ON THE LINGUISTIC CHARACTER OF NON-STANDARD INPUT

Anthony S. Kroch and Donald Hindle
Department of Linguistics
University of Pennsylvania
Philadelphia, PA 19104 USA

ABSTRACT

If natural language understanding systems are ever to cope with the full range of English language forms, their designers will have to incorporate a number of features of the spoken vernacular language. This communication discusses such features as non-standard grammatical rules, hesitations and false starts due to self-correction, systematic errors due to mismatches between the grammar and sentence generator, and uncorrected true errors.

There are many ways in which the input to a natural language system can be non-standard without being uninterpretable.* Most obviously, such input can be the well-formed output of a grammar other than the standard language grammar with which the interpreter is likely to be equipped. This difference of grammar is presumably what we notice in language that we call "non-standard" in everyday life. Obviously, at least from the perspective of a linguist, it is wrong to think of this difference as being due to errors made by the non-standard language user; it is simply a dialect difference. Secondly, the non-standard input can contain hesitations and self-corrections which make the string uninterpretable unless some parts of it are edited out. This is the normal state of affairs in spoken language so that any system designed to understand spoken communication, even at a rudimentary level must be able to edit its input as well as interpret it. Thirdly, the input may be ungrammatical even by the rules of the grammar of the speaker but be the expected output of the speaker's sentence generating device. This case has not been much discussed, but it is important because in certain environments speakers (and to some extent unskilled writers) regularly produce ungrammmatical output in preference to grammatically unimpeachable alternatives. Finally, the input that the system receives may simply contain uncorrected errors. How important this last source of non-standard input would be in a functioning system is hard to judge and would depend on the environment of use. Uncorrected errors are, in our experience, reasonably rare in fluent speech but they are more common in unskilled writing. These errors may be typographical, a case we shall ignore in this discussion, or they may be grammatical. Of most interest to us are the cases where the error is due to a language user attempting to use a standard language construction that he/she does not natively command.

In the course of this brief communication we shall discuss each of the above cases with examples, drawing on work we have done describing the differences between the syntax of vernacular speech and of standard writing (Kroch and Hindle, 1981). Our work indicates that these differences are sizable enough to cause problems for the acquisition of writing as a skill, and they may arise as well when natural language understanding systems come to be used by a wider public. Whether problems will indeed arise is, of course, hard to say as it depends on so many factors. The most important of these is whether natural language systems are ever used with oral, as well as typed-in, language. We do not know whether the features of speech that we will be outlining will also show up in "keyboard" language; for its special characteristics have been little studied from a linguistic point of view (for a recent attempt see Thompson 1980). They will certainly occur more sporadically and at a lower incidence than they do in speech; and there may be new features of "keyboard" language that are not predictable from other language modes. We shall have little to say about how the problem of non-standard input can be best handled in a working system; for solving that problem will require more research. If we can give researchers working on natural language systems a clearer idea of what their devices are likely to have to cope with in an environment of widespread public use, our remarks will have achieved their purpose.

Informal, generally spoken, English exists in a number of regional, class and ethnic varieties, each with its own grammatical peculiarities. Fortunately, the syntax of these dialects is somewhat less varied than the phonology so that we may reasonably approximate the situation by speaking of a general "non-standard vernacular (NV)", which contrasts in numerous ways with standard written English (SWE). Some of the differences between the two dialects can lead to problems for parsing and interpretation. Thus,

* The discussion in this paper is based an on-going study of the syntactic differences between written and of spoken language funded by the National Institute of Education under grants G78-0169 and G80-0163.

subject-verb agreement, which is categorical in SWE, is variable in NV. In fact, in some environments subject-verb agreement is rarely indicated in NV, the most notable being sentences with dummy there subjects. Thus, the first of the sentences in (1) is the more likely in NV while, of course, only the second can occur in SWE:

(1) a. There was two girls on the sofa.
b. There were two girls on the sofa.

Since singular number is the unmarked alternative, it occurs with both singular and plural subjects; hence only plural marking on a verb can be treated as a clear signal of number in NV. This could easily prove a problem for parsers that use number marking to help find subject-verb pairs. A further, perhaps more difficult, problem would be posed by another feature of NV, the deletion of relative clause complementizers on subject relatives. SWE does not allow sentences like those in (2); but they are the most likely form in many varieties of NV and occur quite freely in the speech of people whose speech is otherwise standard:

(2) a. Anybody says it is a liar.
b. There was a car used to drive by here.

Here a parser that assumes that the first tensed verb following an NP that agrees with it is the main verb, will be misled. There are severe constraints on the environments in which subject relatives can appear without a complementizer, apparently to prevent hearers from "garden-pathing" on this construction, but these restrictions are not statable in a purely structural way. A final example of a NV construction which differs from what SWE allows is the use of it for expletive there, as in (3):

(3) It was somebody standing on the corner.

This construction is categorical in black English, but it occurs with considerable frequency in the speech of whites as well, at least in Philadelphia, the only location on which we have data. This last example poses no problems in principle for a natural language system; it is simply a grammatical fact of NV that has to be incorporated into the grammar implemented by the natural language understanding system. There are many features like this, each trivial in itself but nonetheless a productive feature of the language.

Hesitations and false starts are a consistent feature of spoken language and any interpreter that cannot handle them will fail instantly. In one count we found that 52% of the sentences in a 90 minute conversational interview contained at least one instance (Hindle, 1981b). Fortunately, the deformation of grammaticality caused by self-correction induced disfluency is quite limited and predictable (Labov, 1966). With a small set of editing rules, therefore, we have been able to normalize more than 95% of such disfluencies in preprocessing texts for input to a parser for spoken language that we have been constructing (Hindle, 1981b). These rules are based on the fact that false starts in speech are phonetically signaled, often by truncation of the final syllable. Marking the truncation and other phonetic editing signals in our transcripts, we find that a simple procedure which removes the minimum number of words necessary to create a parsable sequence eliminates most ill-formedness.

The spoken language contains as a normal part of its syntactic repertoire constructions like those illustrated below:

(4) The problem is is that nobody understands me.
(5) That's the only thing he does is fight.
(6) John was the only guest who we weren't sure whether he would come.
(7) Didn't have to worry about us.

These are constructions that it is difficult to accomodate in a linguistically motivated syntax for obvious reasons. Sentence (4) has two tensed verbs; (5), which has been called a "portmanteau construction", has a constituent belonging simultaneously to two different sentences; (6) has a wh- movement construction with no trace (see the discussion in Kroch, 1981); and (7) violates the absolute grammatical requirement that English sentences have surface subjects. We do not know why these forms occur so regularly in speech, but we do know that they are extremely common. The reasons undoubtedly vary from construction to construction. Thus, (5) has the effect of removing a heavy NP from surface subject position while preserving its semantic role as subject. Since we know that heavy NPs in subject position are greatly disfavored in speech (Kroch and Hindle, 1981), the portmanteau construction is almost certainly performing a useful function in simplifying syntactic processing or the presentation of information. Similarly, relative clauses with resumptive pronouns, like the one in (6), seem to reflect limitations on the sentence planning mechanism used in speech. If a relative clause is begun without computing its complete syntactic analysis, as a procedure like the one in MacDonald

(1980) suggests, then a resumptive pronoun might be used to fill a gap that turned out to occur in a non-deletable position. This account explains why resumptive pronouns do not occur in writing. They are ungrammatical and the real-time constraints on sentence planning that cause speech to be produced on the basis of limited look-ahead are absent. Subject deletion, illustrated in (7), is clearly a case of ellipsis induced in speech for reasons of economy like contraction and cliticization. However, English grammar does not allow subjectless tensed clauses. In fact, it is this prohibition that explains the existence of expletive *it* in English, a feature completely absent from languages with subjectless sentences. Of course, subject deletion in speech is highly constrained and its occurrence can be accommodated in a parser without completely rewriting the grammar of English, and we have done so. The point here, as with all these examples, is that close study of the syntax of speech repays the effort with improvements in coverage.

The final sort of non-standard input that we will mention is the uncorrected true error. In our analysis of 40 or more hours of spoken interview material we have found true errors to be rare. They generally occur when people express complex ideas that they have not talked about before and they involve changing direction in the middle of a sentence. An example of this sort of mistake is given in (8), where the object of a prepositional phrase turns into the subject of a following clause:

(8) When I was able to understand the
explanation of the moves of the
chessmen started to make sense to
me, he became interested.

Large parts of sentences with errors like this are parsable, but the whole may not make sense. Clearly, a natural language system should be able to make whatever sense can be made out of such strings even if it cannot construct an overall structure for them. Having done as well as it can, the system must then rely on context, just as a human interlocutor would. Unlike vernacular speech, the writing of unskilled writers quite commonly displays errors. One case, which we have studied in detail is that of errors in relative clauses with "pied-piped" prepositional phrases. We often find clauses like the ones in (9), where the wrong preposition (usually *in*) appears at the beginning of the clause.

(9) a. methods in which to communicate with
other people
b. rules in which people can direct
their efforts

Since pied-piped relatives are non-existent in NV, the simplest explanation for such examples is that they are errors due to imperfect learning of the standard language rule. More precisely, instead of moving a *wh*- prepositional phrase to the complementizer position in the relative clause, unskilled writers may analyze the phrase *in which* as a general oblique relativizer equivalent to *where*, the form most commonly used in this function in informal speech.

In summary, ordinary linguistic usage exhibits numerous deviations from the standard written language. The sources of these deviations are diverse and they are of varying significance for natural language processing. It is safe to say, however, that an accurate assessment of their nature, frequency and effect on interpretability is a necessary prerequisite to the development of truly robust systems.

REFERENCES

Hindle, Donald. "Near-sentences in spoken English." Paper presented at NWAVE X, 1981a.

Hindle, Donald. "The syntax of self-correction." Paper presented at the Linguistic Society of America annual meeting, 1981b.

Kroch, Anthony. "On the role of resumptive pronouns in amnestying island constraint violations." in *CLS #17*, 1981.

Kroch, Anthony and Donald Hindle. *A quantitative study of the syntax of speech and writing*. Final report to the National Institute of Education on grant #78-0169, 1981.

Labov, William. "On the grammaticality of everyday speech." unpublished manuscript, 1966.

MacDonald, David "Natural language production as a process of decision-making under constraint." draft of an MIT Artifical Intelligence Lab technical report, 1980.

Thompson, Bozena H. "A linguistic analysis of natural language communication with computers." in *Proceedings of the eighth international conference on computational linguistics*. Tokyo, 1980.