Paraphrasing Using Given and New Information in a Question-Answer System

Kathleen R. McKeown
Department of Computer and Information Science
The Moore School
University of Pennsylvania, Philadelphia, Pa. 19104

ABSTRACT: The design and implementation of a paraphrase component for a natural language question-answer system (CO-OP) is presented. A major point made is the role of given and new information in formulating a paraphrase that differs in a meaningful way from the user's question. A description is also given of the transformational grammar used by the paraphraser to generate questions.

I. INTRODUCTION

In a natural language interface to a database query system, a paraphraser can be used to ensure that the system has correctly understood the user. Such a paraphraser has been developed as part of the CO-OP system [KAPLAN 79]. In CO-OP, an internal representation of the user's question is passed to the paraphraser which then generates a new version of the question for the user. Upon seeing the paraphrase, the user has the option of rephrasing her/his question before the system attempts to answer it. Thus, if the question was not interpreted correctly, the error can be caught before a possibly lengthy search of the database is initiated. Furthermore, the user is assured that the answer s/he receives is an answer to the question asked and not to a deviant version of it.

The idea of using a paraphraser in the above way is not new. To date, other systems have used canned templates to form paraphrases, filling in empty slots in the pattern with information from the user's question [WALTZ 78; CODD 78]. In CO-OP, a transformational grammar is used to generate the paraphrase from an internal representation of the question. Moreover, the CO-OP paraphraser generates a question that differs in a meaningful way from the original question. It makes use of a distinction between given and new information to indicate to the user the existential presuppositions made in her/his question.

II. OVERVIEW OF THE CO-OP SYSTEM

The CO-OP system is aimed at infrequent users of database query systems. These casual users are likely to be unfamiliar with computer systems and unwilling to invest the time needed to learn a formal query language. Being able to converse naturally in English enables such persons to tap the information available in a database.

In order to allow the question-answer process to proceed naturally, CO-OP follows some of the "co-operative principles" of conversation [GRICE 75]. In particular, the system attempts to find meaningful answers to failed questions by addressing any incorrect assumptions the questioner may have made in her/his question. When the direct response to a question would be simply "no" or "none", CO-OP gives a more informative response by correcting the questioner's mistaken assumptions.

The false assumptions that CO-OP corrects are the existential presuppositions of the question.* Since these presuppositions can be computed from the surface structure of the question, a large store of semantic knowledge for inferencing purposes is not needed. In fact, a lexicon and database schema are the only items which contain domain-specific information. Consequently, the CO-OP system is a portable one; a change of database requires that only these two knowledge sources be modified.

III. THE CO-OP PARAPHRASER

CO-OP's paraphraser provides the only means of error-checking for the casual user. If the user is familiar with the system, s/he can ask to have the intermediate results printed, in which case the parser's output and the formal database query will be shown. The naive user however, is unlikely to understand these results. It is for this reason that the paraphraser was designed to respond in English.

The use of English to paraphrase queries creates several problems. The first is that natural language is inherently ambiguous. A paraphrase must clarify the system's interpretation of possible ambiguous phrases in the question without introducing additional ambiguity.

One particular type of ambiguity that a paraphraser must address is caused by the linear nature of sentences. A modifying relative clause, for example, frequently cannot be placed directly after the noun phrase it modifies. In such cases, the semantics of the sentence may indicate the correct choice of modified noun phrase, but occasionally, the sentence may be genuinely ambiguous. For example, question (A) below has two interpretations, both equally plausible. The speaker could be referring to books dating from the '60s or to computers dating from the '60s.

(A) Which students read books on computers dating from the '60s?

A second problem in paraphrasing English queries is the possibility of generating the exact question that was originally asked. If a grammar were developed to simply generate English from an underlying representation of the question this possibility could be realized. Instead, a method must be devised which can determine how the phrasing should differ from the original.

The CO-OP paraphraser addresses both the problem of ambiguity and the rephrasing of the question. It makes the system's interpretation of the question explicit by breaking down the clauses of the question and reordering them dependent upon their function in the sentence. Thus, question (A) above will result in either paraphrase (B) or (C), reflecting the interpretation the system has chosen.

(B) Assuming that there are books on computers (those computers date from the '60s), which students read those books?

(C) Assuming that there are books on computers (those books date from the '60s), which students read those books?

The method adopted guarantees that the paraphrase will differ from the original except in cases where no relative clauses or prepositional phrases were used. It was formulated on the basis of a distinction between given and new information and indicates to the user the presuppositions s/he has made in the question (in the

*For example, in the question "Which users work on projects sponsored by NASA?", the speaker makes the existential presupposition that there are projects sponsored by NASA.

"assuming that" clause), while focussing her/his attention on the attributes of the class s/he is interested in.

IV. LINGUISTIC BACKGROUND

As mentioned earlier, the lexicon and the database are the sole sources of world knowledge for CO-OP. While this design increases CO-OP's portability, it means that little semantic information is available for the paraphraser's use. Contextual information is also limited since no running history or context is maintained for a user session in the current version. The input the paraphraser receives from the parser is basically a syntactic parse tree of the question. Using this information, the paraphraser must reconstruct the question to obtain a phrasing different from the original. The following question must therefore be addressed:

> What reasons are there for choosing one syntactic form of expression over another?

Some linguists maintain that word order is affected by functional roles elements play within the sentence.* Terminology used to describe the types of roles that can occur varies widely. Some of the distinctons that have been described include given/new, topic/comment, theme/rheme, and presupposition/focus. Definitions of these terms however, are not consistent (for example, see [PRINCE 79] for a discussion of various usages of "given/new").

Nevertheless, one influence on expression does appear to be the interaction of sentence content and the beliefs of the speaker concerning the knowledge of the listener. Some elements in the sentence function in conveying information which the speaker assumes is present in the "consciousness" of the listener [CHAFE 76]. This information is said to be contextually dependent, either by virtue of its presence in the preceding discourse or because it is part of the shared world knowledge of the dialog participants. In a question-answer system, shared world knowledge refers to information which the speaker assumes is present in the database. Information functioning in the role just described has been termed "given".

"New" labels all information in the sentence which is presented as not retrievable from context. In the declarative, elements functioning in asserting information that the listener is presumed not to know are called new. In the question, elements functioning in conveying what the speaker wants to know (i.e.- what s/he doesn't know) represent information which the speaker presumes the listener is not already aware of. Firbas identifies additional functions in the question. Of these, (ii) is used here to augment the interpretation of new information. He says:

> "(i) it indicates the want of knowledge on the part of the inquirer and appeals to the informant to satisfy this want.
>
> (ii) [a] it imparts knowledge to the informant in that it informs him what the inquirer is interested in (what is on her/his mind) and [b] from what particular angle the intimated want of knowledge is to be satisfied."
> [FIRBAS 74; p.31]

Although word order vis-a-vis these and related distinctions has been discussed in light of the declarative sentence, less has been said about the interrogative form. Halliday [HALLIDAY 67] and Krizkova* are among the few to have analyzed the question. Despite the fact that they arrive at different conclusions**, the two follow similar lines of reasoning. Krizkova argues that both the wh-item of the wh-question and the finite verb (e.g. - "do" or "be") of the yes/no question point to the new information to be disclosed in the response. These elements she claims, are the only unknowns to the questioner. Halliday, in discussing the yes/no question, also argues that the finite verb is the only unknown. The polarity of the text is in question and the finite element indicates this.

In this paper the interpretation of the unknown elements in the question as defined by Krizkova and Halliday is followed. The wh-items, in defining the questioner's lack of knowledge, act as new information. Firbas' analysis of the functions in questions is used to further elucidate the role of new information in questions. The remaining elements are given information. They represent information assumed by the questioner to be true of the database domain. This labeling of information within the question will allow the construction of a natural paraphrase, avoiding ambiguity.

V. FORMULATION

Following the analysis described above, the CO-OP paraphraser breaks down questions into given and new information. More specifically, an input question is divided into three parts, of which (2) and (3) form the new information.

> (1) given information
> (2) Function ii[a] from Firbas above
> (3) Function ii[b] from Firbas above

In terms of the question components, (2) comprises the question with no subclauses as it defines the lack of knowledge for the hearer. Part (3) comprises the direct and indirect modifiers of the interrogative words as they indicate the angle from which the question was asked. They define the attributes of the missing information for the hearer. Part (1) is formed from the remaining clauses.

As an example, consider question (D):

> (D) Which division of the computing facility works on projects using oceanography research?

Following the outline above, part (2) of the paraphrase will be the question minus subclauses: "Which division works on projects?". Part (3), the modifiers of the interrogative words, will be "of the computing facility" which modifies "which division". The remaining clause

---

* Some other influences on syntactic expression are discussed in [MORGAN and GREEN 73]. They suggest that stylistic reasons, in addition to some of the functions discussed here, determine when different syntactic constructions are to be used. They point out, for example, that the passive tense is often used in academic prose to avoid identification of agent and to lend a scientific flavor to the text.

* Summary by [FIRBAS 74] of the untranslated article "The Interrogative Sentence and Some Problems of the So-called Functional Sentence Perspective (Contextual Organization of the Sentence), Nasa rec 4, 1968.

** It should be noted that Halliday and Krizkova discuss the unknowns in the question in order to define the theme and rheme of a question. Although they agree about the unknowns for the questioner, they disagree about which elements function as theme and which function as rheme. A full discussion of their analysis and conclusions is given in [MCKEOWN 79].

"projects using oceanography research" is considered given information. The three parts can then be assembled into a natural sequence:

(E) Assuming that there are projects using oceanography research, which division works on those projects? Look for a division of the computing facility.*

In question (D), information belonging to each of the three categories occurred in the question. If one of these types of information is missing, the question will be presented minus the initial or concluding clauses. Only part (2) of the paraphrase will invariably occur. If more than one clause occurs in a particular category, the question will be furthered splintered. Additional given information is parenthesized following the "assuming that ..." clause. Example (F) below illustrates the paraphrase for a question containing several clauses of given information and no clauses defining specific attributes of the missing information. Clauses containing information characterized by category (3) will be presented as separate sentences following the stripped-down question. (G) below demonstrates a paraphrase containing more than one clause of this type of information.

(F) Q: Which users work on projects in oceanography that are sponsored by NASA?

P: Assuming that there are projects in oceanography (those projects are sponsored by NASA), which users work on those projects?

(G) Q: Which programmers in superdivision 5000 from the ASD group are advised by Thomas Wirth?

P: Which programmers are advised by Thomas Wirth? Look for programmers in superdivision 5000. The programmers must be from the ASD group.

VI. IMPLEMENTATION OVERVIEW

The paraphraser's first step in processing is to build a tree structure from the representation it is given. The tree is then divided into three separate trees reflecting the division of given and new information in the question. The design of the tree allows for a simple set of rules which flatten the tree. The final stage of processing in the paraphraser is translation. In the translation phase, labels in the parser's representation are translated into their corresponding words. During this process, necessary transformations of the grammar are performed upon the string.

Several aspects of the implementation will not be discussed here, but a description can be found in [MCKEOWN 79]. The method used by the paraphraser to handle conjunction, disjunction, and limited quantification is one of these. A second function of the paraphraser is also described in [MCKEOWN 79]. The set of procedures used to paraphrase the user's query can also be used to generate an English version of the parser's output. If the tree is not divided into given and new information, the flattening and transformational rules can be applied to produce a question that is not in the three-part form. In CO-OP, generation is used to produce corrections of the user's mistaken presuppositions.

* This example, as well as all sample questions and paraphrases that follow, were taken from actual sessions with the paraphraser. Question (A) and its possible paraphrases (B) and (C) are the only examples that were not run on the paraphraser.

A. THE PHRASE STRUCTURE TREE

In its initial processing, the paraphraser transforms the parser's representation into one that is more convenient for generation purposes. The resultant structure is a tree that highlights certain syntactic features of the question. This initial processing gives the paraphraser some independence from the CO-OP system. Were the parser's representation changed or the component moved to a new system, only the initial processing phase need be modified.

The paraphraser's phrase structure tree uses the main verb of the question as the root node of the tree. The subject of the main verb is the root node of the left subtree, the object (if there is one) the root node of the right subtree. In the current system, the use of binary relations in the parser's representation (see [KAPLAN 79] for a description of Meta Query Language) creates the illusion that every verb or preposition has a subject and object. The paraphraser's tree does allow for the representation of other constructions should the incoming language use them.

Each of the subtrees represents other clauses in the question. Both the subject and the object of the main verb will have a subtree for each other clause it participates in. If a noun in one of these clauses also participates in another clause in the sentence, it will have subtrees too.

As an example, consider the question: "Which active users advised by Thomas Wirth work on projects in area 3?". The phrase structure tree used in the paraphraser is shown in Figure 1. Since "work" is the main verb, it will be the root node of the tree. "users" is root of the left subtree, "projects" of the right. Each noun participates in one other clause and therefore has one subtree. Note that the adjective "active" does not appear as part of the tree structure. Instead, it is closely bound to the noun it modifies and is treated as a property of the noun.

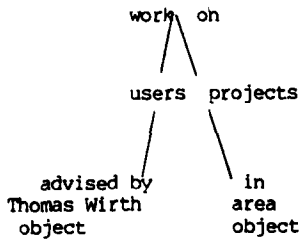
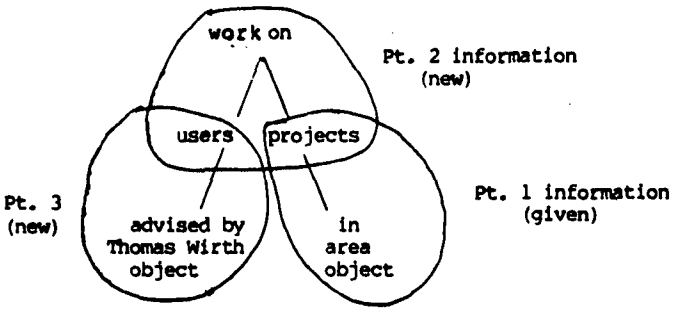


Figure 1

B. DIVIDING THE TREE

The constructed tree is computationally suited for the three-part paraphrase. The tree is flattened after it has been divided into subtrees containing given information and the two types of new information. The splitting of the tree is accomplished by first extracting the topmost smallest portion of the tree containing the wh-item. At the very least, this will include the root node plus the left and right subtree root nodes. This portion of the tree is the stripped down question. The clauses which define the particular aspect from which the question is asked are found by searching the left and right subtrees for the wh-item or questioned noun. The subtree whose root node is the wh-item contains these clauses. Note that this may be the entire left or right subtree or may only be a subtree of one of these. The remainder of the tree represents given information. Figure 2 illustrates this division for the previous example.



Q: Which active users advised by Thomas Wirth work on projects in area 3?
P: Assuming that there are projects in area 3, which active users work on those projects? Look for users advised by Thomas Wirth.

Figure 2

C. FLATTENING

If the structure of the phrase structure tree is as shown in Figure 3, with A the left subtree and B the right, then the following rules define the flattening process:

```
TREE-> A R B
SUBTREE -> R' A' B'
```

In other words, each of the subtrees will be linearized by doing a pre-order traversal of that subtree. As a node in a subtree has three pieces of information associated with it, one more rule is required to expand a node. A node consists of:

(1) arc-label
(2) set-label
(3) subject/object

where arc-label is the label of the verb or preposition used in the parse tree and set-label the label of a noun phrase. Subject/object indicates whether the sub-node noun phrase functions as subject or object in the clause; it is used by the subject-aux transformation and does not apply to the expansion rule. The following rule expands a node:

```
NODE -> ARC-LABEL SET-LABEL
```

Two transformations are applied during the flattening process. They are wh-fronting and subject-aux inversion. They are further described in the section on transformations.

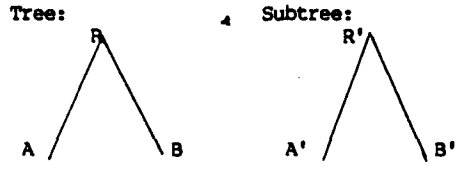


Figure 3

The tree of given information is flattened first. It is part of the left or right subtree of the phrase structure tree and therefore is flattened by a pre-order traversal. It is during the flattening stage that the words "Assuming that there [be] ... " are inserted to introduce the clause of given information. "Be" will agree with the subject of the clause. If there is more than one clause, parentheses are inserted around the additional ones. The tree representing the stripped down question is flattened next. It is followed by the modifiers of the questioned noun. The phrase "Look for" is inserted before the first clause of modifiers.

D. TRANSFORMATIONS

The grammar used in the paraphraser is a transformational one. In addition to the basic flattening rules described above, the following transformations are used:

```
wh-fronting
negation
do-support
subject-aux inversion
affix-hopping
contraction
has deletion
```

The curved lines indicate the ordering restrictions. There are two connected groups of transformations. If wh-fronting applies, then so will do-support, subject-aux inversion, and affix-hopping. The second group of transformations is invoked through the application of negation. It includes do-support, contraction, and affix-hopping. Has-deletion is not affected by the absence or presence of other transformations. A description of the transformation rules follows. The rules used here are based on analyses described by [AKMAJIAN and HENY 75] and analyses described by [CULLICOVER 76].

The rule for wh-fronting is specified as follows, where SD abbreviates structural description and SC, structural change:

```
SD:  X - NP - Y
     1   2    3
SC:  2+1 0    3
condition: 2 dominates wh
```

The first step in the implementation of wh-fronting is a search of the tree for the wh-item. A slightly different approach is used for paraphrasing than is used for generation. The difference occurs because in the original question, the NP to be fronted may be the head noun of some relative clauses or prepositional phrases. When generating, these clauses must be fronted along with the head noun. Since the clauses of the original question are broken down for the paraphrase, it will never be the case when paraphrasing that the NP to be fronted also dominates relative clauses or prepositional phrases. For this reason, when paraphrase mode is used, the applicability of wh-fronting is tested for and is applied in the flattening process of the stripped down question. If it applies, only one word need be moved to the initial position.

When generation is being done, the applicability of wh-fronting is tested for immediately before flattening. If the transformation applies, the tree is split. The subtree of which the wh-item is the root is flattened separately from the remainder of the tree and is attached in fronted position to the string resulting from flattening the other part.

After wh-fronting has been applied, do-support is invoked. In CO-OP, the underlying representation of the question does not contain modals or auxiliary verbs. Thus, fronting the wh-item necessitates supplying an auxiliary. The following rule is used for do-support:

```
SD:  NP - NP - tense - V - X
     1    2       3     4
SC:  1    do+2    3     4
condition: 1 dominates wh
```

Subject-aux inversion is activated immediately afterwards. Again, if wh-fronting applied, subject-aux inversion will apply also. The rule is:

```
SD:  NP - NP - AUX - X
     1    2     3    4
SC:  1    3+2   0    4
condition: 1 dominates wh
```

Affix-hopping follows subject-aux inversion. In the paraphraser it is a combination of what is commonly thought of as affix-hopping and number-agreement. Tense and number are attributes of all verbs in the parser's representation. When an auxiliary is generated, the tense and number are "hopped" from the verb to the auxiliary. Formally:

```
SD:  X - AUX - Y - tense-num-V - Z
     1    2     3         4    5   6
SC:  1    2+4   3         0    5   6
```

Some transformational analyses propose that wh-fronting and subject-aux inversion apply to the relative clause as well as the question. In the CO-OP paraphraser, the head-noun is properly positioned by the flattening process and wh-fronting need not be used. Subject-aux inversion however, may be applicable. In cases where the head noun of the clause is not its subject, subject-aux inversion results in the proper order.

The rule for negation is tested during the translation phase of execution. It has been formalized as:

```
SD:  X - tense-V - NP - Y
     1       2     3    4
SC:  1      2+no   3    4
condition: 3 marked as negative
```

In the CO-OP representation, an indication of negation is carried on the object of a binary relation (see [KAPLAN 79]). When generating an English representation of the question, it is possible in some cases to express negation as modification of the noun (see question (H) below). In all cases however, negation can be indicated as part of the verb (see version (I) of question (H)). Therefore, when the object is marked as negative, the paraphraser moves the negation to become part of the verbal element.

(H) Which students have no advisors?
(I) Which students don't have advisors?

In English, the negative marker is attached to the auxiliary of the verbal element and therefore, as was the case for questions, an auxiliary must be generated. Do-support is used. The rule used for do-support after negation differs from the one used after wh-fronting. They are presented this way for clarity, but could have been combined into one rule.

```
SD:  X - tense-V-no - Y
     1       2      3
SC:  1     do+2     3
```

Affix-hopping, as described above, hops the tense, number, and negation from the verb to the auxiliary verb. The cycle of transformations invoked thru application of negation is completed with the contraction transformation. The statement of the contraction transformation is:

```
SD:  X - do+tense -no - Y
     1      2      3    4
SC:  1    #2+n't#  0    4
```

where # indicates that the result must be treated as a unit for further transformations.

VII. CONCLUSIONS

The paraphraser described here is a syntactic one. While this work has examined the reasons for different forms of expression, additions must be made in the area of semantics. The substitution of synonyms, phrases, or idioms for portions or all of the question requires an examination of the effect of context on word meaning and of the intentions of the speaker on word or phrase choice. The lack of a rich semantic base and contextual information dictated the syntactic approach used here, but the paraphraser can be extended once a wider range of information becomes available.

The CO-OP paraphraser has been designed to be domain-independent and thus a change of the database requires no changes in the paraphraser. Paraphrasers which use the template form however, will require such changes. This is because the templates or patterns, which constitute the type of question that can be asked, are necessarily dependent on the domain. For different databases, a different set of templates must be used.

The CO-OP paraphraser also differs from other systems in that it generates the question using a transformational grammar of questions. It addresses two specific problems involved in generating paraphrases:

1. ambiguity in determining which noun phrases a relative clause modifies
2. the production of a question that differs from the user's

These goals have been achieved for questions using relative clauses through the application of a theory of given and new information to the generation process.

ACKNOWLEDGEMENTS

This work was partially supported by an IBM fellowship and NSF grant MCS78-08401. I would like to thank Dr. Aravind K. Joshi and Dr. Bonnie Webber for their invaluable comments on the style and content of this paper.

REFERENCES

1. [AKMAJIAN and HENY 75]. Akmajian, A. and Heny, F., An Introduction to the Principles of Transformational Syntax, MIT Press, 1975.

2. [CHAFE 77]. Chafe, W. L., "Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Points of View", Subject and Topic (ed. C. N. Li), Academic Press, 1977.

3. [CODD 78]. Codd, E. F., et al., Rendezvous Version 1: An Experimental English-language Query Formulation System for Casual Users of Relational Data Bases, IBM Research Report RJ2144(29407), IBM Research Laboratory, San Jose, Ca., 1978.

4. [CULLICOVER 76]. Cullicover, P. W., Syntax, Academic Press, N. Y., 1976.

5. [DANES 74]. Danes, F. (ed.), Papers on Functional Sentence Perspective, Academia, Prague, 1974

6. [FIRBAS 66]. Firbas, Jan, "On Defining the Theme in Functional Sentence Analysis", Travaux Linguistiques de Prague 1, Univ. of Alabama Press, 1966.

7. [FIRBAS 74]. Firbas,Jan, "Some Aspects of the Czechoslovak Approach to Problems of Functional Sentence Perspective", Papers on Functional Sentence Perspective, Academia, Prague, 1974.

8. [GOLDMAN 75]. Goldman, N., "Conceptual Generation", Conceptual Information Processing (R. C. Schank), North-Holland Publishing Co., Amsterdam, 1975.

9. [GRICE 75]. Grice, H. P., "Logic and Conversation", in Syntax and Semantics: Speech Acts, Vol. 3, (P. Cole and J. L. Morgan, Ed.), Academic Press, N. Y., 1975.

10. [HALLIDAY 67]. Halliday, M.A.K., "Notes on Transitivity and Theme in English", Journal of Linguistics 3, 1967.

11. [HIEDORN 75]. Heidorn, G., "Augmented Phrase Structure Grammar", TINLAP-1 Proceedings, June 1975.

12. [JOSHI 79]. Joshi, A. K., "Centered Logic: the Role of Entity Centered Sentence Representation in Natural Language Inferencing", to appear in IJCAI Proceedings 79.

13. [KAPLAN 79]. Kaplan, S. J., "Cooperative Responses from a Portable Natural Language Data Base Query System", Ph.D. Disseration, Univ. of Pennsylvania, Philadelphia, Pa., 1979.

14. [MCDONALD 78]. McDonald, D. D., "Subsequent Reference: Syntactic and Rhetorical Constraints", TINLAP-2 Proceedings, 1978.

15. [MCKEOWN 79]. McKeown, K., "Paraphrasing Using Given and New Information in a Question-Answer System", forthcoming Master's Thesis, Univ. of Pennsylvania, Philadelphia, Pa., 1979.

16. [MORGAN and GREEN 77]. Morgan,J.L. and Green, G.M.: "Pragmatics and Reading Comprehension", University of Illinois, 1977.

17. [PRINCE 79]. Prince, E., "On the Given/New Distinction", to appear in CLS 15, 1979.

18. [SIMMONS and SLOCUM 72]. Simmons, R. and Slocum, J., "Generating English Discourse from Semantic Networks", Univ. of Texas at Austin, CACM, Vol. 5, #10, October 1972.

19. [WALTZ 78]. Waltz, D.L., "An English Language Question Answering System for a Large Relational Database", CACM, Vol. 21 #7, July 1978.