x is the value of the dialogue field, y is the value of the turn field, and z is the value of the index field. We divide the fields into subsections which match the discussion in section 2. There are several comment fields, and coders did always use the most relevant one.

## A.1 Identifying and locating the repair

`Dialogue:` the name of the dialogue, given in abbreviated form (e.g., EAQ3C3).

`Turn:` positive integer giving the number of the turn as given in the transcript file on the *T line immediately preceding the turn's text.

`Index:` positive integer giving the number of this repair in the turn, starting with one.

`Speaker:` `giver` or `follower`. This information can also be inferred by looking at the *T line in the transcript with the correct turn number, since *TA indicates the giver and *TB, the follower.

`Tape_Counter:` positive integer giving the position of the repair on our copies of the tapes, using our memoscriber. It should now be easier to use the CD to find and play the speech.

`Word_Count:` positive integer giving the number of the "word" boundary at which the MOI occurs, counting the boundary before the first word as zero. For these purposes, "word" is defined as (1) any sequence of characters beginning with { up to and including } and trailing non-whitespace characters, or (2) a non-empty sequence of non-whitespace characters. Punctuation marks, including the angle brackets used to mark overlapped turns, are considered whitespace for this purpose. This definition is meant only to facilitate finding the MOI and not as a serious word count. The first condition counts annotated non-words such as "{m Erm}" as one word, including any trailing punctuation. The position counts were constructed on a version of the transcripts which did not include gaze coding, so any codings beginning with at signs (@) must be stripped out before counting begins.

`Move:` a positive integer giving the number of the move in which the repair occurs from the *M transcript coding, numbering from one. This information is only available for some of the dialogues.

`Overall_Comments:` free text on any subject; since it tended to be filled in last, it was generally used for other comments which had been missed out accidentally or to mark aspects of the repair which are unusual.

## A.2 Classifying the repairs

`Type:` `unfilled_hesitation`, `filled_hesitation`, `replacement`, or `insertion`, as defined above.

**Replacement_Type:** appropriateness, error, or ?, coding whether the repair constructs a more appropriate utterance than the completed OU, whether it corrects an error in the OU, whether it is a recognizable type of repair which does not fit into either category, or whether it is impossible to classify. This field only occurs for replacements.

**Classification_Comment:** free text.

## A.3 Coding the editing phase

**Editing:** any combination from the set {none, silence, ah, erm, sort_of, lengthen, repetition, waffle, ?}, with any amount of whitespace between the designators, giving information about what occurs in and just before the editing phase. Lengthen indicates lengthening of the final word of the OU. ah should be taken to include a range of sounds including "eh", "uh", "aaaah", and short grunts. Sort_of applies only for occurrences of "sort of" which appear primarily to stall for time. Cases of waffle and ? will usually have accompanying comments; for instance, stalling uses of discourse markers such as "well" and clipping are coded as ?.

**Editing_Audibility:** sotto, projected, or n/a, depending on whether there is any editing material which can be projected and whether or not such material is projected or said *sotto voce*.

**Intonation:** n/a, terminal, interrupted, unbroken, or ?, expressing the relationship between the intonation of the editing phase and the OU. N/a is used for cases where the OU is too short to establish a pattern; terminal means that the end of the OU sounds like the end of an utterance. The unbroken and interrupted codings distinguish whether or not it would sound like a fluent utterance, if one could only hear the contour. This information has only been collected quickly "by ear" and should only be taken as a rough guide. Note that for seamless insertions, an "unbroken" contour on the OU will sound terminal; these cases are meant to be coded as unbroken but the coders were not always consistent.

**Rhythm:** n/a, broken, unbroken, or ?, giving an intuitive feel for whether the speech rate was consistent between the end of the OU and the editing phase. N/a is used for cases where the OU is too short to establish a pattern. In general, the rhythm will be coded as broken if and only if the editing phase is coded as containing lengthening or silence.

**Editing_Comments:** free text describing any interesting uncoded information about the editing phase.

## A.4 Coding the Repair's Main Components

**OU:** free text giving the OU of the repair, with hypothetical completions given in square brackets. This representation of the OU is only intended to provide enough information to find the repair in the transcript and to give an idea of what the repair was about; in particular, material at the beginning of the OU may be omitted. In addition, the OU might not be given exactly as it occurs in the transcript but might gloss the effects of previous repairs; these glosses also usually occur in square brackets.

**Repair:** free text giving the repair, possibly with material at the end of the repair omitted and glossing the effects of subsequent repairs.

**Reparandum:** free text giving the reparandum, which must be part of the OU. The reparandum for hesitations must be []. The reparandum for replacements is the set of words being replaced and the reparandum for insertions is the word or set of words to which the insertion attaches.

**Alteration:** free text giving the alteration, which must be part of the repair. The alteration for hesitations must be [].

**Target:** free text giving an utterance which conveys the finally intended meaning that combines the OU and the repair. Material in square brackets glosses over the effects of previous repairs.

**MOI_Placement:** `correct_word`, `incorrect_word`, `between_words`, or `?`, giving where the MOI comes in relation to the word structure of the speech stream. This information is intended for a comparison to Levelt's main interruption rule [10, 11].

**Delay:** `immediate`, `delayed`, `predicted`, or `?`, giving whether the MOI occurs immediately at the point where trouble is realised in the speech stream, is delayed (so that there is material in the OU which occurs after the reparandum), or is anticipated, presumably by monitoring internal speech, and the trouble spot has not yet been reached. This decision is much less reliable for hesitations and insertions than for replacements simply because it is more difficult to tell why the repair occurred.

**General_Comments:** free text commenting on the general form of the repair.

In addition, there are fields coding the syntax of the OU, repair, reparandum, and alteration. The following fields take the values given in appendix B.

**OU_Syntax:** the syntactic type of the largest complete constituent at the end of the OU.

**Repair_Syntax:** the syntactic type of the largest complete constituent at the beginning of the repair.

**Reparandum_Syntax:** the syntactic type of the largest complete constituent contained in the reparandum.

**Alteration_Syntax:** the syntactic type of the largest complete constituent contained in the alteration.

## A.5 Coding the source of the trouble

We included fields to represent the different categories of decisions defined in section 2.5. These fields can have the values `n/a`, `undecided`, `decided`, `changed`, `thrown_away`, or `?`. Each category also has an associated free text comment field. The names of the category fields and their associated comments are as follows:

`Move_Type` and `Move_Comment`;
`Speech_Act` and `Speech_Act_Comment`;
`Strategy` and `Strategy_Comment`;
`Tactics` and `Tactics_Comment`;
`Syntax` and `Syntax_Comment`;
`Realisation` and `Realisation_Comment`; and
`Other_Stage` and `Other_Stage_Comment`.

In addition, the `Trigger` field contains one of the following designations for the defined categories: `move`, `act`, `semantics`, `phrase_strat`, `tactics`, `grammar`, `realisation`, `other`, or `?`. These names only differ from the category names to provide unique first letters. The field `Trigger_Comments` contains free text commenting on the choice of trigger category.

# B The Syntactic Categories

The following syntactic categories are used for all syntactic coding in the data files. Categories 0-A are usually used as terminal categories and B-J as nonterminals.

`0-empty` The category for null ([]) text.

`1-vb` Main verb.

`2-vb[aux]` Auxiliary verb.

`3-prep` Preposition.

`4-det` Determiner.

`5-noun` Noun.

`6-pron` Pronoun.

`7-qual` Qualifier (modifier which serves to qualify another category, such as "probably" and "around").

`8-quant` Quantifier.

`9-adj` Adjective.

`A-conj` Conjunction, including discourse markers such as "well" and subordinating conjunctions such as "because".

`B-clause` Any clause, including imperatives, declaratives, and interrogatives.

`C-prepg` A group of words which together serves as a complex preposition; e.g., "to the left of".

`D-pp` Prepositional phrase.

`E-noung` A group of words which together serves as a complex noun. The intention was that this should describe noun phrases stripped of their determiners.

`G-np` Other noun phrase.

`H-verbg` A group of words which together serves as a complex verb. The intention was that this should describe verb phrases stripped of their objects.

`I-vp` Verb phrase.

`J-parenthetical` A phrase which serves to interrupt the rest of the utterance.

The following examples give the intended syntactic structures for some common utterance forms.

*Go to the left of the swamp.*

| | | | | | | |
|---|---|---|---|---|---|---|
| go | to | the | left | of | the | swamp |
| vb | prep | det | noun | prep | det | noun |
| verbg | | | noung | | | noung |
| | | np | | | np | |
| | | prepg | | | | |
| | | | pp | | | |
| | | | clause | | | |

*Do you have a swamp?*

<table>
<tr><td>do</td><td>you</td><td>have</td><td>a</td><td>swamp</td></tr>
<tr><td>vb[aux]</td><td>pron</td><td>vb</td><td>det</td><td>noun</td></tr>
<tr><td></td><td>np</td><td>verbg</td><td></td><td>noung</td></tr>
<tr><td></td><td></td><td></td><td colspan="2">np</td></tr>
<tr><td></td><td></td><td colspan="3">vp</td></tr>
<tr><td colspan="5">clause</td></tr>
</table>

*Go up about five centimeters.*

<table>
<tr><td>go</td><td>up</td><td>about</td><td>five</td><td>centimeters</td></tr>
<tr><td>vb</td><td>prep</td><td>qual</td><td>quant</td><td>noun</td></tr>
<tr><td>verbg</td><td>prepg</td><td></td><td></td><td>noung</td></tr>
<tr><td></td><td></td><td colspan="3">np</td></tr>
<tr><td></td><td colspan="4">pp</td></tr>
<tr><td colspan="5">clause</td></tr>
</table>

N.B. “About” attaches to “five centimeters”, not “five”, because it serves to mark the distance given as approximate.

*So you go down probably an inch or two.*

<table>
<tr><td>so</td><td>you</td><td>go</td><td>down</td><td>probably</td><td>an</td><td>inch</td><td>or</td><td>two</td><td>[]</td></tr>
<tr><td>conj</td><td>pron</td><td>vb</td><td>prep</td><td>qual</td><td>det</td><td>noun</td><td>conj</td><td>quant</td><td>noun</td></tr>
<tr><td></td><td>np</td><td>verbg</td><td>prepg</td><td></td><td></td><td>noung</td><td></td><td></td><td>noung</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td colspan="2">np</td><td></td><td colspan="2">np</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td colspan="5">np</td></tr>
<tr><td></td><td></td><td></td><td colspan="7">pp</td></tr>
<tr><td></td><td></td><td colspan="8">vp</td></tr>
<tr><td></td><td colspan="9">clause</td></tr>
<tr><td colspan="10">clause</td></tr>
</table>

*It's at the bottom left hand side of the saxon barn.*

*it* *'s* *at* *the* *bottom* *left-hand* *side* *of* *the* *saxon* *barn*

*pron* *vb* *prep* *det* *adj* *adj* *noun* *prep* *det* *adj* *noun*

*np* *verbg* *prepg* *prepg* *noung*

*np*

*pp*

*noung*

*np*

*pp*

*vp*

*clause*

*And then do a sharp turn to the right towards the slate mountain*

and then do a sharp turn to the right towards the slate mountain

conj conj vb det adj noun prep det noun prep det adj noun

conj verbg prepg noung prepg noung

np np

pp pp

noung

np

clause

clause

## C An Example Data File

The following is file eaq3c3.114_1.data.

```
Turn    114
Speaker giver
Index   1
Tape_Counter    323
Word_Count      7
OU      just go about the middle of the
OU_Syntax       4-det
Repair  bottom line
Repair_Syntax   5-noun
Editing silence sort_of
Editing_Audibility      projected
Reparandum      []
Reparandum_Syntax       0-empty
Alteration      []
Alteration_Syntax       0-empty
Target  just go about the middle of the bottom line
General_Comments        filler is "sort of", slightly softer
Delay   immediate
Classification_Comments
MOI_Placement   between_words
Intonation      interrupted
Rhythm  broken
Editing_Comments
Move_Type       decided
Move_Comment
Speech_Act      decided
Speech_Act_Comment
Semantics       decided
Semantics_Comment
Syntax  decided
Syntax_Comment
Strategy        decided
Strategy_Comment
Tactics decided
Tactics_Comment
Realisation     undecided
Realisation_Comment     "bottom line" is troublesome (lexical access)
Other_Stage     n/a
Other_Stage_Comment
Trigger realisation
Trigger_Comments
Overall_Comments
Move    1
Dialogue        EAQ3C3
Type    filled_hesitation
```

## References


[1] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. Thompson, and R. Weinert. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366, 1991.

[2] E. R. Blackmer and J. L. Mitton. Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39:173–194, 1991.

[3] H. Clark and E. Schaefer. Contributing to discourse. *Cognitive Science*, 13:259–294, 1989.

[4] A. E. Hieke. A content-processing view of hesitation phenomena. *Language and Speech*, 24(2):147–160, 1981.

[5] G. Houghton and S. Isard. Why to speak, what to say and how to say it: Modelling language production in discourse. In *Modelling Cognition*, pages 249–267. John Wiley and Sons, 1987.

[6] D. James. Some aspects of the syntax and semantics of interjections. In *Papers from the Eighth Regional Meeting of the Chicago Linguistics Society*, pages 162–172, 1972.

[7] D. James. Another look at, say, some grammatical constraints, on, oh, interjections and hesitations. In *Papers from the Ninth Regional Meeting of the Chicago Linguistics Society*, pages 242–251, 1973.

[8] J. C. Kowtko, S. D. Isard, and G. M. Doherty. Conversational games within dialogue. Technical Report HCRC/RP-31, Human Communication Research Centre, June 1992.

[9] H. Langer. Syntactic normalization of spontaneous speech. In *Proceedings of COLING-90*, volume 3, pages 180–183, 1990.

[10] W. J. M. Levelt. Monitoring and self-repair in speech. *Cognition*, 14:41–104, 1983.

[11] W. J. M. Levelt. *Speaking: From Intention to Articulation.* MIT Press, 1989.

[12] J. Miller. Discourse patterns in spoken English. In *Sheffield Working Papers in Language and Linguistics*, volume 1, pages 10–39, 1984.

[13] C. Nakatani and J. Hirschberg. A speech-first model for repair detection and correction. In *Proceedings of the ACL*, pages 46–53, 1993.

[14] R. J. Passonneau and D. J. Litman. Intention-based segmentation: human reliability and correlation with linguistic cues. In *Proceedings of the ACL*, pages 148–155, 1993.

[15] T. Pizey. Repair strategies in English. Master's thesis, University of Edinburgh, Department of Artificial Intelligence, 1993.

[16] R. J. D. Power. *A Computer Model of Conversation.* PhD thesis, University of Edinburgh, 1974.

[17] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. TOBI: A standard for labeling english prosody. In *International Conference on Speech and Language Processing (ICSLP)*, volume 2, pages 867–870, 1992.

[18] C. van Wijk and G. Kempen. A dual system for producing self-repairs in spontaneous speech: evidence from experimentally elicited corrections. *Cognitive Psychology*, 19:403–440, 1987.