TR-IT-0007

Non-grammatical Phenomena in Real English Conversation

Laurel Fais

July 1993

This report discusses non-grammtical phenomena occurring in spontaneous English conversation. The source data was taken from a published corpus of English conversation and examined with respect to the ability of current machine translation grammars to analyze the constructions found. A number of structures problematical for such grammars are characterized in detail, with suggestions concerning their interface with MT systems.

Contents

1 Introduction

The ultimate of ultimate goals in the field of machine translation is the translation of spontaneous speech. But before that goal can even be considered, researchers must determine what formalism or approach is the most likely to yield successful results. In the process of making that determination, researchers have tended to construct small systems, limiting input to the system in a variety of ways: restricting the semantic scope of utterances to specific, goal-oriented tasks such as getting information about a conference or train schedules; editing out from speech input any ungrammatical or incomprehensible productions; constraining participants to using utterances selected from a set list.

However, once a reasonable level of success has been achieved with these limited systems, the goal of translating spontaneous conversation remains. In order to fulfill that goal, it will be necessary first to have a clear understanding of the nature of the phenomena of spontaneous conversation, and of those phenomena which will not readily yield to machine translation systems as they are currently constructed.[1]

This paper describes the results of the analysis of three natural conversations among educated adults in various settings. Some fully implemented, comprehensive machine translation system incorporating syntactic, semantic and lexical components was assumed. In particular, the syntactic component was assumed to be similar to most syntactic systems now in use in that it was considered to operate by matching the discovered structure of the input string to some

[1] Strictly speaking, what is required is an understanding, not of spontaneous conversation between human beings, but of "natural" conversation between a human being and whatever machine translation structure is in place. Clearly, people tend to use clearer and more grammatical speech when in the presence of another person whose command of the language is imperfect; I strongly suspect that these accommodation techniques would also be applied in a case where a speaker knew his/her speech was to be translated by a machine. (And, in fact, Svartvik and Quirk (1980) noted adjustments in speech depending upon the presence or absence of a microphone or the use of the telephone in their elicited material.) However, since this information is not yet available, the phenomena reported on here can be taken to be a "worst case scenario" for ungrammatical utterances in human/machine translation system conversations.

member of a pre-defined set of "acceptable" or "grammatical" language patterns.[2] The conversation analyses were made in order to determine what structures are present in natural conversation that could not be handled by such a machine translation system. It is these structures which pose the next hurdle to achieving the goal of machine translation of spontaneous conversation.

2 Data

Three conversations were chosen from Svartvik and Quirk (1980). The elicitation techniques used to gather this corpus are described in detail in Greenbaum and Quirk (1970). The source corpus consists of 34 conversations each comprised of 5,000 running words, transcribed from recordings with notations for simultaneous speech, incomprehensible utterances, tone units, nuclear tone, relative pitch levels, stress, pauses, and phonetic transcription of "deviant" pronunciations. In the examples cited below, pauses are indicated by additional spaces between words, or by "-", and all other prosodic information has been omitted. The indications of simultaneous speech, (the use of "*'s" or "**'s" to bracket two phrases uttered at the same time) however, are sometimes relevant to the description of the phenomena and thus are retained in many cases.

The designations used below for the conversations selected from the source corpus are those of Svartvik and Quirk (1980). These conversations were selected to represent a variety of participants and intimacy level. Speakers designated by capital letters were unaware of being taperecorded; those designated in lowercase letters were "non-surreptitious speakers," were aware of the recording, and were responsible for keeping the conversation going. The participants in the conversations discussed in this paper are as follows (description quoted from Svartvik and Quirk 1980, pp.27 and 29):

[2] This can be interpreted generally enough to include both pattern-matching and rule-based approaches to syntactic analysis. An example- or knowledge-based system might have greater success with "peripheral" cponstructions, depending critically, of course, upon the scope of the knowledge base.
In a few cases, the ability of a machine translation system to handle the phenomena described will be significantly affected by the presence or absence of a dialogue manager or discourse plan component. It will be clear from the discussion below when those cases arise.

Conversation	Speaker identity symbol	Speaker identity
S.1.1	A	male academic, age c. 44
(1211 tone units)	B	male academic, age c. 60
S.1.5	A	female secretary, age c. 21
(1310 tone units)	B	female academic, age c. 25
	C	female secretary, age c. 35
	D	female secretary, age c. 21
S.2.10	A	male merchant banker, age c. 30
(1462 tone units)	B	female housewife, age c. 30 (A's wife)
	c	male computer specialist, age c. 30
	C	same speaker as c, but in the last part of the text he is no longer aware of being recorded and is called C...
	d	female research worker, age c. 25

From the descriptions of the participants, and from the analysis of the conversations, some general conclusions can be drawn as to the social tenor of these conversations. S.1.1 is a somewhat academic and academic-political conversation between two colleagues who seem to know each other fairly well, but who may not actually be friends socially. In S.1.5, B, C, and D are also colleagues in an academic setting, though at different social level, and they are discussing with a new person, A, the way their work environment is structured, giving advice, etc. The participants in S.2.10 are much more familiar with one another, clearly friends of some standing, and are conversing about day-to-day subjects over dinner.

3 Analysis

Because no actual machine translation system such as the one assumed here was available to use for analysis, it was necessary to cull "ungrammatical" phenomena from the chosen conversations by hand. To adopt Schiffrin's (1987) adaptation of Abraham Kaplan's phrase, the "logic in use" that directed the selection of these phenomena was as follows: any structure, utterance or piece of utterance that was felt could not be analyzed by the assumed machine translation system was listed (in almost all cases selected, the choice is uncontroversial and only loosely dependent upon the sophistication of the grammar assumed). As the examination of the first conversation progressed, those phenomena which were repeated were tentatively categorized together, while new phenomena were added to the list. This process of adding and grouping continued through the first conversation. Occasionally, differences that had seemed minor (e.g., between "yes" and "yeah") were reassessed to be important, and necessary readjustments to categories were made. The grammatical structures of the second and third conversations were analyzed keeping in mind the categories found in the first, but with the recognition that different conversational contexts may, and in fact did, give rise to different phenomena, to different weights of frequency and importance for similar phenomena, and to different configurations of groupings of phenomena.

Because of the varied nature of the conversations selected for analysis, it is possible to examine the relative frequencies and functions of particular phenomena across conversations with respect to intimacy level, age and gender differences. Though some suggestions in that area are made here, the bulk of that work is still being undertaken. The goal of the present paper is simply to explanatorily characterize the phenomena that were discovered.

There are two major ramifications of the method of analysis utilized in this paper. The first and most obvious is that, since the analysis was done by a human being, the very real chance of human error is introduced. The most likely type of error is omission. It may certainly be the case that some individual examples were overlooked; however, the possibility that an entire category of phenomena could have been omitted is highly unlikely, given the amount and variety of data examined.[3]

The second consideration is, in some ways, the opposite issue. That is, there are a number of structures that are, strictly speaking, grammatical, and yet that are typical of spoken rather than written forms of expression. Because these structures are of particular concern to grammars written to handle spontaneous conversation, despite the fact that they are not strictly "ungrammatical," they are noted and described below.

While the application of the criteria described above seems rather straightforward, it in fact required a liberal use of linguistic and conversational intuition. Certainly there were a number of cases in which absolute categorization could not be made, either because the exact nature of

[3] Another area for error is misunderstanding or misinterpretation of the function or importance of certain phenomena. It was to guard against this error that three quite different conversations were selected rather than, say, one or even two. It was felt that the greater variety and sheer weight of the data would prevent this type of error.

the structure could not be understood, or because the example seemed to lend itself to more than one category. The former problem is a by-product of working with transcribed conversations, though it may be a hazard even when the researcher is herself involved in the conversation. The second concern speaks simply to the nature of conversation and conversational structures: they may not necessarily be uni-functional, and in fact, given the requirements of discourse coherence (Schiffrin 1987, p.315-6), we would not expect them to be.

Although this was not a major aim of this analysis, applying the criterion "handleable by a comprehensive machine translation system" to the data allows us to make an estimation of the number of utterances occurring in spontaneous conversation which are in fact grammatical. To make a statistically valid count would require a far more rigorous definition of "utterance" than has yet been proposed, but an idea of the situation can be gotten from the examination of a typical piece of conversation such as the one included as Appendix A. All structures that could not be handled by the English analysis system developed in the Interpreting Telecommunications Department of ATR (Fais 1993) are in bold type. A perusal of that example gives an indication of the relative numbers of grammatical and ungrammatical utterances in these conversations.

The relative amount of grammatical phenomena can be compared to estimates by Labov (1972): "the great majority of utterances--about 75 percent--are well-formed sentences by any criterion. When rules of ellipsis are applied, and certain universal editing rules to take care of stammering and false starts, the proportional of truly ungrammatical and ill-formed sentences falls to less than two percent." It is with that 25 percent, and with those "rules of ellipsis" and "universal editing rules to take care of stammering and false starts," that we must be concerned.

4 Description of phenomena

4.1 Syntactic violations

Though they are not common, clear syntactic violations do occur in spontaneous conversation. The following are a selection of examples:

555-7 B: but you'll be amazed actually if you go to some of these seminars **Ø** the things that people say (S.1.5)[4]

803-4 C: there was a rather peculiar situation **that** they advertised for a secretary (S.1.5)

735-6 C: Sidney Heath sort of lives upstairs but **he's** really seems to work more with Hart (S.1.5)

1015-1018 A: I must say if one wants to be have a success a successful job[5] and to be successful in whatever field one enters **and** I'm absolutely convinced now that it's important to know at the age of eighteen (S.1.5)

346-7 B: but he **was** a great **thing about** structuralism though **isn't** he (S.2.10)

310 c: I'd quite like a modern sort of single lens reflex **stuff** (S.2.10)

There is little systematic that can be said about these examples; in the three conversations examined, only about ten such errors were found. The only pattern that repeated involved complementizers such as those in 555-7 and 803-4. This sort of phenomenon, straightforward mistakes in usage, will not be a significant factor in MT efforts to handle real conversation since they occur so infrequently.

4.2 Structural ambiguity

Another prominent concern in discussions about adapting MT systems to handle real speech is ambiguity. In the course of these three conversations, there were no cases where an ambiguous utterance caused any clear communication difficulties. However, what MT systems and what human beings regard as ambiguities are often very different; humans clearly have an extensive range of strategies available for interpreting what would be ambiguous in an MT system. In the absence of access to a system with which to detect such ambiguous utterances,

[4] The numbers that appear at the head of each example are the designations for the tone units in which the example occurs in the conversation. The numbers at the end of each example are the numbers of the conversation from which the example was taken. Aspects of the example which are under discussion are in boldface.

[5] This sort of correction is considered below.

the extent of MT ambiguity in these conversations remains unclear.

4.3 The starts and stops of conversation

Of more obvious concern, however, are the numerous occasions on which speakers break off, interrupt, correct, or repeat utterances. In each case described below, an effort is made to indicate any consistent patterning that might afford an MT system the leverage with which to deal with these phenomena.

4.3.1 Breaks

There are two general cases in which speakers omit material from their utterances.[6] In one case, either the semantic properties or the lexical item or both are recoverable in some way; these are called "knowable omissions" and are discussed below. The other case involves what are called "breaks": instances in which material is deleted that cannot be recovered. In effect, the speaker has broken off one structure/idea without bringing it to completion.

These are generated in two ways in spontaneous conversation: first, the speaker gets interrupted and doesn't finish, or, second, the speaker in essence interrupts himself, either muddying the syntactic waters and losing track of the initial thought, or simply abandoning the utterance in midstream.

The former type, turn-boundary breaks[7], are not an issue for an MT system. Presumably, in the context in which an MT system would be utilized, turn-taking structure will be at the very least quite marked, and possibly even rigidly pre-determined. Thus, the opportunity for interruption simply will not arise given the structure of the interaction. However, the speaker-internal break could still occur, and poses in fact a much more difficult problem for identification than the simple case of turn-boundary breaks.

[6] Clearly, from the point of view of the conversations, these are not omissions at all; they become "omissions" only when analysed from the point of view of an MT grammar.

[7] I will not use the more common term "interruption" for these cases because they also include examples such as the following in which the break occurs at the end of the speaker's term, but is not occasioned by another speaker pre-empting the turn ("-" indicates pause):

528-9 A: This is very tricky - I should have thought there were -
530-2 B: yes well quite they do that sort of thing you see (S.1.5)

There is a small number of breaks of this kind scattered through the three conversations.

The following example contains at least two speaker-internal breaks, one after "how many people have you got for the," and the other before "we haven't seen each other...":

756-8 A: how many people have you got for the Ø you know if you incidentally Ø we haven't seen each other since that peculiar meeting with the language lecturers, remember? (S.1.1)

Possibly, an MT system could be set up that would simply discard ungrammatical pieces of utterances such as the entire string before "we haven't seen each other..." However, as we will see below, similar "fragments" may play a part in the semantic import of utterances, though they may have only a loose if any connection to the syntactic structures of the utterance.

There was a total of nearly 100 breaks identified in the three conversations. In both S.1.1 and S.2.10, there was an equal number of speaker-internal and turn-boundary breaks, with the latter being far fewer in S.1.5, a fact which may have been affected by the presence of a new person in S.1.5. Conversations S.1.1 and S.1.5 each have half the total number of breaks found in the more casual conversation S.2.10, which indicates that breaks are less likely to be found in more formal situations, which a machine translation context would be expected to be.

4.3.2 Knowable omissions

The second case, alluded to above, in which a speaker deletes material from an utterance is the situation in which that material can in some way be recovered, either by syntactic reconstruction or by inference from semantic context or both. In terms of MT system analyzability, knowable omissions fall into two categories. In the first, syntactic material necessary to the well-formedness of the utterance is deleted. In the second, adjunctival material is omitted.

The first case is similar to what has been termed a break. However, in the case of a knowable omission, the missing material can be extrapolated either syntactically or semantically. In the casual and friendly conversation of S.2.10, the vast majority of omissions consist of the subject and frequently the copula or another verb form. The omissions are most often in the first person, but may be third person referential or non-referential and sometimes second person as well:

26 A: I mean I wouldn't like to live in Paris Ø think Paris is a fabulous city but - Ø

wouldn't be easy to live there. [omissions of first person and non-referential third person]

235 B: I think my desk actually is a little too much for a present, isn't it
238 A: Ø wouldn't get up the stairs [referential third person]

367 A: you don't read enough darling Ø just sit and drink [second person]

643 B: Murder on the Orient Express is now on the ABC Shaftesbury Avenue with the Godfather Ø much the same thing really [subject and copula]
647 c: the old Godfather Ø first Godfather [article]
648 A: Ø be a pretty good double bill that actually [subject and modal "would"] (S.2.10)

Example 647c illustrates another, somewhat common omission: that of the article. All of these elements, i.e., subjects, verbs, and article, are syntactically necessary to the well-formedness of the sentence. In the case of a break, where material is omitted that is crucial to the sentence structure, the partial sentence can simply be ignored since it contributes little to the conversation. However, it is clear that these examples cannot be ignored; despite the fact that they are syntactically incomplete, they are an integral part of the conversation. Instead, in real conversation, the syntactic and semantic components operate in conjunction to supply the missing material: the syntactic component by supplying the category information (say, "article"), and the semantic by supplying (at least) the relevant semantic information (say, "definite").

There are two major constructions within which crucial syntactic material can be supplied from context: short answers and lists. Short answers have long been analyzed as sentences which omit material repeated across utterances, much as comparatives have been analyzed intrasententially.[8] Broadly speaking, verification queries could be analyzed analogously, supplying the missing material from the statement previous to the query. List type structures also tend to omit repetitious material:

[8] Comparatives such as "I am as tall as he is " have been analyzed as omitting the adjective "tall" from the second clause. However, in conversation, the amount of deletion may extend even further:

811 A: I don't find myself getting as irritated Ø. I'm more amused Ø you know

1326 C: can you while it's in your mouth bite the cherry off and then spit out the stone still connected to the stalk
...1332 A: oh of course you can't Ø [short answer] (S.2.10)

585 B: heard his name mentioned ... by Darlington while I was down there
591 A: did you Ø [verification query] (S.1.1)

(in a discussion of the linguistic status of the word "worth")
684 A: you can either say that it's like "like" and some of the other pseudo-prepositions "it's not like me" "it's not like his wife" "it's not worth the trouble" "Ø not worth fivepence" [list structure] (S.1.1)

The difficulty for MT systems, of course, is in recognizing these conversational moves. Short answers may be identified on the basis of the previous question, and verification queries on the basis of the previous statement, but list structures will be more difficult to identify[9].

Sometimes knowable omissions border on the idiomatic:

606 B: "it's not worth the trouble." - how do you analyze "worth" ...
616 A: Ø good question (S.1.1)

310 c: I got a letter from her this morning which I haven't read yet
312 B: oh golly wouldn't you like to read it
314 c: Ø good idea (S.2.10)

These might lend themselves to an idiomatic analysis (with the problems that these sorts of moderately productive idioms entail: what about "good point," good thinking," "good answer"...?). But another "idiomatic" omission structure is much too complex to be handled via the lexicon, namely, the omission of the alternate clause to an initial "if" clause, a quite

[9] Short answers are usually identified on the basis of their position as a second speaker's reply to a question uttered by a first speaker. However, allowance will have to be made for cases in which the same speaker both asks and answers a question. A small number of cases like this appear in these conversations:

1385 B: These aren't English cherries no they can't be Ø can they (S.2.10) common phenomenon in these conversations:

257 B: I don't know if he dropped that - - Ø (S.1.5)

667 B: it's interesting but it's purely academic and if you're happy with that Ø ... (S.1.5)

These structures, as well, are integral parts of the conversation. Their alternative clauses can be understood by the hearer and thus need to be included in an MT analysis. The mechanism for accomplishing this, however, is unclear.

Another problematical area involving knowable omissions includes cases in which the syntactic structure of the existing material is well-formed, but gives rise to an incorrect interpretation because it does not take into account material that was, in fact, deleted. Consider the following example:

368 A: but there are lots of precedents for moving round in posts in Brighton aren't there - -
look at - Zimmerman for example - (Ø) from - Turnwick - to Lord Warden - to PP
376 B: yes well Turnwick to Lord Warden wasn't a very obvious one was it (S.1.1)

A's utterance beginning with "look at" could be syntactically analyzed as analogous to "look at Zimmerman from top to middle to toe," that is, an interpretation in which the three prepositional phrases following the object modify the verb "look at." However, the correct interpretation would be two clauses with the subject and verb omitted from the second clause: "Look at Zimmerman. [He moved] from Turnwick to Lord Warden to PP." In fact, only in that way can there be an appropriate semantic assignment for the antecedent of "one" in B's utterance, namely "move," (although this is the type of referential construction reviled by high school English teachers inasmuch as the antecedent to "one" is the verb "move" rather than the noun "move").

There are a number of cases, however, where the human analyst can identify the general import of missing material, but in which the connection between that material and the syntactic and even semantic contexts is so tenuous as to be undefinable in a way rigorous enough for an MT system to utilize. These may occur whether or not there is any syntactic ill-formedness (the first example is edited for brevity):

(discussing setting questions for a qualifying exam at university)
15 B: what you do is to make sure that ... there's something that your own candidate can handle
21 A: ah you mean that the papers are more or less set ad hominem are they
25 B: they shouldn't be but I mean one sets one question (Ø) now I mean this fellow's doing... [syntactically well-formed] (S.1.1)

It seems clear that B's utterance could have been completed: "one sets one question that one's own candidate can answer," but how to formulate that in specific enough parameters for an MT system to interpret appropriately is at this stage impossible to know.

Similarly, this sort of vague connection can be accompanied by syntactic ill-formedness, though with similar difficulties surrounding the possibility of formulating a procedure for identifying the omitted import (also edited for brevity):

469 B: it's what they call ILA tests which stands for investigating language acceptability and they've done these on groups of undergraduates...and asked them Ø - there are various types of tests they give them (S.1.5)

A listener could easily interpret that the missing object of "asked" should be "questions," but how that can be accomplished in an MT grammar is still unclear.

There is no sharp line of demarcation between examples in which omitted material is recognizable or knowable and examples which should be analyzed as a break, implying that the missing material is irretrievable. Instead, there is a spectrum of syntactic and semantic vagueness ranging from cases where it is possible to supply missing lexical items, to those in which semantic imports can be supplied, to those which require human-like powers of inference to detect a semantic connection or implication, to cases where such inference is impossible.

Knowable omissions were quite common in all three conversations. The less casual conversations, S.1.1 and S.1.5, had 15 and 24 respectively, exhibiting a wide range of types, especially the omission of "if" clause alternatives, the omission of material repeated in list structures, and constructions similarly retrievable from context. By contrast, however, the casual conversation S.2.10 had more than 80 knowable omissions, almost all of them subject, subject and copula, or subject and verb omissions.

4.3.3 Interjections

Occasionally a speaker breaks his utterance in order to change direction, but then returns to the breaking point and resumes the thought. These examples are typically known as interjections. Sometimes they consist of a well-formed syntactic structure and are surrounded by what would be a well-formed structure in the absence of the interjection:

1326-30 C: Then can you **while it's in your mouth** bite the cherry off and then spit out the stone still connected to the stalk (S.2.10)

However, where the interjection is an independent clause or is somewhat complex, there is often repetition of some of the initial phrase when the speaker returns to the first construction:

13-16 B: well what you do is to - - **this is sort of between the two of us** - - what you do is to make sure that your own candidate... (S.1.1)

The interjection may be a phrase as in the first two examples below, or, commonly, a vocative as in the last two:

213-4 A: Gordon, where **simple question** where are you going to put it all (S.2.10)

1411-14 A: when I was little we had a garden and we had **in Persia** and we had about fourteen cherry trees (S.2.10)

90-1 B: I wouldn't want it before the end of June anyhow **Reynard** because I'm going to Madrid (S.1.1)

427-430 B: as I've tried to tell you **darling** all these times in fact it's very succinct isn't it (S.2.10)

Notice that what defines these as interjections is that they appear in positions within the string that are not well-formed clause or adjunct phrase adjunction sites[10]. Thus, despite the fact

[10] With the possible exception of the vocatives. However, vocative phenomena in general are not incorporated into current syntactic systems and thus their use is still problematical for such a system, although it might be possible to characterize their adjunction sites as clause boundaries at some specified level.

that they themselves are well-formed clauses or phrases, and their matrix contexts are also well-formed, their presence within that context cannot be interpreted correctly by current MT systems.

Interjections as such are infrequent in conversation. If the speaker loses track of the initial syntactic thread, "interjections" will in fact end up simply as breaks, since the original structure will not be completed, as in the following typical (though long) example:

181-96 B: Now if these papers come by the twenty-ninth of June and you send them through to me at Loughton then between the twenty-ninth and uh let me see we're having this meeting of CSC assistants on the fourth of July which is a Saturday. I'll have about half a day's work to look at some odd scripts before then and then I shan't get any scripts from the assistants before about let me see four five six seven about the eighth so I shall have roughly from the twenty-ninth of June to the eight of July [which I can spend on those papers]

Notice that the structure "between X and Y," begun in "between the twenty-ninth and..." is never completed. The long interjection introduced by "let me see," although it could be completed and the original structure resumed, masks the original structure so that it is never completed and must be considered, instead, a break. (However, notice that a similar structure is picked up at the end, where it *is* completed: "from the twenty-ninth of June to the eight of July.")

Another factor which reduces the number of interjections which are identified here is that standard phrases such as "you know" or "I mean" which might also be interpreted as interjections are considered in another category below. They pose similar problems in terms of adjunction site.

The distribution of interjections in these three conversations is similar to that for breaks discussed above. S.1.5 has only one interjection, while S.2.10 has the greatest number, ten, and S.1.1, four. This is in direct proportion to the amount of intimacy among the participants in the conversations. Vocatives and clauses or phrases are more frequent than independent clauses, as might be expected from the discussion above.

4.3.4 Corrections

Closely related to breaks and interjections are corrections, which involve switching from one syntactic direction to another. These shifts may range from a fairly minor adjustment to complete abandonment of a structure or direction:

1015 A: I must say, if one wants to **be** have a **success** a successful job one must... (S.1.5)

573 A; it obviously is a matter of seeing whether **one gets** - - one's sufficiently interested in a thing (S.1.5)

Corrections differ from breaks in that they maintain the same semantic direction across the syntactic shift; they differ from interjections in that they do not return to the original construction. They are similar to knowable omissions; for example, in 1015 above, it seems clear from context that the gap after "if one wants to be" could be filled by "a success." Thus, in a sense, it is a knowable omission. Yet because the utterance continues the same semantic content in, in fact, two other different syntactic guises, resulting finally in a syntactically well-formed and complete utterance, the shifts in this case are considered to be corrections.

Corrections pose the same sort of problem for MT systems that breaks pose: they contain an ungrammatical, and thus unanalyzable utterance that must be ignored in the analysis process. Occasionally, of course, corrections will be made in a completely conscious and, thus, completely well-formed way by speakers:

445 c: have I have I got *it around*
446 d: *you must* have it. I gave you **a copy**
447 c: **I know you** **I know I said have I got *my* copy around. I didn't say have I got *it*.** (S.2.10)

Occasionally, a word or phrase will cue a correction, as "I think" does in the first example below, and "no" does in the second:

430 B: she's done an MA and is now on I think I'm not sure if she's doing a PhD. (S.1.5)

768 B: how on earth are we going to get into that tomorrow night - - no - friday night (S.2.10)

But even for these cues to be useful to an MT system, their function as a signal for a correction must be disambiguated from their other functions in marking discourse structures in a conversation (see below for more on "I think" especially).

The distinction between correction and break is a subjective one, the crucial criterion for which is the perception of a semantic link between the material on both sides of the omitted material, with all the possible questions that judgment entails: how much of a link is a link? how far can you go to find such a link? In the first example below, the only link between "he's a hell of a..." and "he's a big head" seems to be a general feeling of deprecation, but there is certainly no specific sense in which the first phrase has a particular semantic link to the second. In the second example, the clue that the missing word after "I might get um terribly..." might be "busy" occurs four tone units away in the form of a very vague inference from "we may need you to do some work...":

1041 B: I mean he's a nice fellow normally but he's a hell of a - he's a big head in some ways (S.1.1)

221 A: he said oh well you know I might get um terribly - - you know I'm - I'm just hanging on now and could take you on permanently - we may need you to do some work in the evening (S.1.5)

Regardless of the exact classification, a rough estimate can still be made of the frequency of corrections in spontaneous conversation. The most academic conversation, S.1.1, contained approximately 18 corrections; the more casual S.1.5, 23; and the most casual, S.2.10, 36.

4.3.5 Repetitions

Often accompanying corrections (and interjections) are repetitions. In making a correction, a speaker may repeat material around the portion of the phrase he/she is restructuring; in constructing an interjection, the speaker often repeats material from the original construction in order to establish the pattern he/she is resuming.

Repetitions may serve other functions as well. They may be made for effect, emphasizing an utterance contribution, either speaker-internally or across speakers[11]:

[11] See Tannen (1989) for a full discussion of the use of repetition in conversational discourse.

150 B: oh well, that's **very good** - **very good** (S.2.10) [speaker-internal]

223 B: I'll give you **a couple of hundred tiles** maybe. I have got a particularly re*volting coffee service*
225 c: *what do you mean **a couple of hundred tiles*** why do you have **a couple of hundred tiles** (S.2.10) [across speakers]

Repetitions by a second speaker may be used to confirm his/her understanding of the statement of a first speaker (this occurred in all three conversations):

9 B: find out the right seminars to go to that's what I did when I first came
11 A: **the right seminars** yes (S.1.5)

They may also function in the negotiation of turns, usually because there has been an overlap between the first and second speaker and the speaker who takes over the next turn must repeat what she/he has said to ensure that it was heard. But less frequently, a second speaker will simply repeat a phrase from the first speaker and then continue, retaining his/her turn as in this first example (see also cooperative structures below):

64 B: he has a way of having [seminars] at a horrible time
63 A: yea
66 C: like five-fifteen
68 B: **like five fifteen** when you want to go home [B regains turn] (S.1.5)

197 c: so we won't see them until September no
198 B: *so you*
199 c: *and they wouldn't* **they wouldn't** come earlier you see [c retains turn] (S.2.10)

564 c: it's written for people who've got a smattering of knowledge anyway of music *<2-3 syllables>*
566 B: *oh I thought* **I thought** it was written for people who have a sense of humor as well [B takes over turn] (S.2.10)

Many repetitions, however, are simply that: the repeated utterance of a particular word or phrase:

121 A: ...the other man, Chomley, ought **ought ought** also to have got his in on time (S.1.1)

Simple repetitions, as in this last example, and turn-internal repetitions as in 199c and 566B above, can safely be ignored in some way by an MT grammar. However, repetitions made to signify understanding, or for rhetorical emphasis make a substantial contribution to the import of the sentence and should be taken into account in a translation. It is impossible to conceive at this time how an MT system might go about the task of differentiating between the two classes of repetitions.

Repetitions occur in all three conversations, though to a greater extent in the most casual conversation, S.2.10. The other two conversations had approximately 20 repetitions each, primarily simple repetitions, but also a small number for confirmation or emphasis. Conversation S.2.10, on the other hand, had approximately 80 simple repetitions, with another ten or so each for effect and in turn-taking negotiations.

4.3.6 Discussion

It is clear that there is a great deal of overlap and vagueness in the delineation of the five phenomena discussed in this section (breaks, knowable omissions, interjections, corrections, and repetitions). Regardless of how they are defined, however, it is clear that the processes that they represent pose a problem for MT systems. The structures preceding breaks, corrections and repetitions are structures that are in essence replaced by what follows and so should be ignored by the MT system. However, the situation with interjections and knowable omissions is slightly different. The structure preceding an interjection, though the interface between that structure and the interjection may look like a break, must be retained for unification with the structures following the interjection. Clearly, systems which abandon and discard analysis of ill-formed structures will not retain the partial analysis of the initial structure needed to complete the full structural description of an interjection with its matrix clause.

This type of system poses similar problems in the case of knowable omissions. While these structures are indeed syntactically ill-formed, their structure plus the semantic and syntactic inferences possible to make from that structure and from the context carry important discourse information and functions. An MT system dealing with spontaneous speech must recognize the essential contribution of semantic structures that are only inferred and incorporate some means of arriving at these inferences for incorporation into the structural description of the utterance.

4.4 Noun phrase phenomena

There are a number of uses of noun phrases that require special attention from an MT system. Short answers have already been alluded to above; it will require the incorporation of illocutionary force information for the single NP in utterance 249 below to be correctly identified as an answer to 247, rather than to be discarded as an incomplete syntactic structure:

246 B: ...there may be an interview round about January
247 A: yeah. You heard anything about this?
249 B: Ø nothing at all yet (S.1.1)

Structures such as topicalization, left and right dislocation, what used to be called "Heavy NP Shift," and the use of appositives are all, strictly speaking, grammatical, yet they are also optional phenomena whose use is more often described in rhetoric texts than in MT grammars. However, they are fairly common phenomena in the conversations examined here:

393 B: I once knew an American who used to leave pieces of French lying around just to impress people ... even I don't do that
400 A: no, **your books** you hide [topicalization] (S.2.10)

583 B: **whereas Pickering and the linguistic group**, they just set out to do purely scientific texts [left dislocation] (S.1.5)

1 A: isn't this going to be a strange and impossible task for me **picking up linguistics** [right dislocation] (S.1.5)

418 A: I once found hidden in the back of one of Deb's drawers **a little book called How to Bluff Your Way Through Music** ["Heavy NP Shift"] (S.1.5)

1872 A: ...I had a letter, **an official letter**, ages ago from Miss Baker [appositive] (S.1.5)

These examples follow the usual characterization of these structures, but there are many other variations on these types of construction which do not. These other variations function usually to clarify or amplify a noun phrase already mentioned or about to be mentioned. The samples below illustrate some of the structural differences that make subsuming these utterances under the traditional frameworks problematical.

401 B: it really was Beryl that did it I think **Beryl Martin** (S.1.5): the NP "Beryl Martin" cannot be analysed as an appositive because it is not placed proximally to the NP "Beryl."

462 B: I purely program, largely for Bill
463 A: oh I see
464 B: **Bill and Hart** (S.1.5): Again, intervening material, this time from another speaker, prevents this from being analyzed as an appositive. Further, the NP "Bill and Hart" does not refer to the same entity as the NP in the matrix sentence, "Bill," and so cannot be considered an appositive or a dislocated NP.

143 B: I've got about a week of fairly hard work after the fourth of July, **this CSC stuff** you see and after that ... (S.1.1): In the above examples, there is some identity of noun phrase contents that might allow an analysis to recognize the link between the matrix NP and the moved NP. In this case, the link is a tenuous semantic one, at best. The analysing grammar must recognize that "this CSC stuff" is in some way linked to "fairly hard work" in order to find a place for the former NP in the analysis of this sentence.[12]

1136 A: my grandmother was the only one of our family who knew anything about sex and she was too old for it
1141 B: **Granny Bunn**
1142 A: yea. ooh she was gorgeous (S.2.10): This example poses several problems. First of all, assigning the same referent to "my grandmother," "she," and "Granny Bunn," may be problematical (see footnote 12). Second, the amplifying or clarifying use of the NP occurs in the turn of a speaker other than the speaker who uses the NP that is being clarified. Third, "Granny Bunn" could be analysed as filling two different functions: a right dislocated phrase

[12] This is the same sort of capability that will be required to assign the same referent identities to the two noun phrases in examples such as:

Arnold Schwarzenegger stepped up to the platform and waved. **The big man** dwarfed the reporter standing in front of him.

from utterance 1136 or a left dislocated phrase from 1142.

In these cases, there is at least a possibility that a very sophisticated analysis system could identify a structure in a matrix clause with which to associate the outlier NP, thereby assigning it some sort of legitimate place in the structure as some kind of adjunct. However, there are other cases where noun phrases simply occur alone; not as appendages to other utterances, and not as (short) answers to questions, but simply as what appears to be a statement of topic. A number of examples may illustrate the type of function these lone NP's seem to fulfill:

64 A: **one other thing. Sam Delaney, a Canadian who's graduated.** Delaney's the Canadian student, remember, last year? (S.1.1)

445 B: Harold is just winding up his Ph.D. but has been teaching for longer
447 A: mm
448 B: but he last year was my first year and he was certainly teaching before last year
450 A: mm
451 B: and he's the sort of next one, you know, next senior one after Hart
454 A: mm **Harold** (S.1.5)

690 C: well have we decided then, **the grand tour** (S.1.5)

422 c: this is a very *good book*
423 d: *oh **that one*** (S.2.10)

614 B: hey what about Janis Joplin? She's really great. Monty Python, he's amusing
618 c: **God father Part Two** (S.2.10)

An MT grammar flexible enough to allow partial analyses, without rejecting structures that are not well-formed sentences would give these an adequate analysis. However, distinguishing them from noun phrases that are uttered and then corrected and thus ought to be ignored will be a difficult task, requiring at the very least, some sort of discourse topic tracking.

4.5 Sentence level issues

Along with breaks, corrections, and repetitions, fragmentary exclamatory phrases are often touted as examples of phenomena that pose problems for an MT system. There are a great number of these sorts of exclamations sprinkled liberally throughout the three conversations; in Appendix B are listed all the types found in these conversations, sorted alphabetically, by frequency of appearance across conversations, and by overall frequency. In fact, however, a close examination of this list reveals certain groupings of expressions which lend themselves to analysis in various ways. Each of these groups will be discussed below; Appendix C gives a complete listing by group.

4.5.1 Idiomatic phrases and structures

A large number of these expressions are singular, that is, occurring alone, without arguments or structural attachment to the matrix sentence, and may be entered into the lexicon as idioms. Expressions such as "bless you," "golly," or "hey," can be recognized wherever they occur in an utterance and their semantic contribution incorporated into the semantic structure of the utterance in an appropriate way. A few expressions, however, also have functions as "normal" words and their use as expressions must be disambiguated from this other function in some way: "God," "good," "goodness," "look," "lovely," etc. Some of these expressions may co-occur with "oh," in which case their function as exclamatory expressions is clear: "oh great," "oh look," "oh heavens," etc.

There are also a small number of idiomatic structures which could simply be registered as such in an MT grammar:

have NP: 1254 B: darling *have some cherries*
1255 d: *have a hand*ful (S.2.10)

how AP: 797 B: but we can't find [a certain movie] - - I mean it just - its - just not on any more - and Richmond [theater] doesn't know where it's gone
800 d: oh how boring of them (S.2.10)

like NP 199 c: we wondered whether to try to organize it all quite a bit earlier...
200 A: what like next week (S.2.10)

what NP 860 c: and [I met] an enormous eighteen year old firefighting for the summer
862 B: what a chap to meet (S.2.10)

The latter three structures, of course, also occur within syntactically well-formed, complete sentences.[13]

A number of idiomatic structures are used as hedges. As in the case of illocutionary acts, in which surface syntactic structure is secondary to semantic force, so with these structures, which consist of a conjunction and some additional element, but whose semantic contribution is not to conjoin but to indicate the speaker's attitude toward his utterance. Examples are: "and so on," "and stuff," "or so," and "or something."

These conversations also contained phrases idiomatic for specific purposes such as showing gratitude or making apologies. Examples of the former include "thanks awfully," and "thank you very much indeed," and of the latter, "sorry."[14] In addition, there are a number of phrases which are very explicit, standard markers for discourse functions; "what's his name," and "what was the other thing I wanted to ask you?" are not intended as questions requiring an answer from the interlocutor. Instead, they are rhetorical. Other examples of explicit discourse function phrases occurring in these conversations are: "that finishes that," "let me tell you a story," and "I've got a problem for you."

4.5.2 Yes/no answers

A wide variety of expressions are used to signify agreement or disagreement with, or understanding of a previous utterance. "All right," "goodness no," "OK," "why not," "yea," "yes, of course/quite/that's so/ that's right," and "right" are all used in these ways. In any

[13] The differentiation of these particular uses of these expressions from other uses is by no means trivial. Illocutionary force type information could help to disambiguate the Offer use of "have NP" from its interpretation as a question containing the knowable omission of "do you" as in "[do you] have some coffee?" Intonation information could disambiguate the Question "how lovely?" from the Exclamation "how lovely!" Similarly, the intonations and pausal structures of "what, freedom?" [="what are you talking about? freedom?"]; "what freedom?" [="what freedom do you mean?"]; and "what freedom!" could differentiate those meanings. However, that would imply not only access among all levels of analysis in the system, but also specifications of structure and function as yet unknown.

[14] "Sorry" may be an example of an utterance deriving from a knowable omission, in this case, the omission of subject and copula, a common omission type. The idiomatic analysis of similar phrases has been discussed above for examples such as "good idea." "One other thing" is another phrase that could be analysed either as an idiom or as a knowable omission.
There is a large number of these types of expressions, each particular to certain social contexts. So, for example, telephone conversations will have opening and closing idioms, idioms which will differ from the opening and losing phrases used in face-to-face information-gathering tasks.

system dealing with these expressions it will be important to differentiate the agreement/disagreement function from the understanding function; despite the fact that "yes" and "no" seem to be contraries and could not be used together to denote agreement (or disagreement), they are sometimes used together to signify understanding:

1182 D: because being over here we tend to be a bit isolated
1184 A: yeah
1186 D: especially as we don't go to to coffee over in the main building you see
1187 A: **no yes** (S.1.5)

4.5.3 Discourse markers

Schiffrin (1987) defines discourse markers as "**sequentially dependent** elements which bracket units of talk," where "sequentially dependent" "indicate[s] that markers are devices that work on a discourse level: they are not dependent on the smaller units of talk of which the discourse is composed." Such markers may serve to signal information states, participation possibilities for interactors, or ideational relations within the discourse. While these markers may also have grammatical roles, their function is much better described in non-syntactic terms as elements which signal and regulate turn-taking and conversational coherence. Schiffrin discusses a number of examples which I have called exclamatory phrases and which appear in the conversations under examination here: "oh," "well," "now" and "then." Other such expressions appear in these conversations as well: "as a matter of fact," "at any rate," "let me see," "say" and "why,..." These have been discussed above (though see Schiffrin (1987) for a detailed discussion of their discourse contributions).

There are two other syntactic categories of discourse markers. One involves subject/verb expressions which, if taken in their literal sense, form a matrix clause for a complement clause. Her examples are "I mean" and "y'know;" others found in these conversations include: "you see," I don't know/think/suppose," "I must say," "I'll bet," "I agree," and "I'm pretty sure." That the discourse functions of these expressions must be differentiated from their literal meanings is clear; consider this example:

223 B: I'll give you a couple of hundred tiles maybe ...
225 c: what do **you mean** a couple of hundred tiles. why have you got a couple of hundred tiles
226 B: oh **I don't know** you just get left with these things **you know.** **I mean** bath tiles **you know**, nice ones. (S.2.10)

The use of "you mean" is a literal use; c is asking for clarification of B's reference to tiles. B's use of "I don't know," on the other hand, is a discourse use and not a literal use; he, in fact, clearly does know something, namely what he is telling c. Furthermore, despite B's use of "you know," c does *not* know, since he asked the question in the first place; "you know" has its discourse function here. B then uses "I mean" in its literal use to clarify "tiles." In that sense, "I mean" generally follows another phrase which is being clarified, as it does in this example ("these things"). But there is no previous utterance that could be considered clarified by the clause following "I mean" in the following:

121 B: When we were cos when you're engaged **I mean** people want to see you the whole time.... (S.2.10)

The second other syntactic category of discourse markers are conjunctival elements. Schiffrin (1987) lists "and," "but," "so," and "because," and other examples from this corpus are "except," "or," and "yet." These are problematic because, although they are members of a standard grammatical category and have grammatical functions, in some cases, they do not serve those functions, but rather mark discourse relations.[15] That this is the case can be seen in the following example in which "and " and "but," grammatically mutually exclusive elements, co-occur :

1222 D: **but** in fact I was in the office two years in a junior post and then they shoved shoved me up
1225 A: m yea
1227 D: **and** **but** I haven't had anyone for the last three gosh three months or so (S.1.5)

The standard example of the use of conjunctival discourse markers is one in which the conjunctive element is utterance-initial, in the absence of any previous clause which it could be analyzed as conjoining to the following clause. The following are typical examples of each of the expressions listed above (these expressions when used in their grammatically established sense are shown in parentheses):

[15] Actually, it might be more correct to say that their *primary* function is to mark discourse relations in these cases; Schiffrin (1987) discusses the role that the influence of the basic meanings of these expressions plays in their discourse functions.

398 B: *it's just* how it's grown up you **see**
400 C: ***and** uh* **yes**
401 B: it's really was Beryl that did it I think Beryl Martin
403 C: **but** surely I mean they can't *<<4-5 syllables>>*
405 B: ***and** I think Marilyn's* changed it a bit since Beryl left **and** I don't know you might find that you don't agree with various things. I should say so
413 A: **but** what functions do people variously fill? (S.1.5)

218 A: **cos** he tried to get me in in fact. I was slightly annoyed (but) actually it rather amused me. **but** um he said oh well you know I might get terribly - - you know I'm I'm just hanging on now (and) could take you on permanently - we may need you to do some work in the evening (so) I said fine being obliging **so** I got a peremptory command over the phone right - when can you come - **so** I said oh I'll come when it suits you (S.1.5)

574 A: I'm sure that you know he would be awfully grateful if you could see him in your office sometime
576 B: well, I'd like to have a chat with <<3 syllables>>
577 A: <<2 syllables>> **because** if if he doesn't work in close collaboration with you (and) - - (and) try to get - your experience - - he's going to go badly - at sea (S.1.1)

485 A: Anyway I used to go into the hospital in the evenings and find her - - sort of in real great pain because she'd laughed so much she'd burnt a couple - burst a couple of stitches - **except** that's the other thing about How to Bluff your Way Through Music it's the sort of book that people hide (S.2.10)

194 B: when are you going to see your parents? **or** not before the wedding (or) are you going over in the summer? (S.2.10)

777 B: God I thought it was old Joe Wright who'd walked in at first
780 A: It is extraordinary, isn't it yes
783 B: **yet** I gather they're they've got quite a good opinion of him there (S.2.10)

4.5.4 Cooperative structures

Sometimes conjunctive discourse markers also function in the larger set of cooperative constructions (also called "joint productions," see Ferrara (1992)). These are constructions in which the syntactic structure begun by one speaker is completed by another speaker. Cooperative structures and the use of conjunctions to take over the conversational turn blend into one another as in this example:

572 B: the ones that aren't amusing are things like Bluff Your Way Through Accountancy
574 c: (laughs)
575 B: cos that matters - if you're trying to be an accountant
577 c: **and** doesn't matter if you're not - at all

If MT systems were to be used to generate translations of conversations that had already occurred, these examples, and the many others in which speakers complete a wide variety of each other's syntactic structures, would pose a problem for the one-utterance-per-turn view of conversation. However, since the most reasonable and likely use for MT systems involves incorporation of the machine into the conversational structure, turn-taking will necessarily be rigidly structured enough so that these sorts of overlaps will not occur.

4.5.5 Combinations

All of the expressions, exclamatory, clausal and connective, discussed in this section can also occur in combination. Depending upon the expressions involved and the intonation with which they are used, they may simply compound their individual meanings or the combination may create a singular function/meaning differing from those of both of the combined elements. "Well" and "oh" combine with other elements especially frequently. Examples of combinations include "well quite," "well now," "well then," "well yes," "yes well," "oh yes," "oh no," "no now," "sort of you know," etc., practically *ad infinitum.*

5 Conclusion

In the current state of the development of MT systems, the goal of the translation of spontaneous conversation is still a long way off. The examination of phenomena in spontaneous conversation that do not yield to analysis by current MT grammars is an important step in determining future directions and goals of the MT effort. It can be hoped that speakers in an MT context will automatically make accommodations that will make the translation of their speech a more accessible goal, but the nature of these accommodations has yet to be determined. A very few of the problems discussed above can be handled by the fine-tuning of currently utilized grammar systems; most of the difficulties pointed out here challenge the MT research community to create other avenues through which these phenomena can be successfully attacked: the incorporation into the system of speech act, discourse, and cognitive information; the use of more flexible systems tolerant of partial analyses; the creation of accessibility among all analysis levels within the system; the exploitation of a variety of interactional modalities; the incorporation of feedback and correction possibilities both from the computer to the human and and vice versa. An effective and efficient combination of all of these options, and others, will be necessary in order to build a machine translation system that can begin to accommodate natural conversation.

Appendix A: Extract from Natural Conversation, Annotated for Grammaticality

The following excerpt from conversation S.1.5 (Svartvik and Quirk, 1980; p. 133-134) has been analyzed with respect to the ability of the English analysis grammar as currently developed (Fais, 1993) to produce an analysis for the utterances it contains. Any structures that that system can not handle have been set in boldface print. The incorporation of some of those phrases or structures is a trivial matter (e.g., "yes"), while some of the other examples pose serious problems for any MT grammar. (See text for full discussion.)

338: C: **and** I don't envy you the slips so I'd better not say any more about those *laughs*
B: *well Grace did some 3 sylls* that was the way um when I first came Beryl Martin Mervyn was the sort of mathematician cum programmer **you see** and I started working with him and his wife was English and she went through some first stage analyses with me just so I'd get the idea and we did some slips together it's pretty tedious though isn't it
A: m
B: it was terribly intriguing
A: m
B: *well **actually***
A: *what are you doing you're checking them **or putting I mean putting them into** (??)
C: well initially **I suppose** you just type them **I mean** I don't *know any* English or
A: *yes I suppose you do*
C: linguistics **so I was just** you just accept things to type because you know eventually you have to sort of sort things out yourself
B: **finding examples and underlining things *and* then working out**
A: ***yes* yea**
B: **what's what**
A: **oh yea but as** even **a** as a research assistant in six months' time that's what I'll be doing
C: **yes in that case as well**
B: **and then you may get your you** know you may not quite **think quite** agree with their system
A: **yea**
B: **now even as a non-English person at times sort of wanting to * say***
A: *m*
B: well **now what about**

A: **yea** ***yea yea***

C: well I'm *glad it* wasn't just me because there was some peculiar convention about hyphens which just

B: seemed quite *arbitrary*

C: *it was* absolutely illogical

B: **it's just** how it's grown up you +see+

C: ****and uh**** **+yes+**

B: it's really Beryl that did it I think **Beryl Martin**

C: **but** surely **I mean** they *can't 4-5 sylls*

B: ***and** I think Marilyn's* changed it a bit since ** . ** since Beryl left

C: **m**

B: **and I don't know** you might find that you don't agree with various things I should say so

Appendix B: Distribution of Fragmentary Expressions

Sorted alphabetically:

Expression	Number of occurrences	In number of conversations
• about	1	1
• absolutely	1	1
• absolutely not	1	1
• ah	3	1
• ah yes	1	1
• all right	1	1
• and so on	1	1
• and stuff	1	1
• anyway	11	3
• as a matter of fact	2	1
• as sort of you know	1	1
• as you know	4	3
• as you say	3	2
• at any rate	1	1
• at least	2	1
• bless you	1	1
• blimey	1	1
• certainly	1	1
• come on	1	1
• cor	1	1
• damn	1	1
• eh?	1	1
• exactly	2	1
• fancy that	1	1
• god	1	1
• god damnation	1	1
• golly	2	1
• good	4	1
• good heavens	3	2
• good idea	1	1
• good lord	1	1
• goodness	1	1
• goodness no	1	1
• gosh	2	2

• have NP	5	1
• hey	1	1
• how AP	3	1
• I agree	1	1
• I don’t know	7	3
• I don't suppose	1	1
• I don’t think	4	2
• I found	1	1
• I imagine	1	1
• I know	4	2
• I mean	60	3
• I must say	4	3
• I say	1	1
• I see	4	2
• I should think	1	1
• I shouldn’t think	1	1
• I suppose	7	3
• I think	15	3
• I’ll bet	1	1
• I’m just saying	1	1
• I’m pretty sure	1	1
• in any case	3	1
• in fact	6	1
• kind of	3	1
• let me see	2	1
• like NP	5	1
• look	1	1
• lovely	1	1
• much to my shame	1	1
• my God	2	1
• name	3	1
• never mind	1	1
• no	36	3
• no quite	2	2
• no thanks	1	1
• no, now	1	1
• now	3	3
• of course	1	1
• oh	53	3
• oh christ	1	1

• oh crikey	1	1
• oh fantastic	1	1
• oh god	2	1
• oh golly	2	1
• oh good	2	1
• oh great	1	1
• oh honestly	1	1
• oh I beg your pardon	1	1
• oh I don’t know	1	1
• oh I know	1	1
• oh I see	5	2
• oh look	1	1
• oh no	6	3
• oh really	2	1
• oh sorry	2	1
• oh well	8	3
• oh yea	4	1
• oh yes	10	2
• oh you know	1	1
• OK	2	1
• ooh gosh	1	1
• ooh heavens	1	1
• or so	1	1
• or something	2	1
• probably	1	1
• quite	7	1
• quite good	1	1
• really	4	2
• right	2	2
• say	2	2
• sorry	1	1
• sort of	45	2
• terribly yes	1	1
• thank you	1	1
• thank you very much indeed	1	1
• thanks awfully	2	1
• then	1	1
• well	104	3
• well actually	1	1
• well at least	1	1

• well good	1	1
• well now	3	3
• well quite	1	1
• well yes	1	1
• well you know	1	1
• well, no	4	2
• well, say	1	1
• well, then	2	1
• well, you see	1	1
• what NP	4	1
• what?	2	1
• why	1	1
• why not	2	1
• yea	30	3
• yes	53	3
• yes but	2	1
• yes exactly	2	2
• yes well	2	2
• yes, of course	1	1
• yes, quite	1	1
• yes, that's right	1	1
• yes, that's so	1	1
• you know	42	3
• you see	33	3

Appendix B, cont'd: Distribution of Fragmentary Expressions

Sorted by conversation frequency:

Expression	Number of occurrences	In number of conversations
• well	104	3
• I mean	60	3
• oh	53	3
• yes	53	3
• you know	42	3
• no	36	3
• you see	33	3
• yea	30	3
• I think	15	3
• anyway	11	3
• oh well	8	3
• I don't know	7	3
• I suppose	7	3
• oh no	6	3
• as you know	4	3
• I must say	4	3
• now	3	3
• well now	3	3
• sort of	45	2
• oh yes	10	2
• oh I see	5	2
• I don't think	4	2
• I know	4	2
• I see	4	2
• really	4	2
• well, no	4	2
• as you say	3	2
• good heavens	3	2
• gosh	2	2
• no quite	2	2
• right	2	2
• say	2	2
• yes exactly	2	2
• yes well	2	2
• quite	7	1
• in fact	6	1
• have NP	5	1
• like NP	5	1
• well	5	1
• good	4	1

• oh yea	4	1
• what NP	4	1
• ah	3	1
• how AP	3	1
• in any case	3	1
• kind of	3	1
• name	3	1
• as a matter of fact	2	1
• at least	2	1
• exactly	2	1
• golly	2	1
• let me see	2	1
• oh god	2	1
• oh golly	2	1
• oh good	2	1
• oh really	2	1
• oh sorry	2	1
• OK	2	1
• or something	2	1
• thanks awfully	2	1
• well, then	2	1
• what?	2	1
• why not	2	1
• yes but	2	1

One occurrence in one conversation:

- about
- all right
- at any rate
- come on
- fancy that
- good idea
- hey
- I imagine
- I'll bet
- lovely
- no thanks
- oh crikey
- oh I beg your pardon
- oh you know
- probably
- thank you
- well actually
- well yes
- why
- yes, that's so
- absolutely
- and so on
- bless you
- cor
- god
- good lord
- I agree
- I say
- I'm just saying
- much to my shame
- no, now
- oh fantastic
- oh I don't know
- ooh gosh
- quite good
- thank you very much indeed
- well at least
- well you know
- yes, of course
- absolutely not
- and stuff
- blimey
- damn
- god damnation
- goodness
- I don't suppose
- I should think
- I'm pretty sure
- my God
- of course
- oh great
- oh I know
- ooh heavens
- sorry
- well good
- well, say
- yes, quite
- ah yes
- as sort of you know
- certainly
- eh?
- good
- goodness no
- I found
- I shouldn't think
- look
- never mind
- oh christ
- oh honestly
- oh look
- or so
- terribly yes
- then
- well quite
- well, you see
- yes, that's right

Appendix B, cont'd.: Distribution of Fragmentary Expressions

Sorted by overall frequency:

Expression	Number of occurrences	In number of conversations
• well	104	3
• I mean	60	3
• oh	53	3
• yes	53	3
• sort of	45	2
• you know	42	3
• no	36	3
• you see	33	3
• yea	30	3
• I think	15	3
• anyway	11	3
• oh yes	10	2
• oh well	8	3
• I don't know	7	3
• I suppose	7	3
• quite	7	1
• in fact	6	1
• oh no	6	3
• have NP	5	1
• like NP	5	1
• oh I see	5	2
• as you know	4	3
• good	4	1
• I don't think	4	2
• I know	4	2
• I must say	4	3
• I see	4	2
• oh yea	4	1
• really	4	2
• well, no	4	2
• what NP	4	1
• ah	3	1
• as you say	3	2
• good heavens	3	2
• how AP	3	1
• in any case	3	1
• kind of	3	1
• name	3	1
• now	3	3
• well now	3	3

• as a matter of fact	2	1
• at least	2	1
• exactly	2	1
• golly	2	1
• gosh	2	2
• let me see	2	1
• no quite	2	2
• oh god	2	1
• oh golly	2	1
• oh good	2	1
• oh really	2	1
• oh sorry	2	1
• OK	2	1
• or something	2	1
• right	2	2
• say	2	2
• thanks awfully	2	1
• well, then	2	1
• what?	2	1
• why not	2	1
• yes but	2	1
• yes exactly	2	2
• yes well	2	2

One occurrence in one conversation: as above

Appendix C: Classification of Fragmentary Expressions

Exclamations--idioms

- ah
- bless you
- blimey
- come on
- cor
- damn
- eh?
- fancy that
- god
- god damnation
- golly
- good
- good lord
- good heavens
- goodness
- gosh
- hey
- look
- lovely
- my God
- never mind
- with oh
  - oh christ
  - oh crikey
  - oh fantastic
  - oh god
  - oh golly
  - oh good
  - oh great
  - oh honestly
  - oh look
  - ooh gosh
  - ooh heavens
- quite good

Idiomatic structures

- have NP
- how AP
- like NP
- what NP

Qualifiers/hedges

- Phrasal
  - and so on
  - and stuff
  - or so
  - or something
- Modifiers
  - about
  - at least
  - exactly
  - kind of
  - quite
  - really
  - sort of
- Degree of certainty
  - absolutely
  - absolutely not
  - certainly
  - of course
  - oh really
  - probably

Idiomatic phrases for specific purposes

- no thanks
- thank you very much indeed
- thank you
- thanks awfully
- sorry

From knowable omissions

- good idea
- one other thing
- sorry

Yes/no

- all right
- goodness no
- no quite
- no
- OK
- right
- terribly yes
- why not
- yea
- yes exactly
- yes, of course
- yes, quite
- yes, that's right
- yes, that's so
- yes

Discourse markers

- oh
- anyway
- as a matter of fact
- as you know
- as you say
- at any rate
- in any case
- in fact
- let me see
- now
- much to my shame
- oh well
- say
- then
- well
- well actually
- why

Matrix clause type discourse markers

- you know
- you see
- oh you know
- I agree
- I don't know
- I don't suppose
- I don't think
- I found
- I imagine
- I know
- I mean
- I must say
- I say
- I see
- I should think
- I shouldn't think
- I suppose
- I think
- I'll bet
- I'm just saying
- I'm pretty sure
- with oh
  - oh I don't know
  - oh I know
  - oh I see
- well I mean

Conjunctive elements

- and
- but
- so
- because
- except
- or
- yet

Combinations

- ah yes
- as sort of you know
- no, now
- oh no
- oh sorry
- oh yea
- oh yes
- with well
  - well at least
  - well good
  - well now
  - well quite
  - well yes
  - well no
  - well say
  - well, then
  - well you know
- yes but
- yea, well

Bibliography

Fais, Laurel. 1993. An English analysis grammar in a unification-based framework. ATR Technical Report TR-I-0351. Kyoto, Japan: Advanced Telecommunications Research Institute.

Ferrara, Kathleen. 1992. The interactive achievement of a sentence: Joint productions in therapeutic discourse. *Discourse Processes*, **15**:2.

Greenbaum, Sidney, and Randolph Quirk. 1970. *Elicitation experiments in English: Linguistic studies in use and attitude*. London: Longman.

Labov, William. 1972. *Sociolinguistic patterns*. Oxford: Basil Blackwell.

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

Svartvik, Jan, and Randolph Quirk. 1980. *A corpus of English conversation*. Sweden: C W K Gleerup Lund.

Tannen, Deborah. 1989. *Talking voices: Repetition, dialogue, and imagery in conversational discourse*. Cambridge: Cambridge University Press.