Internal Use Only

002

TR-IT-0040

Structures in Spontaneous English Conversation

Laurel Fais

February 1994

This report details a comprehensive list of structures, grammatical and ungrammatical, found in spontaneous English conversation collected in the EMMI experiment. Specific suggestions are made for filtering non-contributing speech noises; for analyzing problematical grammatical structures and for handling discourse phenomena in a realistic way.

Contents

Introduction 1

EMMI experiment 1

Structures to be filtered 2

Filled pauses 3

False starts 3

Breaks 4

"OK" 4

"And" 6

Structures to undergo grammatical analysis 7

Structures involving discourse considerations 11

Coordination of the approaches 14

Future directions 14

Conclusion 15

Figure 1. Relative system dependencies 16

Appendix A: List of structures 17

Appendix B: Transcription of telephone conversation 23

Appendix C: Transcription of multi-media conversation 27

Bibliography 31

Introduction

It is commonplace for linguists and engineers alike in the field of machine translation to make broad statements about the nature of spontaneous speech. It is tempting to claim that natural native speaker conversation contains too many errors, too many non-standard structures, or too many fragmental utterances to be dealt with in a machine translation environment.

But is this actually the case? We made a thorough examination of two of the 16 conversations collected during an experiment set in the ATR Environment for Multimodal Interaction (EMMI) (Loken-Kim et al., 1993a, b) to catalogue exactly what kinds of structures were found in native-speaker, spontaneous speech. While it is certainly the case that such speech contains a number of problematical characteristics not found in formal or read speech, it is possible to make suggestions as to how these structures can be dealt with. Further, spontaneous speech contains features absent in formal speaking styles which might actually be used to advantage in machine translation. Seligman et al. (1993) explore the use of filled pauses and pauses for segmenting Japanese utterances in an automatic interpreting system. Some suggestions will be made below for a similar study for English.

This report presents an exhaustive listing of the structures contained in the two conversations examined and specific proposals for how to handle them in a natural language processing system. Of course, even if these proposals are fully implemented, they may not, in fact, be adequate for additional examples of spontaneous speech. However, the machinery that these proposals entail should be useful in a number of cases not specifically present in the conversations under examination here.

Below we will briefly describe the nature of the conversations examined and discuss the cataloguing of the structures found in those conversations. (The full listing of structures is given in Appendix A; the transcriptions of the conversations are given in Appendix B and Appendix C.) Then a multi-pronged approach is described for handling the structures, utilizing filtering, grammatical analysis, and illocutionary considerations.

EMMI experiment

The conversations were collected in an experiment comparing speaker performance by telephone only and via a multi-modal communication environment, using the ATR EMMI. North American native speakers of English were asked to imagine that they were arriving for the first time in Kyoto and had to find their way to a conference center. They called the "Conference Office" and talked to an "agent" (who remained the same throughout the experiments) in each of the two environments. Their speech was recorded and transcribed. Details concerning EMMI can be found in Loken-Kim et al. (1993a); details of the experiment can be found in Fais et al. (forthcoming).

Because the agent was practiced and the clients naive, it was felt that the clients' speech structures would be characteristic of a more natural, spontaneous speech style. For the work reported here, one of the clients was selected and the two conversations in which he took part were analyzed. The client was selected because he showed a relatively high percentage of self-initiated talk and thus it was possible to sample a greater range of structures in his conversation than in conversations in which the client took a more passive role.[1]

The conversations were examined by hand, using as a rough starting point the categories tabulated in Fais (1993c). Actually, the range of structures in the current investigation is not as broad as that of the previous work, probably because the conversations were more constrained.

Once the structures were identified, it became necessary to answer the question: how could each structure be handled in a natural language processing system? Looked at in this light, the structures seemed to fall more or less naturally into three groups: those that should be "filtered out" of consideration before grammatical analysis begins; those that are best described in grammatical terms; and those that involve discourse considerations. The structures that fall in each of these groups will be discussed below, and the coordination of the approaches suggested for dealing with each one will then be addressed.

Structures to be filtered

There is a certain amount of speaker production that makes no contribution to the conversation, and so can appropriately be filtered out of the analysis input. This statement, so baldly put, is clearly wrong; every uttered sound makes a contribution to a conversation, from the lowliest throat clearing to the most Latinate substance word. It may be more accurate to put the argument this way: at the level at which a language processing system will render the structure of its input strings, the fine-grained distinctions signified by such utterances as "ah" or "um," for example, will be unrecognized. Clearly it isn't the case that they make no contribution in conversation in the real world; they simply cannot make a contribution in the world of automatic language processing because language processing systems are not equipped to recognize and exploit their contribution. It is these utterances that we are proposing be filtered out of the input to the grammatical analysis component.

These utterances fall into three categories: filled pauses, false starts, and breaks in sentences: The first two types of utterance have been discussed in the literature as "disfluencies."

[1] In fact, the agent was the dominant speaker in all conversations; in the case chosen here, however, the client had a relatively greater number of utterances than other clients.

Filled pauses

"Filled pauses" refers to non-word sounds that speakers typically make to fill silence when taking time to consider a structure, lexical item, or conversational direction. In English, they typically take the form of "um," "ah," "oh," and somewhat less frequently, "er." Until recently, these utterances had to be filtered by hand , but some attempts are currently being made to recognize and filter them automatically out of input to structural analysis (Shriberg et al., 1992; Woszczyna et al., 1994; Kikui, personal communication). The general approach is to recognize filled pauses as words are recognized and to eliminate them via a look-up process that identifies them as belonging to a class of "unnecessary words."

False starts

The term "false starts" as used here actually combines three categories distinguished by Nakatani and Shriberg (1993): repetitions, repairs, and fresh starts. In each case, initially uttered material is "replaced" by a following utterance. In the case of repetitions, the replacement is identical to the original:

Client (C): OK **there's a** **there's a** taxi stand in the bus station[2]

In the case of repairs, the replacement corrects a lexical item:

Agent (A): ...a lot of foreign people stay there ... and they go **they visit** there

In the case of fresh starts, the replacement material corrects a phrase or sentential level construction:

C: I'm trying to figure out how to get to the International Conference Center where **I where** um **the conference** is I believe

Kikui has also tentatively proposed an approach for dealing with these kinds of problems, using the concept of "keys." The language processing system searches morphologically labelled data for "keys," which are pairs of words (w1, w2) such that w1 constitutes a (proper or improper) substring of w2. The search takes place within a "window" of some length of a specified number of words. Once the keys are located, their context is searched forward and backward for more keys, until a pair of maximally similar strings is identified. The system then "ignores" or "discards" the first string of the pair. Refinements to the procedure include retaining the similar strings if they are involved in some two-argument structure such as coordination or a "from...to..." type

[2] All examples given here are taken from the two EMMI conversations analyzed for this report, unless noted otherwise.

structure.

Breaks

While this procedure may be productive in cases where there is some repetition, it will not be effective in the case of breaks, in which there is no repetition. Breaks are characterized, in fact, by the *lack* of continuity between the initial, discarded utterance and the following re-start:

A: OK I'm going to ah **where are you located in the station** do you know

It is possible that prosodic and/or discourse information can be used to identify these cases so that they can be eliminated from the input to structural analysis, but there are no definitive results in this area as yet.[3]

"OK"

For the most part, the filtering approaches described above use some sort of pattern-matching-based algorithm to identify the structures involved. In addition, those structures are unambiguously redundant, that is, they are filtered out of the input to grammatical analysis in *all* their occurrences. However, there are some conversational elements which have multiple uses, only some of which need to be filtered out. One such element is "OK."

Unlike in the case of repetitions or corrections, the identification of the word "OK" is relatively straightforward given accurate speech recognition. However, also unlike repetitions or corrections, the decision to filter out "OK" does not depend solely upon that recognition. Instead, it requires additional information regarding the contextual use of "OK" as well.

"OK" occurs very frequently in the conversations examined, an average of once for about every three turns. In these conversations, the use of "OK" performed three major functions. The least frequent, accounting for two of the 35 uses of "OK" examined, is the use of "OK" with question intonation in order to ask whether the hearer understands, acquiesces to, or agrees with the previous utterance. These instances are clearly marked by intonation, make a significant contribution to the conversation, and as such, need to

[3] The first step in designing a language processing system that will handle spontaneous speech is to make the grammatical analysis more flexible so that it can accommodate the fragmentary utterances that occur legitimately in spontaneous speech. Breaks, on the other hand, would be better served by the traditional grammatical analysis approaches: if it isn't a complete sentence, throw it out. In the evolution of systems for spontaneous speech, the first change must be to make the system more flexible, and thus more capable of handling the great number of fragmentary structures in such speech; only after that has been accomplished can we then look for ways to restrict or augment this flexibility to recognize such less common phenomena as breaks.

be retained for processing.

The second most frequent use of "OK" occurs as an acknowledgment to a response to a request for information. That is, the first speaker asks for information; the second speaker gives him the information; and the first speaker acknowledges her response with "OK." These uses accounted for 20% of the occurrences of "OK" in the conversations examined. A typical example follows:

A: ...you can get to the International Conference Center by a few different ways you can either go by subway bus or taxi how would you prefer to travel <request for information>

C: [umm] let's take a bus <response>

A: **OK** <acknowledgment>

These occurrences of "OK" perform a necessary function; they inform the person making the response (the client in the example) that that response has been understood or accepted. If these utterances were filtered out, the conversation could not continue naturally. Thus, these utterances make a significant contribution to the dynamics of the conversation. As such, they need to be retained in processing.

The third use of "OK" is by far the most frequent, accounting for nearly 75% of the uses of "OK" in these conversations. This occurrence is difficult to characterize; it seems to be used to mark the beginning of a new move in the conversation, either a new piece of information or a new direction, often begun by a question.[4] The following examples give an idea of this use of "OK:"

A: **OK** let me look at the maps we have here to help you

C: [uhum]

A: **OK** I'm gunna take a look at the station where you're at and the area **OK** do you see the map on the screen

While this sort of subtle discourse information is probably used by human conversants to chart the flow of the conversation, it is too vaguely defined to contribute to natural language processing. It can be filtered out with no effect on the content or naturalness of the conversation. Notice how the example cited above, this time with the "OK's" removed, still sounds perfectly normal:

[4] As such, this use is most common in the speech of the agent, since she is responsible for introducing the most new information (i.e., giving directions). It may also, in fact, be a peculiarity of her speech that it occurs as often as it does, but without speech from another agent for comparison, we have no way of knowing.

A: Let me look at the maps we have here to help you

C: [uhum]

A: I'm gunna take a look at the station where you're at and the area do you see the map on the screen

Consider, however, what is required in order to carry out such a filtering. First of all, the utterance "OK" must be recognized as such. Then, its intonation must be examined to determine if it is a question. If it is, it is retained. If not, the discourse functions of the two previous utterances must be examined to determine if they make up a question/response pair. If so, the "OK" is also retained. If these requirements do not hold, the OK may be ignored in further processing.[5] Considerations concerning the coordination of filtering and analysis mechanisms are discussed in greater detail below (see also Figure 1 for a schematic of the interrelationships involved in the variety of processing considerations proposed here).

"And"

Although the use of "and" as a discourse marker is much less frequent than these uses of "OK," it can be handled in a similar way. Spontaneous speech is characterized by frequent use of the conjunction "and" to string together clauses in a single turn. "And" may be used as well to initiate a turn which is intended as a continuation of information from a previous turn (usually of the same speaker, but occasionally not; see Fais (1993a), Schiffrin (1987)). Although this use of "and" caries important discourse information, it makes little contribution to the semantic import of the utterance and may be filtered out in utterance-initial position. This includes those utterances in which "and" is not strictly utterance-initial, but becomes so after the filtering of "OK," false starts and filled pauses, as in the following example:

C: I've never been in Kyoto before

A: **OK and** <false start> **[ah] and** <discourse marker> you wanna get to the International Conference Center

This implies at least a partial ordering of the application of various filtering parameters (Again, see below and FIgure 1 for further considerations of the coordination of mechanisms). Of course, "and" also has syntactic functions; the implications of those for the syntactic analysis of these structures will be take up in the next section.

[5] One further problem occurs where "OK" is the only utterance in a turn and yet does not carry question intonation or function as an acknowledgment. Provision should be made for these cases to receive some default translation, such as "hai" for Japanese.

Structures to undergo grammatical analysis

Once the recognized string has been filtered for disfluencies and unnecessary "OK's" and "and's," standard grammatical analysis can proceed. But what are the remaining structures with which this analysis must work?

The full list of structures is given in Appendix A, but we will summarize each here and discuss those that seem to be peculiar to, and problematic for, spontaneous speech.

As noted in the discussion of "and," spontaneous speech is marked by an extended use of **sentence conjunction**. The use of "and" intra-utterance poses no difficulty for syntactic analysis. Traditionally, however, an utterance-initial "and" is a syntactic anomaly; the function of "and" is to connect two units and no such connection can be made if "and" does not occur between those units (unlike, say, the subordinate conjunctions "if" or "because"). Thus, utterance-initial "and" has no standard syntactic analysis. Above we motivated the filtering out of "and" on the grounds that it contributes information not utilizable by a language processing system, information that has little impact on the semantic content of the utterance. Here we see that for ease of syntactic processing as well, the elimination of utterance-initial "and" is desirable.[6]

The **subordinate conjunctions** used in the conversations examined were predominantly "so," "if," and "because." The structures they occurred in are amenable to standard syntactic analysis. The same is true for the limited instances of **imperatives**, although the placement of "please" can be problematical (see Fais, 1993b; Lepage and Fais, 1993) The **question** structures that appear are also all standard, though they may represent differing illocutionary force types and need to be more finely analyzed in any system that tracks illocutionary force. (Suggestions for a means for doing this were made in Fais and Kikui, 1991; see below for further discussion of illocutionary force.)

Declarative sentences on the whole are quite simple in structure, containing few **relative clauses**, and only occasional **prepositional phrases**. The majority of these sentences contain **copular verb** forms as well.

Short answers ("yes I do") and **confirmation tags** ("that's coins, right?") both appear infrequently; their analysis is unproblematical. An approach analogous to tag

[6] Although the use of "and" as a discourse marker with the function of "continu[ing] a speaker's action" (Schiffrin 1987) is usually discussed vis-a-vis its occurrence in sentence-initial positions, it clearly carries that function within an utterance as well. That is, the use of "and" can be seen overall as a way to mark continuation, continuation of the speaker's action of expressing a number of semantic units, and the occasions on which it fulfills this function utterance-initially can be seen simply as a special case of the overall function. In actual fact, most cases of the use of "and" for *sentence* conjunction, i.e., intra-sententially, contribute very little semantic content to the utterance either and could probably be filtered out along with the utterance-initial occurrences. However, because a syntactic analysis for the inter-sentential instances does exist, that option is not pursued here.

analysis might be employed to handle the single case of a postposed **belief clause** in a natural way: "...where the conference is, I believe."

What Yeager and Den (1993) call "**logophoric reflexives**" occurs once in the conversations examined ("I think the easiest way would be taxi myself"). They make some suggestions for how to incorporate an analysis of these pronouns into a grammatical system by using intonational information. The "**let me**" construction seems to be a feature of the agent's conversational style and occurs fairly frequently, along with one instance of the use of "let's" by the client. Syntactically, these are straightforward. What is interesting about them, however, is that their use in these conversations runs counter to their expected discourse use. These constructions have been interpreted as classic examples of syntactically coded speech acts: "let me" is usually unambiguously interpreted as an offer and "let's" as a suggestion. However, they do *not* always encode those speech acts in these conversations. In fact, "let me" functions as a suggestion in only one of the five cases where this construction occurs in these two conversations. The more frequent use of "let me" is equivalent to "I will:"

A: **let me** take a look at the maps we have here to help you

as is the use of "let's:"

A: how would you prefer to travel

C: **let's** take a bus

Interestingly enough, Seligman et al. (1993) found the same results for the analysis of "ne" in Japanese spontaneous conversation. The textbook explanation of "ne" is that it functions to seek confirmation; however, Seligman et al. found in their corpus that it was "almost always used for hesitation instead." These findings should ring a loud cautionary note for any analysis of illocutionary force type based solely on surface syntactic structure.

There are two syntactic phenomena which do not lend themselves to a straightforward analysis under standard assumptions. One is the utterance of **unattached noun phrases**, that is, NP's which do not play a functional part in a sentence structure. These were discussed extensively in Fais (1993c) as they also played a prominent role in the conversations examined in that work.[7] Clearly, a grammar

[7] The functions performed by the single NP's in the conversations under discussion here seem to be easier to circumscribe, however. They are limited to a statement of identification ("Good morning **Conference Office** can I help you"), or of affirmation:

C: [ah] do you know how much it should cost

A: yes the taxi [ah] is running about ten thousand yen right now

C: **ten thousand yen**

that requires the resolution of each structure to a sentential level is not adequate for the characterization of these utterances. The grammar must be flexible enough to recognize the validity of a free-standing noun phrase if it is not possible to incorporate the phrase into a sentential structure (although see footnote 3 for some reservations).

In the discussion above, we rather glibly glossed over the issue of the analysis of short answers. In fact, short answers of the sentential type are rather straightforward, requiring only a provision for the attachment of "yes" or "no" to what is otherwise a simple sentential structure, such as "yes I do." However, short answers may also consist of single noun phrases.[8] This case also requires a grammar that accepts a free-standing noun phrase.

To take it one step further, of course, short answers may be free-standing phrases of any type; the conversations examined here contain an example with prepositional phrases: "on the sides and the front of the bus." And free-standing clauses are not limited to the case of short answers; these conversations contained one example of a free-standing "if" clause, a construction quite common in the work done in Fais (1993c):

C: if you think I'll be able to communicate with the taxi driver

Thus, it is clear that the grammar cannot be restricted to the requirement that all utterances be headed by S, that is, that all utterances be analyzed as sentential units. However, how to limit the possibilities for acceptable uppermost level categories is unclear at this point.[9]

The last problematical case may admit of no clear solution at present. This is an example of what we would loosely call **postposing phenomena**. These kinds of examples are discussed at length in Fais (1993c) because they are, in fact, a fairly common phenomena there, though there is only one example in the conversations examined here.

A: that's where you can pick up a taxi **right there**

This is an example of cataphoric reference in which the pronoun "that" precedes its antecedent "right there," which in turn is a free-standing phrase. It *might* be possible to characterize the syntactic structure of this utterance accurately; it is a fairly standard

A: ten thousand yen

[8] of which another slightly problematical response, a single answer "yes" or "no," might be considered a special case.

[9] The option of expanding possible ultimate-level categories from just S to S, NP, PP and possibly others is only one possible solution. Another approach, one taken by Hirst and Ryan , is to allow incomplete analyses. In the first approach, stand-alone NP's would be analyzed as complete structures, headed by NP. In Hirst's approach, they would be analyzed as incomplete structures, S's having an NP but lacking a VP. Hirst's approach is also discussed in another context below.

example of what has been described as "right dislocation." However, there are other possible types of examples of this phenomenon which are more problematic (see Fais, 1993c for a detailed discussion). In addition, it will almost certainly pose problems for referent tracking; until the mechanisms for such tracking are better understood, we will leave the problem here.

Although syntactic errors are less frequent than is commonly thought in spontaneous speech, they do occur. The following two utterances are examples:

C: I Ø never been in Kyoto before

A: this is the bus station here and Ø the middle of the station you'll want to catch bus number five

In a real-time operating language processing system, it is clearly unacceptable simply to leave these utterances unprocessed. The system must either be flexible enough to make a partial attempt or have the option to query the speaker for a reformulation, or a combination of both. The design of this aspect of the system could benefit from experiments in which users must deal with faulty or no translations for their utterances (Seligman, personal communication). At present, these examples are merely noted and not resolved.

Likewise, a careful flexibility will be needed in the semantic interpretation of sentences, especially prepositions. A number of "odd" uses of prepositions were recorded:

A: I'm at Kyoto

C: I've just gotten into Kyoto at the Kyoto Station

While these are perfectly correct grammatically, the use of "at" with "Kyoto" in the first sentence and "at" with "gotten" in the second (in the sense meant here) are unnatural[10].

One possible approach to this problem is suggested by Hirst and Ryan (1992) for use in a different kind of natural language processing context. In their approach, there are two representations for any given text ("utterance" in a spoken language processing context): a natural language representation, i.e., the text itself, and the system's representation of the text, stored in parallel and linked at appropriate points. In cases where there is no determinant representation in the system's formalism, the natural language version is used. Hirst and Ryan are envisioning use of this type of "mixed-depth encodings" in the domain of large-text queries, but the concept is a useful one, on

[10] The use of "the" with "Kyoto Station" is another example of semantic "unnaturalness" and contributes to the "strangeness" of this utterance.

a more limited scale, in natural language processing. Certainly a real-time machine translation system, for example, cannot afford to leave large amounts of the analysis represented by the natural, source language. However, being able to manipulate non-understood pieces of utterance within the appropriate structural context is an essential aspect to an appropriate user-query function. So, for example, take the case in which the grammar of the system were such that it could not understand "I'm at Kyoto." If it had access to *some* representation of that utterance, it could use that information to generate an appropriate query such as "where are you?" It could generate this query by recognizing subject and verb from the structural analysis it *could* do, and generating an appropriate question word "where" to match at least the higher level semantic import of "at."

This type of capability might be useful in dealing with lexical disambiguation. There are a number of cases in which correct translation depends crucially upon the system's ability to disambiguate the sense of a lexical item from context; the correct interpretation of the following depends upon the use of the appropriate meaning for "take" vis-a-vis its use with "exit six:"

A: you wanna **take** exit six across the street to the bus station

That is, it cannot have its "pick up and carry with you" sense, but rather must have its "go through" sense. Whether a semantic analysis system can be designed that is robust enough to be able to discriminate between these two senses on the basis of contextual information alone is still an open question. Boitet and Loken-Kim (1993) have suggested that it may be more reasonable to expect that speakers will have to actively work with the system to resolve these sorts of ambiguities. If this is the case, then the system must have some means for isolating the difficulty and generating a query. If it were able to work with partial representations, then it might locate the two senses of "take," and formulate a query such as "do you mean 'carry the exit across the street' or 'go through exit six and across the street'" or some similar sort of question. The degree to which this capability will be necessary of course, depends crucially upon the proportion of sentences in which the system finds unanalyzable structures or undecidably ambiguous expressions.

Structures involving discourse considerations

There is a range of utterances which function not grammatically, but on a discourse level. We have already discussed "OK" and the discourse use of "and" above, but other examples include phatic utterances such as "good morning," "hello," and "you're welcome,." and responses and acknowledgments such as "great," yep," and "got it." These utterances form a rough continuum, from those having only a phatic function, such as "hello" or "yep," to those which may function either in a "literal" sense ("it's **great** weather outside") or a conventional sense ("I'll make a reservation for you;" "**great**"). The difficulty for machine analysis, of course, is discriminating between the two uses and thus, between the two possible meanings for the expressions. At present, it is not possible to automatically and unambiguously assign appropriate meanings to these expressions. However, it is possible to weight meaning assignments according to how frequently the expression appears in either its phatic or literal function. So, "hello" might receive an assignment of high phatic probability, while "good" would be ranked as neutral, and "bus" (which presumably would not have any phatic function), would receive a phatic probability of zero.

Another factor that can aid in discriminating between these two uses is the fact that these expressions tend to have their phatic meanings when they are singular expressions, and tend to have their "literal" meanings when embedded in another utterance. Thus, "good morning" uttered alone is almost certainly phatic, while in "It's a good morning," "good morning" has its literal sense.

In the conversations examined, there were ten utterances which could have phatic functions: "good morning," "thanks," "bye," "have a good time," "well," "then," "let's see," "all right," "got it," and "fine" (listed here in [intuitive] order of decreasing probability of carrying phatic meaning). All of these, with the exception of "fine" and "right," were used only as non-embedded utterances with phatic function. In no case were these eight found with phatic function in an embedded position in a sentence.

"Fine" occurred as a singular utterance functioning as a response to a suggestion:

A: Let me tell you how to get to a taxi stand right now[11]

C: OK **fine**

It also occurred embedded in a sentence. In this case it had its literal meaning:

A: if you need to make change you can just put in a thousand yen note ... and it will give you change back

C: OK I thank you very much I think I think that'll be **fine**

Similarly for "all right;" it appeared once, unembedded, with its phatic meaning, and a number of times embedded with three different literal meanings: as an intensifier ("right now"); as a confirmatory tag ("that's coins, right?"); and in a copular sentence ("that's right"). Thus, the generalization that an utterance expresses its literal meaning when it occurs in an embedded context, and expresses its phatic meaning when it is not embedded seems to hold without contradiction in these conversations. This distinction enables a clear-cut decision to be made between those two types of meaning for the fragmentary utterances with phatic functions in these conversations.

While it is certainly the case that utterances may have two types of meaning, a phatic

[11] Recall the discussion of "let's" above. This is the one case in which it is found in its typical role as the structure of a suggestion.

one and a "literal" one, it is not the case that they must be handled by two different mechanisms. Although great progress has been made in the area of discourse tracking, it is not yet possible to incorporate a sophisticated discourse analysis module into a working natural language processing system. Indeed, in some cases, it might not even be desirable to do so. That is, it is possible to take the analysis of so-called discourse phenomena a long way simply by exploiting their surface syntactic structures.

Fais and Kikui (1991) discuss in detail the particular factors involved in the generation of English responses appropriate to discourse context, focusing on a small number of responses and acknowledgments. The approach proposed there links a possible illocutionary force type with the surface syntactic structure of the utterance; rules determining the surface form of the response have access to the illocutionary force type features (there called "intention features") of the previous utterances and use those, along with other syntactic information, to generate an appropriate response. This approach could easily be extended to include the wider range of responses found in spontaneous speech simply by adding the proper entries to the lexicon.

Both this approach and the technique described above for discriminating phatic and literal uses of fragmentary utterances require minimal discourse level information, but they do presuppose a unique assignment of intention feature value to surface syntactic structure. While this is clearly something of an oversimplification, (as we saw in the case of "let me" and "let's" above) it may be an efficient procedure in a great number of cases. Shriberg et al. (1992) discuss an alternative method for distinguishing two uses of the words "no" and "well:" their literal (embedded) uses and their use as signals for repairs. They found these two uses to be "quite distinguishable ... on the basis of simple prosodic features." Clearly, there is further work to be done in this area.

Topic or focus tracking is another problematical area in which discourse information seems to play a part. In Fais (1993c), a troublesome number of what were called there "knowable omissions" were noted. These are cases in which some "knowable," focused element is elided from the utterance. Tracking focused elements through the course of a conversation might enable us to "re-construct" the full form of the utterance, with the focused element re-inserted. However, in the conversations examined here, only one such example was noted[12]:

C: the numbers are written on the bus?

A: yes the numbers are in English

C: **where on the bus?**

The tracking and inferencing required to interpret this sentence as "where are the

[12] In Fais (1993c), "knowable omissions" also included the result clause which is often omitted from "if" statements. The one example of this sort found in the conversations examined here was discussed above under "Structures to undergo grammatical analysis."

numbers (written) on the bus" are beyond the scope of present systems.

Coordination of the approaches

The schematic sketched in Figure 1 represents the logical dependencies among the approaches described above. All subsequent processes depend upon the process described above them in the figure. Thus, all processes depend upon the speech recognition phase to render the speech signal into identified words. Some of those words will be, in Kikui's term, "redundant," that is , they will be filled pauses and false starts. These will be discarded by a pattern matching filter along the lines of that proposed by Kikui (personal communication) or Shriberg et al. (1992).

In order for the discourse marker filter to correctly retain substantive uses of "OK" and "right," it is necessary that intonational information be available. Since one of the criteria for discarding the discourse marker use of "and" is that it appear utterance-initial, the ordering of this deletion process after that for filled pauses and repetitions (and even "OK") is crucial, as we saw above.

After filtering and intonational tagging has been done, the grammar can operate on the resulting input strings. The grammar will still have to be made somewhat more flexible than current versions in order to accommodate the various structures described above; in addition, it must be able to make use of intonational information in order to correctly assign Intention features to utterances. Intention features are assigned as part of the grammatical analysis on the basis of surface syntactic structure, intonational information, and, in some cases, the value of the intention feature of the preceding utterance.

It is impossible to correctly gauge the effectiveness of this type of system without implementation and testing. However, it *is* possible to guess that further discourse information and tracking might be of use in resolving residual difficulties. It is reasonable to think that a discourse manager could make use of the output of the grammar to achieve these aims.

Figure 1 is not meant as a representation of the actual organization of such a system, although, of course, if the processes are carried on in a strictly sequential manner, it would be. Instead, as explained above, it is a schematic of the logical relationships among these processes. This sort of system could easily be implemented in a "whiteboard architecture" such as that described by Seligman and Boitet (1993).

Future directions

Clearly the first step required once each of the areas described above is examined in detail, is the implementation of one or more of the processes discussed, in order to determine how effective they are. Only through implementation and testing can the questions raised here be confidently answered. Future collaborative work with Loken-Kim and others to put these suggestions into workable form should help provide some of those answers.

Specifically, one area crucial for the implementation of effective discourse analysis is the definition of the unit of analysis. As we have seen, the sentence can no longer be considered the unique unit of grammatical analysis for spontaneous speech since many well-formed utterances consist of structures smaller than or other than a sentence. We have begun work to identify basic syntactic units using as cues "and," "OK," and pause information in much the same way that pause information is used to segment Japanese spontaneous speech in Seligman et al. (1993), or auxiliary sequences are used to identify "stars" in Japanese by Tomokiyo et al. (1993).

English speakers seem to chunk their spontaneous speech using "and" and "OK" as boundary markers; however, where these cues are absent, other means, probably phonetic information involving pause and F_0, must be employed. While it has been assumed here that (fairly) conventional current grammars will be enabled by the filtering techniques described, it may well be the case that a grammar written for the "pause unit" will have significantly different characteristics from those of conventional grammars. This may necessitate adjustments to the filtering described here.

Conclusion

Because spontaneous speech contains a number of characteristics not found in more formal styles of speech, it is necessary to propose a greater variety of approaches for processing such speech in an automatic system. The conventional role of grammatical analysis must change; although in formal language processing, grammatical analysis could be the primary means for dealing with phrasal structures, in spontaneous language analysis, it must be augmented by filtering procedures which sort the structures to be made available to grammatical analysis. On the other hand, there are a number of discourse level phenomena found in spontaneous speech that may, in fact, yield to grammatical analysis. The proposals made here combine specific parameters for filtering non-contributing utterances with suggestions for new orientations in grammatical processing, which together constitute a configuration of approaches for processing the particular characteristics of spontaneous speech. Future work in the implementation of the proposals made here should provide insights as to the feasibility of these approaches.

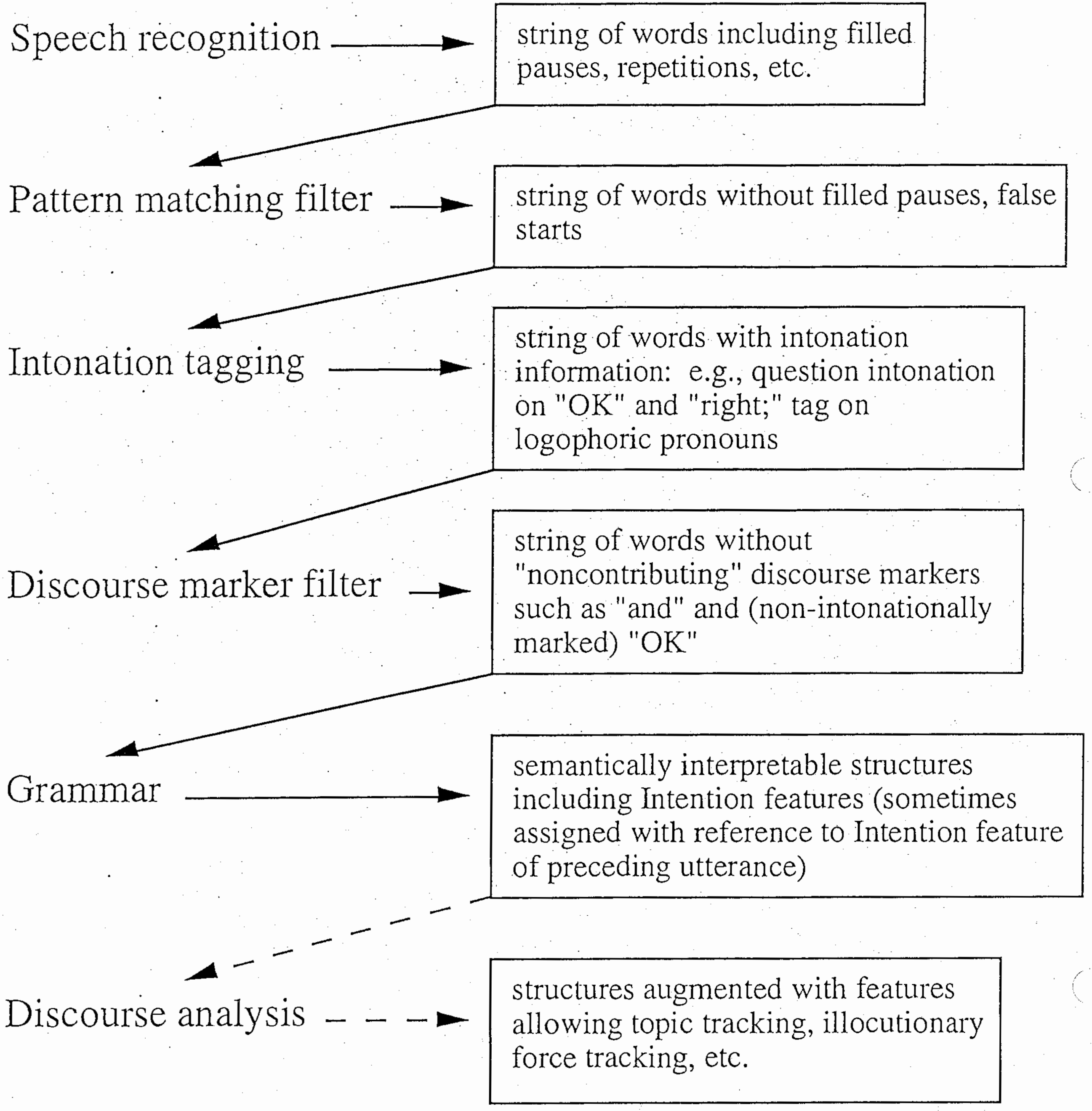
Speech recognition
string of words including filled pauses, repetitions, etc.
Pattern matching filter
string of words without filled pauses, false starts
Intonation tagging
string of words with intonation information: e.g., question intonation on "OK" and "right;" tag on logophoric pronouns
Discourse marker filter
string of words without "noncontributing" discourse markers such as "and" and (non-intonationally marked) "OK"
Grammar
semantically interpretable structures including Intention features (sometimes assigned with reference to Intention feature of preceding utterance)
Discourse analysis
structures augmented with features allowing topic tracking, illocutionary force tracking, etc.

Figure 1. Relative system dependencies

Appendix A: List of structures

The actual examples of the structures discussed in the text are listed here. Effort was made to list the structures in the same order (and under the same headings) as in the text. However, all discourse-related structures, though they are discussed under both "grammatical" and "discourse" headings in the text, are listed under "discourse considerations" here.

In the list below, notations such as "x2" indicate that the expression was found twice in the combined conversations. The use of "?" denotes a glottal stop. The word transcribed as "thi" is the full pronunciation of the word "the," with [i:] rather than schwa.

Notes on conventions: although effort was made during transcription to subjectively register small but discernible differences in the pronunciation of filled pauses (e.g., "um" vs. "umm"), I have not attempted to reproduce those differences here. Nor have I retained the various transcription conventions noted at the beginning of Appendix B. Please see those notes for additional questions concerning the transcriptions.

Structures to be filtered

Filled pauses

- ah x29
- eh x4
- hm
- mmhu*
- oh x2
- uh x2
- uhuh x3
- uhum x6
- um x9

False start

- a?
- and
- I
- I think
- I've never
- if you need
- is
- it's on
- Kyoto Center
- let me
- let me
- so that's
- th
- that's x2
- the bus
- there's a

False start, cont'd

- they go
- where I
- y
- you can travel
- you're l
- yu

Break

- if you go directly into the front doors of the bus station the tae (interrupted)
- I'm going to

Structures to undergo grammatical analysis

Conjunction

- you're going to be coming up this street here and you'll be going down Sanjo dori and coming up to the Conference Center about right there
- exit six will take you to the north side of the building and right across the street is the bus station
- I am on the first floor and that's about all I can tell
- I'm at Kyoto station now and I'm trying to figure out how to get to the conference
- let me go to a different map and I believe we can call up thi International Center and show you little bit about where the center is in relation to the area
- my name is Smith and I've just gotten i' to Kyoto at thi Kyoto Station and I'm trying to figure out how to get to thi International Conference Center where the conference is I believe
- that's right and I don't speak any Japanese
- the bus ticket will cost you five hundred yen and you can pay for it right on the bus
- there is change on the bus and if you need to make change you can just put in a thousand yen note and it will give you change back
- there's only one number five bus and it leaves every half hour
- you're attending a conference at thi International Conference Center and you're at Kyoto Station right now
- you're right in this terminal here and you can get to the International Conference Center by a few different ways

Subordinate conjunction

- if I take the number five bus it definitely goes there
- the taxi drivers do speak English so I don't think there will be a problem
- there is change on the bus and if you need to make change you can just put in a thousand yen note and it will give you change back
- they're very familiar with that Center because a lot of foreign people stay there and they visit there and go to conferences there so that shouldn't be a problem
- you're on the first floor I believe if you're using this phone

Imperative

- say please take me to the International Conference Center

Question

- and where am I now
- can I help you
- do they make change on the bus
- do you know
- do you know how much it should cost
- do you see the map on the screen
- does that Center appear on my map here
- how can I help you
- how would you prefer to travel
- is there anything else I can help you with
- what would you prefer to do
- where are you located in the station
- which would be easiest

Declarative sentence

- I assume it's OK if I speak English
- I can handle that
- I see x3
- I thank you very much
- I think that'll be fine
- I think thi conference is at thi International Conference Center
- I'm going to take a look at the station where you're at and the area
- I'muna show you where we're at
- in the bus station there's a taxi stand
- it's on the east side of the bus station
- my name is Smith
- that should be fine
- that's right x3
- that's what I am
- the numbers are in English
- the numbers are written on the bus
- the taxi is running about ten thousand yen right now
- there's no other number five bus
- there's a taxi stand in the bus station
- there's a taxi stand in the bus station
- thi International Center is right there
- thi International Conference Center's the first stop
- this is an English-speaking agent
- we have three different ways to get to the International Conference Center
- you can either go by subway bus or taxi

- you can take the subway bus or taxi
- you wanna get to the International Conference Center
- you wanna go out exit six
- you're at the station
- you're at Kyoto Station
- you're at Kyoto Station

Short answer

- no
- on the sides and the front of the bus
- yes I do

Tag

- that's coins right?

Belief clause

- where the conference is I believe

Reflexive

- I think thi easiest way would be taxi myself

Let me/us

- let me go to a different map and I believe...
- let me pull up my maps to help you here
- let me take a look at the maps we have here to help you
- let me tell you how to get to the taxi stand right now
- let's take a bus

Free-standing NP

- bus number five in the middle of the bus station
- Conference Office x2
- first floor
- five hundred yen
- how much money
- taxi stand on the east side of the bus station
- ten thousand yen x2

Free-standing PP

- in the bus station
- on the first floor
- on the sides and the front of the bus

Postposing

- all I have to say to the taxi driver is International Conference Center
- that's where you can pick up a taxi right there

Problematic sentences

- I never been in Kyoto before
- I've just gotten into Kyoto at the Kyoto Station
- if you look around you there should be exit six
- it says here on my flyer
- this is the bus station here and the middle of the station you'll want to catch bus number five
- you wanna take exit six across the street to the bus station
- you're at Kyoto

Structures involving discourse considerations

Phatic utterances

Fragment

- all right
- fine
- got it
- let's see
- then
- well

Idiom

- bye x2
- bye bye x2
- good morning x2
- have a good time x2
- thanks again
- thanks very much

Discourse marker

- and ah you wanna go out exit six
- and how much should I have ready
- and it's on ah it's on the east side of the bus station
- and that should take you to the International Conference Center
- and that will be the bus route
- and they're very familiar with that Center because ah a lot of foreign people stay there
- and where am I now
- and you wanna get to the International Conference Center

Knowable omission

- and how much should I have ready
- if you think I'll be able to communicate with the taxi driver
- where on the bus

Yes/no

- aw right
- no
- OK x37
- yea x5
- yep
- yes x7

Appendix B: Transcription of telephone conversation

Transcription conventions: In both this conversation and the one following in Appendix C, "A" designates the Agent; "C" designates the client, and the turns are numbered for ease of reference. The name of the client has been changed to protect anonymity.
Square brackets, [], surround filled pauses; parentheses, (), surround material that is judged to have been a false start (either a repetition, repair or fresh start, see text).
Slashes, //, surrounded descriptions of non-speech noises made by the speakers such as lip smacks (noted as "ls"), breaths, or laughter.
Curly braces, { }, and occasionally plus marks, ++, mark the boundaries of speech uttered simultaneously with speech uttered by the other speaker. Plus marks were used where a number of such instances occurred close together in order to minimize confusion as to which speech segments were simultaneous with which. The determination of the extent of the overlap of spontaneous speech was made by ear, and may not be entirely precise, although great efforts were made to be as accurate as possible.
Partial pronunciations or deviant pronunciations were marked with an apostrophe (e.g., "l' me" for "let me"); the fusions of "want to" and "going to" were transcribed as "wanna" and "gonna" respectively. Some other less common fusions were also transcribed as words. The full pronunciations of "the" and "a" were transcribed as "thi" and "e" respectively.

1. A: Good morning conference office how can I help you

2. C: (y') yes my name is [ah] Smith {[eh] (is)} I assume it's OK if I speak English

3. A: {[uhuhn]} yes [ah] this is an English speaking agent

4. C: {[oh]}

5. A: {that's} what I am {/laughing/}

6. C: {All right [um] I'm at [ah]} Kyoto station now and I'm trying to figure out how to get to the conference I think thi conference is at thi International Conference Center

7. A: O{K}

8. C: {it} says here on my flyer

9. A: OK [ah] you're at {Kyoto}

10. C: {[ah]} yes (I've never) [ah] I never been in Kyoto before

11. A: OK (and) [ah] and you wanna get to the International Conference Center

12. C: that's right {and I don't} speak any Japanese

13. A: {/breath/} OK let me pull up my maps to help you here

14. C: all right

15. A: OK /ls/ OK you're at (Kyoto Center)

16. C: that's {right}

17. A: {[eh] Kyoto} station and (you can travel) we have three different ways to get to the International Conference Center you can take the subway bus or taxi what would you prefer to do

18. C: [um] which would be easiest

19. A: I think thi easiest way would be taxi myself

20. C: OK [ah] if you think I'll be able to communicate with the taxi driver

21. A: [ah] the taxi drivers do speak English so I don't think there will be a problem (let me) [ah] let me tell you how to get to the taxi stand {right} now

22. C: {/creaky voice/} OK fine

23. A: OK [ah] you're at the station you're on the first floor I believe if you're using this phone

24. C: that's right first {floor}

25. A: {and} [ah] you wanna go out exit six exit six will [ah] take you to the north side of the building and right across the street is the bus station in the bus station there's a taxi stand if you [ah] go directly into the front doors of the bus {station} +the tae+

26. C: {OK} (+there's+ a) there's a taxi stand in the bus sta{tion}

27. A: {yes the}re's a taxi stand in the bus station

28. C: [mmhu]

29. A: and (it's on [ah]) it's on the east side of the bus station that's where you can pick up (a?) a taxi right there

30. C: got it

31. A: OK

32. C: yep

33. A: OK [ah] is there anything else I can help you with

34. C: [um] well let's see all I have to say to thi taxi driver is International Conference Center

35. A: yes say please take me to thi International Conference Cent{er}

36. C: {I} can handle that {/laughing/}

37. A: {O /laugh/ and they're very familiar with that Center because [ah] a lot of foreign people stay there (so that's) and (they go) they visit there and go to conferences there so that shouldn't be a problem

38. C: OK taxi stand on the east side of the bus station

39. A: [uhuh]

40. C: [ah] do you know how much it should cost

41. A: yes the taxi [ah] is running about ten thousand yen right now

42. C: ten thousand yen

43. A: ten thousand yen

44. C: [uhuh] OK [ah] that should be fine

45. A: OK

46. C: all right [ah] thanks very much then

47. A: OK have a good time

48. C: OK bye-bye

49. A: bye

Appendix C: Transcription of multi-media conversation

1. A: Good morning Conference Office can I help you

2. C: [eh] yes my name is Smith and [uh] I've just gotten (l) to Kyoto at thi [uh] Kyoto Station and I'm trying to figure out how to get to thi International Conference Center

3. A: O{K}

4. C: ({wh}ere I) where [um] the conference is I believe

5. A: OK you're attending a conference at thi International Conference Center and you're at Kyoto Station right now

6. C: that's right

7. A: OK let me take a look at the maps we have here to help you

8. C: [uhum]

9. A: OK I'm going to take a look at the station where you're at and the area OK do you see the map on the screen

10. C: yes I do

11. A: OK I'muna show you where we're at (you're l) you're at Kyoto Station you're {right in} this terminal +here+ and you can

12. C: {[uhum]} +I see+

13. A: get to the International Conference Center by a few different ways you can either go by subway bus or taxi how would you prefer to travel

14. C: [umm] let's take a bus

15. A: OK

16. C: /ls/

17. A: OK (I'm going to [ah]) wherer you located in the station do you know yu on the {first floor}

18. C: {[umm] I} am on the first floor and that's about all I can tell

19. A: OK if you look around you [ah] there should be exit six you wanna take exit six across the street

20. C: {[uhum]}

21. A: {to the bu}s station

22. C: I see

23. A: OK

24. C: yea

25. A: [ah] this is the bus station here and the middle of the station you'll want to catch bus number five

26. C: [uhum] bus number five in the middle of thi

27. A: in the

28. C: {bus station y}ea

29. A: {bus station} and that should take you to thi International Conference Center

30. C: OK [ah] does that Center appear on my map here

31. A: no (let me) [ah] let me go to a different map and I believe we can call up thi International Center and show you little bit about where the center is in relation to the area

32. C: [uhum]

33. A: OK thi International Center is right there

34. C: yea and where am I now

35. A: OK [eh] you're going to be coming up this street here

36. C: {[hm]}

37. A: {and} you'll be going down Sanjo dori

38. C: [uhuh]

39. A: an coming up to the Conference Center about right there

40. C: {I see}

41. A: {and that will be} the bus route (that's) thi International Conference Center's the first stop

42. C: If I take the number five bus [ah] (I) it definitely goes there there's no other number five bus

43. A: no there's only one number five bus and it leaves every half hour

44. C: awright [um] the numbers are written on the bus

45. A: yes the {numb}ers +are in Engl+ish

46. C: {Ith?} +where on the bus+ {yea}

47. A: {[ah]} on the sides an the front of the bus

48. C: OK [umm] and how much should I have ready

49. A: ({the bus})

50. C: {how much money}

51. A: the bus ticket will cost you five hundred yen and you can pay for it right on the bus

52. C: OK five hundred yen (that's) [um] that's coins right do they make change on the bus

53. A: [ah] yes there is [ah] change on the bus yea and (if you need) if you need to make change you can just put in a thousand yen note

54. C: [uhum]

55. A: and it will give you change back

56. C: OK I thank you very mu{ch (I think}) I think that'll be fine

57. A: {OK} great have a good time

58. C: OK thanks a{gain} bye bye

59. A: {bye}

Bibliography

Boitet, Christian and Kyung-Ho Loken-Kim, 1993. Human-machine-human interactions in interpreting telecommunications. Proc. International Symposium on Spoken Dialogue-93, 247-250.

Fais, Laurel, 1993a. Conversation as collaboration: Some syntactic evidence. Proc. International Symposium on Spoken Dialogue-93, 133-136.

Fais, Laurel, 1993b. An English analysis grammar in a unification-based framework. ATR Technical Report TR-I-0351. Kyoto, Japan: ATR Interpreting Telephony Research Laboratories.

Fais, Laurel, 1993c. Non-grammatical phenomena in real English conversation. ATR Technical Report TR-IT-0007. Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories.

Fais, Laurel, Kyung-ko Loken-Kim, Yoshihiro Kitagawa, forthcoming. Spontaneous speech in multi-media and speech-only environments. ATR Technical Report. Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories.

Fais, Laurel and Gen-ichiro Kikui, 1991. Determining surface form for indirect speech acts in English. ATR Technical Report TR-I-0235. Kyoto, Japan: ATR Interpreting Telephony Research Laboratories.

Hirst, Graeme, and Mark Ryan, 1992. Mixed-depth representations for natural language texts. In Paul S. Jacobs, ed., *Text-based intelligence systems*. Hillsdale, NJ: Lawtrence Erlbaum Associates, 59-82.

Lepage, Yves, and Laurel Fais, 1993. Syntactic trees for the sentences of ten dialogues. ATR Technical Report TR-IT-0036. Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories.

Loken-Kim, Kyung-ho, Fumihiro Yato, Kazuhiko Kurihara, Laurel Fais, and Ryo Furukawa, 1993a. EMMI - ATR environment for multi-modal interactions. ATR Technical Report TR-IT-0018. Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories.

Loken-Kim, Kyung-ho, Fumihiro Yato, Laurel Fais, Kazuhiko Kurihara, Ryo Furukawa, Yoshihiro Kitagawa, 1993b. Transcription of spontaneous speech collected using a multi-modal simulator--EMMI, in a direction-finding task (Japanese-Japanese; English-English). ATR Technical Report TR-IT-0029. Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories.

Nakatani, Christine, and Elizabeth Shriberg, 1993. Draft proposal for labelling disfluencies in TOBI.

Schiffrin, Deborah, 1987. *Discourse markers.* Cambridge: Cambridge University Press.

Seligman, Mark, and Christian Boitet, 1993. A "whiteboard" architecture for automatic speech translation. Proc. International Symposium on Spoken Dialogue-93, 243-246.

Seligman, Mark, Junko Hosaka, and Harald Singer, 1993. Pauses and hesitations in spontaneous Japanese dialogue. Draft.

Shriberg, Elizabeth, John Bear, and John Dowding, 1992. Automatic detection and correction of repairs in human-computer dialog. Proc. DARPA Speech and Natural Language Workshop-92.

Tomokiyo, Mutsuko, Laurel Fais, and K.H. Loken-Kim, 1993. Natural utterance analysis in Japanese-English machine translation based on pragmatics. ATR Technical Report. Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories.

Woszczyna, M., N. Aoki-Waibel, F.D. Buø, N. Coccaro, N. Horiguchi, T.Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Schultz, B. Suhm, M. Tomita, and A. Waibel, to appear. Janus 93: Towards spontaneous speech translation. Proceedings ICASSP-94.

Yeager, Brian A., and Yasuharu Den, 1993. Some Irregularities in English conversation. ATR Technical Report TR-IT-0008. Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories.