Discourse Cues In Narrative Text: Using Production To Predict Comprehension

David Allbritton
University of Pittsburgh
Learning Research and Development Center
3939 O'Hara St.
Pittsburgh, PA 15260
allbritt+@pitt.edu

Johanna Moore
University of Edinburgh
Human Communication Research Centre
2 Buccleuch Place
Edinburgh, UK EH8 9LW
J.Moore@ed.ac.uk

Abstract

Discourse cues (e.g., *because, since, therefore*, and *so*) express structural and semantic relationships between parts of an explanation, and are used extensively by human tutors. Previous research by Moser and Moore (in preparation) revealed patterns of cue occurrence, placement, and selection in the speech of expert tutors during tutorial dialogues. An experiment measuring both reading times and recognition memory for cued and uncued sentences in a narrative text was conducted to test the hypothesis that patterns of cue usage observed in the speech of expert human tutors could predict comprehension. Small but reliable effects on comprehension were observed. The results are discussed in relation to strategies for generating comprehensible text in an Intelligent Tutoring System, differences between narrative and expository texts, and the role of distributional analyses of naturally occurring dialogue in understanding discourse comprehension.

Introduction

The findings reported here are part of a program of research intended to improve the quality of automated text generation systems used in intelligent tutoring systems (ITSs). Although ITSs have been shown to improve student test scores by as much as 1 standard deviation (Anderson *et al.* 1995), skilled human tutors can achieve improvements of up to 2 standard deviations (Bloom 1984). A general strategy for increasing the effectiveness of ITSs and bridging this gap is to examine what effective human tutors do and then implement the findings in an ITS. The research described here is one step in the process of applying this strategy to a specific aspect of tutorial dialogues: the use of discourse cues by human tutors.

Discourse cues (e.g., *because, since, therefore*, and *so*) express structural and semantic relationships between parts of an explanation, and are used extensively by human tutors. With the goal of incorporating the appropriate use of discourse cues in the text generated by an ITS, we are following a three-stage approach. The first stage involves examining corpora of tutorial dialogues from skilled human tutors to identify patterns of cue usage. The second stage is to experimentally test the assumption that the patterns of cue usage most frequently observed in production by tutors do indeed lead to better comprehension. The third stage is to implement in an automated text generation system the strategies for cue usage identified in the first two stages. The research presented here is part of the second stage of this project, comparing comprehension of text segments that contain more vs. less frequently observed patterns of cue placement.

The patterns of cue production for which comprehension was tested were derived from a corpus study (Moser & Moore in preparation) that employed the Relational Discourse Analysis (RDA) coding scheme (Moser, Moore, & Glendening 1996). In RDA, segments of a discourse are coded to indicate their intentional structure. Each segment consists of a *core* – the element that most clearly expresses the purpose of the segment – and one or more *contributors* that support the purpose expressed in the core.

The following example from the corpus illustrates several features of core:contributor relations and the use of discourse cues to signal those relations. The dialogues between tutors and students in the corpus were collected during electronics trouble-shooting simulations and subsequent debriefing sessions. This excerpt is from one of the debriefing sessions (Moser & Moore, p. 18):

Student: Why is testing pins in part3 not a good thing?
Tutor: <u>Although</u>
A. you know that part1 is good,
B. you should eliminate part2 before troubleshooting in part3.
<u>(This is) because</u>
C. 1. part2 is moved frequently
<u>and thus</u>
2. is more susceptible to damage.
<u>Also,</u>
D. it is more work to open up part3 for testing.

Statement *B* is the core of the entire segment, with *A*, *C*, and *D* all functioning as contributors in support of *B*. In *C* we find an embedded relation, with *C1* as the contributor to the core *C2*. "Although" in *A*, and "because" in *C1* are both cues signalling that what follows is a contributor, while "and thus" in *C2* is a cue signalling that what follows is a core.

We would expect that core-contributor order should be easier to understand than contributor-core, based on the different processing demands that each would presumably impose on listeners, as Moser and Moore (in preparation) predict. Hearing the core first should help guide the interpretation of the contributor that follows it, much the way having a macro-proposition for the gist of a text can guide interpretation of subsequent propositions in Kintsch's Construction-Integration Theory of text comprehension (Kintsch 1988). Also, assuming that something like the CI Model's short-term memory buffer is required to hold information until the listener can tell how to integrate it into her representation of the discourse, a contributor that precedes its core places a greater load on STM because its role in the discourse is not apparent until the core has been comprehended. Furthermore, comprehension data from Stevenson (Stevenson & Urbanowicz 1995) can be interpreted as supporting the conclusion that core-contributor order is easier to comprehend.

For these reasons, one might expect that contributor-core ordering would occur infrequently, and that when it did occur, there would be a discourse cue on either the contributor or the core to help the reader interpret it correctly. Moser & Moore (in preparation) found, however, that although relations in contributor-core order are in fact likely to have a cue (55% of the time, as opposed to 37% of the time in core-contributor order), they are not at all rare, constituting 58% of the relations in Moser and Moore's corpus. The sample text shown above, for example, contains two relations in contributor-core order. Why should an ordering that is more difficult to comprehend occur so frequently?

It must be remembered that cores and contributors do not occur in isolation, but within a larger discourse structure, which has several levels of organization that constrain the ordering of sentences and clauses. The contributor in *A* for example, is related to its core *B* as a concession, and concessions are known to have a tendency to precede their main clauses (Mann & Thompson 1988; Noordman in press). As for the other contributor *C2* which precedes its core *C1*, in the context of clause *D* a core-contributor ordering would have introduced an ambiguity in the assignment of *D* – as a contributor to *B* or to *C2*. Another discourse-level influence on ordering could be the need to smoothly change the focus of attention without abrupt shifts in focus (McKeown 1985; Sidner 1981). So, for example, once part2 has been introduced in *B*, the speaker finishes talking about part2 in *C1* and *C2* before moving on to talk about part3 in *D*.

In addition, in both instances of contributor-core ordering in this segment of dialog, the contributor can reasonably be assumed to be *given* information (in the sense of hearer-old, or already part of the hearer's knowledge), and therefore required to be presented before the new information in the core (Clark & Haviland 1977). As Oberlander & Moore (1999) have noted, such information packaging constraints may be partly responsible for the tendency for contributors to precede cores in natural discourse contexts. Consideration of constraints on discourse structure and processing theories of discourse comprehension, then, does not lead to the conclusion that contributors should never precede cores in well-formed discourse, but rather that when they do the contributor should be marked with a cue to indicate its function in the information structure of the discourse.

In light of our goal of using corpus analyses to guide the construction of comprehensible texts, we asked the following question: Can a distributional analysis of naturally occurring discourse shed further light on the effects of cues on comprehension, apart from more theoretically-based processing considerations such as those discussed above? The first step in answering this question is to identify relevant patterns in the distributional analysis of the dialogue corpus.

Moser & Moore's (in preparation) analysis of core:contributor relations in their tagged corpus of tutorial dialogues revealed patterns of cue <u>occurrence</u> (whether or not to use a cue), cue <u>placement</u> (where to place a cue when one is used) and cue <u>selection</u> (which cue to use). Some of the observations that emerged from this analysis were:

- **Cue occurrence**: Cues are more likely to be used when a core follows a contributor than when the contributor follows a core (55% of the contributor-core relations were cued vs. 37% of the core-contributor relations).
- **Cue placement**: Cues tend to occur on the contributor more often than on the core (40% of cues occur on cores, 60% on contributors).
- **Cue placement**: Cues tend to occur with the second relatum more often than the first (27% of cues occur on the first relatum, 73% on the second).
- **Cue placement**: Cues are never placed on the core of a relation that is in core-contributor order.
- **Relata order**: Cores usually follow contributors (58% of the time overall).

Do theses patterns of cue placement represent useful guidelines for producing comprehensible text? Although this is the assumption that motivated the corpus study, we can not be certain in the absence of experimental tests of comprehension. An experiment examining both reading time and recognition data for cued and uncued relations was conducted to test this assumption. Participants in the experiment read a narrative text containing 20 sentences that expressed a core:contributor relation. We manipulated the order of

the core and contributor and the placement of discourse cues in the sentences. Reading times and a recognition memory test were used to assess the effects of order and cue placement on comprehension. The results are first compared to what would be predicted from the corpus study under the assumption that the forms produced most frequently by speakers should also be easiest to comprehend. Other possible influences on comprehension are then discussed, along with differences between narrative and expository text that may be relevant to the results of this study.

Method

Participants

A total of 95 University of Pittsburgh undergraduates participated in Experiment 1. All were students enrolled in Introductory Psychology and participated to fulfill a course requirement. Reading times were collected from all 95 participants, and 79 also completed a recognition memory test after reading the passages (16 did not have time to do the memory test or were not given the test due to technical problems).

Materials and Design

To examine the comprehensibility of discourse segments (core:contributor relations) as a function of cue occurrence and cue placement, a narrative text was constructed containing 20 sentences that each expressed a core:contributor relation. The text described an extended conversation between two characters concerning a variety of everyday topics familiar to college students. (Although tutorial dialogues are more similar to expository than narrative texts, we chose to use a narrative text for this experiment because it provided a natural context for presenting a number of distinct arguments about topics that would be easily understood by our undergraduate subjects.) Each item was expressed in a single sentence. In versions in which no cue was present a semicolon was used to connect the core and contributor. Although the core and contributor could have more naturally been written as two separate sentences, to do so would have confounded the presence vs. absence of discourse cues with whether the core and contributor occurred in the same vs. separate sentences.

The presence and placement of discourse cues were varied to create several versions of each of the 20 items in the passage. Our original intent was to create six versions of each item, crossing the factors of cue location (cue on core, cue on contributor, or no cue) and order of relata (core-contributor vs. contributor-core) in a 3 x 2 design. We discovered, however, that it was not possible to create acceptable versions of the items in one of the six resulting conditions. All attempts to create items in core-contributor order with a cue placed on the core resulted in (at best) very awkward sentences such as: "With the result that they almost always had a steady stream of stimulating conversation, they had just the right combination of shared interests and conflicting opinions." Note that this condition corresponds to the pattern that was never observed in the corpus. The following excerpt from the beginning of the experimental passage displays for one of the items the five remaining versions that were used in the experiment. In this example the core is "They almost always had a steady stream of stimulating conversation" and the contributor is "They had just the right combination of shared interests and conflicting opinions." Underlining indicates the presence and location of discourse cues.

> Jerry and Sheila were walking home from the theatre towards their apartment in Greenwich Village. They had been roommates throughout college and, although never romantically involved, they had become particularly close friends.

core-contributor: They almost always had a steady stream of stimulating conversation; they had just the right combination of shared interests and conflicting opinions.

<u>core</u>-contributor: *not used; never occurred in corpus*

core-<u>contributor</u>: They almost always had a steady stream of stimulating conversation, <u>because</u> they had just the right combination of shared interests and conflicting opinions.

contributor-core: They had just the right combination of shared interests and conflicting opinions; they almost always had a steady stream of stimulating conversation.

contributor-<u>core</u>: They had just the right combination of shared interests and conflicting opinions, <u>so</u> they almost always had a steady stream of stimulating conversation.

<u>contributor</u>-core: <u>Because</u> they had just the right combination of shared interests and conflicting opinions, they almost always had a steady stream of stimulating conversation.

The discourse cues used in this experiment were *because* and *since* to signal contributors, and *therefore* and *so* to mark cores. Only one of the two cues for a contributor and one of the cues signalling a core was used for each item – it was never the case that *because* was used for one version of an item and *since* was used with another version of the same item, for example. Thus cue selection was not included as a factor in the experiment. Moser and Moore's results suggest the existence of selectional restrictions influencing which cues are produced in which locations, however, and the effect of cue selection on comprehension will need to be addressed in future research (Moser & Moore 1995).

The five conditions in the reading-time study were counterbalanced across subjects and items in a Latin Square design, so that each subject saw 4 items in each of the five versions, and each item appeared an equal number of times in each version across subjects.

In the sentence recognition test that participants were given after reading the passage, an altered version of each item they had read in the passage was presented, so that the correct response to the experimental items in the recognition test was always "new." Each test sentence always had the order of the core and contributor clauses reversed from that which had been presented in the passage, so that an item that had been in core-contributor order in the passage would be presented in contributor-core order in the recognition test, and vice-versa. After reversing the order of the core and contributor, discourse cues were then used to manipulate the meaning of the test items, so that each test item had either the *same* meaning as the corresponding sentence from the passage, or had a *different* meaning resulting from a reversal of the roles of the core and contributor. Item 1's core, for example, was "They almost always had a steady stream of interesting conversation," which in the passage might have been cued as a core with "so." In the recognition test this clause could have been cued with "therefore," giving the sentence the *same* meaning it had in the passage, or cued (inappropriately) with "since," marking it as a contributor and therefore giving the sentence a *different* meaning than it had in the passage. To the extent that participants understood and remembered the nature of the core-contributor relationship in the original sentence and understood the relationship in the test sentence, they should have found test sentences with *different* meanings easier to reject as "new" compared to *same*-meaning test sentences. The difference between false alarm rates for the *same* vs. *different* meaning conditions, therefore, provided a measure of the effectiveness of cues in signalling core-contributor relationships.

All sentences in the recognition test were cued, and the cue always preceded the second clause of the sentence, with cue selection determining whether the meaning of the test sentence was the same as or different from the corresponding sentence in the passage. Thus, there were 10 versions of each item used in the recognition test (the 5 versions that appeared in the passage, crossed with whether the core:contributor relationship was the *same* or *different*). These 10 conditions were counterbalanced across subjects and items.

For each item, the cue word that had appeared in the passage was never used in the recognition test for that item. For item 1, for example, the cues "because" and "so" were used in the passage, but the items created for the recognition test used "since" and "therefore" as cues instead. This was done to keep subjects from using the cue words themselves as signals to respond "old" or "new" in the recognition test. The following are the test sentences used for item 1.

core-contributor order in passage; same meaning in recognition test:

They had just the right combination of shared interests and conflicting opinions, <u>therefore</u> they almost always had a steady stream of stimulating conversation.

core-contributor order in passage; different meaning in recognition test:

They had just the right combination of shared interests and conflicting opinions, <u>since</u> they almost always had a steady stream of stimulating conversation.

contributor-core order in passage; same meaning in recognition test:

They almost always had a steady stream of stimulating conversation, <u>since</u> they had just the right combination of shared interests and conflicting opinions.

contributor-core order in passage; different meaning in recognition test:

They almost always had a steady stream of stimulating conversation, <u>therefore</u> they had just the right combination of shared interests and conflicting opinions.

Another 20 sentences from the passage were included as "old" filler items for the recognition test, along with 20 "new" filler sentences that had not appeared in the passage. Each of the "new" fillers contradicted information that had been stated in the passage. Thus the recognition test contained 60 items – 20 old and 40 new – with half of the test sentences consistent with the meaning of the passage and half inconsistent.

Procedure

All materials were presented on a 15 inch computer monitor in text mode, and all responses were collected via the computer keyboard. Presentation of materials and collection of responses was controlled by the Cogsys system for experiments running on a 386 or Pentium PC (Ratcliff, Pino, & Burns 1986).

In the first stage of the experiment, the subject read the experimental text in a self-paced moving-window presentation. The subject first read instructions presented on the screen. The instructions included several example sentences presented with the self-paced moving-window technique to familiarize the subject with the procedure. After reading the instructions, the subject then pressed the spacebar to initiate presentation of the experimental passage. The first word of the passage then appeared on line 10 (about half way down the screen vertically) at the left edge of the screen. After reading that word, the subject then pressed the spacebar to go on to the next word. The screen was then cleared and the next word of the sentence was presented one space to the right of the end of the previous word. The position for the presentation of the next word was reset to the beginning of line 10 again at the beginning of each sentence in the passage, and whenever the next word of a sentence would extend past the right edge of the screen. Subjects were instructed to read at a comfortable pace, and were told that they would be tested on the content of the passage later in the experiment.

After the subject finished reading the passage, the instructions for the recognition test appeared on the screen. Subjects were instructed to decide whether each test sentence was one that had appeared in the passage or not and to press one of two keys labelled "Yes" and "No" (the . and /? keys respectively) to indicate their response using the first two fingers of their right hand. The instructions emphasised that they were to say "Yes" only if that exact sentence had appeared in the passage, and say "No" if it had been reworded or re-arranged in any way. Subjects were instructed to be as fast and accurate as possible. No feedback was given regarding the accuracy of subjects' responses during the recognition test.

After reading the instructions, the subject pressed the spacebar to initiate presentation of the first test sentence. The screen was then cleared for one second, after which the test sentence was presented in the middle of the screen (on lines 10 and following) until the subject pressed a key. When the subject pressed a key, the screen was cleared for one second and then the next test sentence appeared, and so on until the subject had responded to all 60 test sentences.

Results

Reading Times

Mean per-word reading times were calculated for each sentence in the passage by dividing the total reading time for the sentence by the number of words in the sentence (see Table 1). For the 20 experimental items, the discourse cues were excluded from these calculations so that only reading times for words that appeared in all versions of an item would be compared in the analyses. The resulting per-word reading times for the experimental items were entered into one-way analyses of variance (ANOVAs). One sentence (out of 1900 total experimental sentences) had a per-word reading time of over 1000 ms., which was truncated to 1000 ms. Six sentences had per-word reading times of less than 200 ms. and were discarded as errors. Version (core-contrib, core-contrib, contrib-core, contrib-core, or contrib-core, where underlining indicates the location of a discourse cue) was a within-subjects factor in the subjects (F_1) analysis and a within-items factor in the items (F_2) analysis. The main effect of version was reliable by subjects but not by items, probably because of low power: $F_1(4, 360) = 2.98$, $MSE = 678$, $p < .05$, power = .76; $F_2(4, 60) = 1.75$, $MSE = 233$, n.s., power = .10.

Paired comparisons of means revealed only the following significant ($p < .05$) differences: Versions core-contrib and core-contrib were read faster than version contrib-core, and version core-contrib was read faster than version contrib-core. The difference between versions contrib-core and contrib-core was statistically marginal ($p < .10$).

Did the patterns of cue placement and relation ordering in the corpus analysis predict the pattern of reading time data? The most straightforward way that corpus data might predict comprehension would be for constructions that are more frequent in the corpus to be read more quickly. If the corpus analysis were to predict comprehension in this straightforward sense, we could make several predictions based on the generalizations observed by Moser and Moore (in preparation).

First, because more relations occurred in contributor-core than core-contributor order in the corpus, we could predict that reading times should be faster for versions in contributor-core order (contrib-core, contrib-core, and contrib-core) than in core-contributor order (core-contrib and core-contrib). In fact, the opposite was found. The mean reading time per word for core-contributor ordered items was 370 ms., vs. 377 ms. for contributor-core order. A planned contrast showed this difference to be reliable by subjects but not by items, $F_1(1, 90) = 7.32$, $MSE = 267$, $p < .01$; $F_2(1, 15) = 1.6$, $MSE = 172$, n.s. This finding was not surprising given that there are reasons to expect contributor-core ordering to place greater processing loads on comprehenders, and in fact, this result is exactly what Moser and Moore's account would predict.

A second prediction that could be made from the corpus analysis is that, reflecting the fact that cues occur more often on the second relatum, reading times should be faster when the second relatum is cued. A planned contrast did confirm this prediction. Reading times were faster when a cue occurred on the second relatum (core-contrib and contrib-core, 371 ms. per word) than when a cue occurred on the first relatum (contrib-core, 377 ms. per word), $F_1(1, 90) = 2.48$, $MSE = 693$, $p < .12$, n.s.; $F_2(1, 15) = 8.38$, $MSE = 81.3$, $p < .05$.

A third prediction that could be made from the corpus frequencies that items with cues on the contributor should be read more quickly than items with cues on the core. This prediction was not supported. A contrast comparing items with cued cores (contrib-core, 374 ms.) to items with cued contributors (core-contrib and contrib-core) was not significant, Fs < 1.

A fourth prediction from the corpus frequencies could be that cues should be of more benefit to comprehension when the core follows the contributor than when the core precedes the contributor. This prediction follows from the observation that cues occur more often on relations in contributor-core than core-contributor order. A contrast testing this prediction (that the difference between core-contrib and core-contrib is less than the difference between contrib-core and (contrib-core or contrib-core) was not significant, Fs < 1.

A fifth planned contrast comparing cued (core-contrib, contrib-core, and contrib-core, 373 ms.) to non-cued items (core-contrib and contrib-core, 377 ms.) was statistically marginal, $F_1(1, 90) = 2.44$, $MSE = 283$, $p < .13$; $F_2(1, 15) = 1.87$, $MSE = 90.1$, $p < .20$. This contrast was not directly related to any prediction from relative frequencies in the corpus analysis, but instead tested the general hypothesis that discourse cues made the text segments easier to comprehend.

Version	core-contrib	core-contrib	contrib-core	contrib-core	contrib-core
Mean	373	368	381	374	377
Std.Error	8.7	7.6	9.2	8.7	9.4

Table 1: Reading time per word (excluding cue words) in milliseconds. Underlining indicates the location of a discourse cue.

Version	core-contrib	core-contrib	contrib-core	contrib-core	contrib-core
Same meaning	.58	.63	.69	.73	.73
Different meaning	.41	.32	.34	.30	.32

Table 2: Proportion false alarms in sentence recognition test. Chance = .50

Sentence Recognition

The proportion of "Yes" responses (false alarms) to the experimental test sentences was calculated for each subject (F_1 analysis) and each item (F_2 analysis) and these proportions were entered in two-way ANOVAs. Version (the version in which the sentence had appeared in the passage) and meaning (whether the test sentence had the *same* meaning as the sentence that had appeared in the passage or had the core:contributor relationship reversed and thus a *different* meaning) were both within-subjects and within-items factors in the analyses. The proportions of false alarms for the 10 conditions in the recognition test are presented in Table 2. The overall false alarm rates were .505 for the experimental items and .103 for the new filler sentences (chance = .50). The hit rate for the old filler sentences was .779.

The data in Table 2 suggest that subjects responded primarily on the basis of the experimental test sentences' meanings, rather than on the basis of the exact wording of the sentences. The ANOVAs supported this interpretation. Collapsing across meaning, the mean false alarm rates for all 5 versions were near chance (.50) and the main effect of version did not approach significance, $F_1(4, 312) < 1$, $MSE = .087$; $F_2(4, 76) < 1$, $MSE = .035$. The main effect of meaning was reliable, with twice as many false alarms when the meaning of the test sentence was the same as the sentence in the passage compared to when the meaning was different, .67 vs. .33 respectively; $F_1(1, 78) = 166.6$, $MSE = .135$, $p < .001$; $F_2(1, 19) = 96.1$, $MSE = .06$, $p < .001$.

It is assumed that the difference in false alarm rates between the same-meaning and different-meaning conditions reflects the degree to which subjects remembered the nature of the relationship between the core and contributor in the passage. This assumption seems reasonable given that the different-meaning condition was created by changing the cue used in the same-meaning condition in a way that reversed the core:contributor relationship. Given this assumption, the version-by-meaning interaction provided the critical tests of hypotheses about which cue placements led to better memory for the meanings of the experimental items. The interaction was reliable by subjects and items, $F_1(4, 312) = 4.28$, $MSE = .099$, $p < .01$, power = .92; $F_2(4, 76) = 3.25$, $MSE = .027$, $p < .05$, power = .80.

In order to directly compare the recognition data to the pattern of results observed in the reading times, we analysed the differences between same- and different-meaning false alarm rates for each of the 5 versions in one-way ANOVAs. The main effect of version in these analyses of the difference scores is functionally equivalent to the version-by-meaning interaction in the original analyses. Re-casting the analyses in terms of difference scores, however, makes it easier to relate the recognition data to the reading time results. Figure 1 plots the false alarm difference scores and the mean reading times for the 5 versions together. The scale for the false alarm differences is inverted to facilitate visual comparison to the pattern of the reading time data.

The most apparent difference in Figure 1 is the reversal of the effect of relata order between the reading time and recognition data. In the reading time data, items in core-contributor order were read more quickly. In the recognition test, however, items in contributor-core order were better remembered: difference scores of .24 vs. .40, $F_1(1, 78) = 9.47$, $MSE = .103$, $p < .01$; $F_2(1, 19) = 8.50$, $MSE = .024$, $p < .01$. Of the other contrasts performed on the reading time data, only the comparison of cued vs. uncued items was reliable in the recognition analyses: Cued items were better remembered than uncued (difference scores of .39 and .26 respectively), $F_1(1, 78) = 7.01$, $MSE = .089$, $p < .01$; $F_2(1, 19) = 7.72$, $MSE = .016$, $p < .05$.

Discussion

The primary question this research addressed was whether patterns of discourse cue production observed in a corpus of tutorial dialogues predict the comprehensibility of text segments with and without cues. In the limiting case of a construction that never occurred in the corpus, the answer was clearly "yes." Relations in core-contributor order never occurred with a cue on the core anywhere in the corpus, and when we attempted to create such constructions for use in an experiment the resulting sentences were clearly awkward and difficult

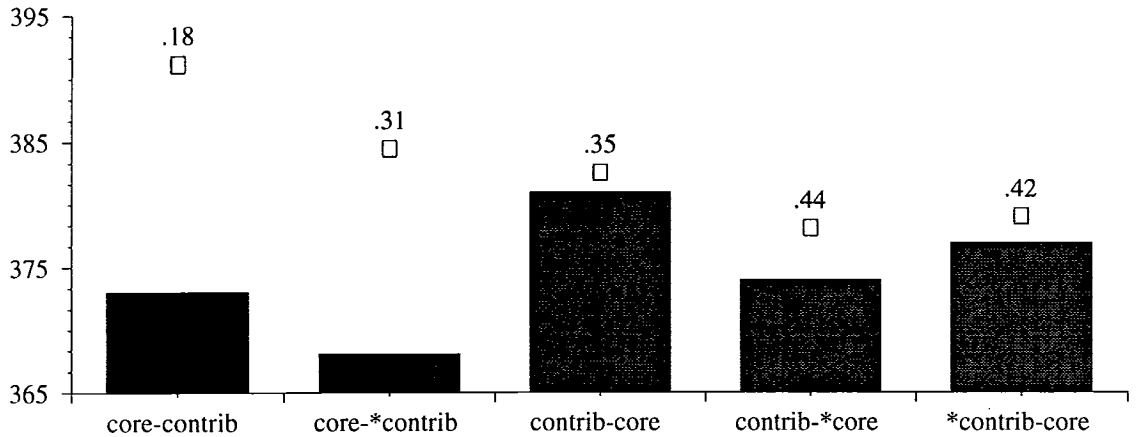

Figure 1: Reading time (bars) and recognition data (boxes). Reading times are milliseconds per word. Recognition data are same-meaning false alarms minus different-meaning false alarms, plotted on an inverted scale. This difference score reflects how sensitive participants were to a reversal of the core:contributor relationship between study (reading the passage) and test. The version presented in the passage is labelled on the x-axis, with a * indicating the location of a cue.

to understand – so much so that these sentences could not be used in the study.

The reading time and recognition data from the experiment addressed the question of whether more subtle patterns of cue production could predict comprehension. In particular, three generalizations about cue placement drawn from the corpus analysis were used to try to predict comprehension: cues occur more often on contributors, cues occur more often on the second relatum, and relations in contributor-core order are more likely to contain a cue. Of these three generalizations, only the placement of a cue on the second relatum appeared to affect comprehension in the manner directly predicted by frequency of occurrence in the corpus.

A very specific prediction concerning the contrib-core and contrib-core conditions could be motivated by both production frequencies in the corpus and a plausible account of differences in processing mechanisms engaged by the two versions of the sentence. It could be argued that contrib-core sentences, in addition to being less frequent than contrib-core in the corpus (40% of contributor-core ordered relations that were cued had a cue on the contributor, 60% on the core), impose an additional processing load compared to contrib-core in that the initial cue on the contributor creates an expectation that information later in the text will have to be integrated with the first clause. The additional processing load associated with maintaining the contributor in memory while anticipating the core could slow processing of the text. Both the production frequencies and these processing considerations would tend to predict that contrib-core sentences would be read faster than contrib-core sentences. While the trend in the data was consistent with this interpretation, the difference was not statistically significant.

Relata order (core-contributor vs. contributor-core) did not affect comprehension in the way that would directly reflect production frequencies. In the reading time data in particular, the less frequent ordering of relata produced the fastest reading times. This pattern of results – the less frequently occurring ordering being easier to comprehend – was predicted by Moser and Moore (in preparation), however. They hypothesised that in core-contributor order (the less frequent ordering) the core provides a context for interpreting the contributor and therefore makes comprehension easier. They further reasoned that the relative difficulty of contributor-core ordered relations could explain why speakers are more likely to use a cue for relations in contributor-core order. Our finding that contributor-core order is indeed more difficult to understand is also consistent with earlier findings by Stevenson and Urbanowicz (1995), in which sentences that we would identify as being in core-contributor order were read more quickly than similar sentences in contributor-core order. We did not, however, find any evidence to support Moser and Moore's hypothesis that cues should be of more benefit to comprehension for relations in contributor-core order than those in core-contributor order.

Oberlander and Moore (1999) suggest that the reason the less easily understood contributor-core order occurs so frequently may lie in the pragmatic constraints on the ordering of given and new information in a discourse. Information that is "given," or previously introduced in the discourse, generally must precede new information. Thus it is because contributors are frequently the "given" information in the relation that the contributor-core order is so frequent, on their account. An analysis of the given-new relationships in the materials used in this study did not, unfortunately, shed any further light on this hypothesis. None of the core or

contributors in the experimental items were discourse-old information, and very few (at most 5) could reasonably have been assumed to be hearer-old information in the context of the story. Therefore we were unable to examine the effects of discourse cues and ordering in the presence of "given" information, in either sense of given-ness that we considered.

In the sentence recognition data, the effect of relata order was reversed from that found in the reading time data. The more frequent contributor-core ordering was better remembered although it had been read less quickly. If contributor-core relations are indeed more difficult to comprehend, it could be that readers were more likely to remember them because of the extra processing that they carried out in order to comprehend them. Such an explanation can not completely account for the recognition results, however, since the pattern of results within each ordering of relata was quite similar for the two measures – there was not a simple negative correlation between reading time and subsequent memory for meaning. At least some patterns were consistent across the two measures, such as the trends for faster reading times and better memory for cued than uncued relations.

This research has demonstrated small but reliable effects on comprehension of narrative text that can be partially related to patterns of cue production in a corpus of tutorial dialogues. It has also illustrated how mismatches between production frequencies and comprehension can point to productive candidates for further analysis and research. Although this research yielded only small effects of cues on comprehension, recent research by Fox Tree and Schrock (1999) found significant facilitation of comprehension resulting from the presence of "oh," a discourse cue that serves very different functions from the cues examined here. Cueing of cores and contributors may be a function that is particularly important in expository text, rather than narrative text. Future research will attempt to extend these findings to the comprehension of expository text, which is more similar to tutorial dialogues. Because accurate identification of core:contributor relationships would be expected to be more crucial for understanding expository texts, which tend to take the form of arguments rather than the form of stories, we expect to find more robust effects of discourse cues on comprehension. In addition, future studies will need to examine not only cue placement, but cue selection as well.

Acknowledgments

The research presented in this paper was supported by Grant #N00 014-91-J-1694 from the Office of Naval Research, Cognitive and Neural Sciences Division.

References

Anderson, J. R.; Corbett, A. T.; Koedinger, K. R.; and Pelletier, R. 1995. Cognitive tutors: Lessons learned. *Journal of the Learning Sciences* 4(2):167–207.

Bloom, B. S. 1984. The 2 Sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher* 13:4–16.

Clark, H. H., and Haviland, S. 1977. Comprehension and the given-new contract. In Freedle, R., ed., *Discourse production and comprehension.* Hillsdale, New Jersey: Lawrence Erlbaum Associates. 1–40.

Fox Tree, J. E., and Schrock, J. C. 1999. Discourse markers in spontaneous speech: Oh what a difference an oh makes.

Kintsch, W. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review* 95:163–182.

Mann, W. C., and Thompson, S. A. 1988. Rhetorical Structure Theory: Towards a functional theory of text organization. *TEXT* 8(3):243–281.

McKeown, K. R. 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text.* Cambridge, England: Cambridge University Press.

Moser, M., and Moore, J. D. 1995. Investigating cue selection and placement in tutorial discourse. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 130–135.

Moser, M., and Moore, J. D. in preparation. On the correlation of cues with discourse structure: Results from a corpus study. Submitted for publication.

Moser, M.; Moore, J. D.; and Glendening, E. 1996. Instructions for Coding Explanations: Identifying Segments, Relations and Minimal Units. Technical Report 96-17, University of Pittsburgh, Department of Computer Science.

Noordman, L. G. M. in press. On the production of causal-contrastive although-sentences in context. In Saunders, T.; Schilperenol, J.; and Spooren, W., eds., *Text and presentation: Linguistic and psycholinguistic aspects.*

Oberlander, J., and Moore, J. D. 1999. Cue phrases in discourse: further evidence for the core:contributor distinction. Paper presented at the Levels of Representation in Discourse Workshop, University of Edinburgh, `http://www.hcrc.ed.ac.uk/~lorid99`.

Ratcliff, R.; Pino, C.; and Burns, W. T. 1986. An inexpensive real-time microcomputer-based cognitive laboratory system. *Behavior Research Methods, Instruments, & Computers* 18:214–221.

Sidner, C. L. 1981. Focusing for interpretation of pronouns. *American Journal of Computational Linguistics* 7(4):217–231.

Stevenson, R. J., and Urbanowicz, A. J. 1995. Structural focusing, thematic role focusing and the comprehension of prounouns. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, 328–332. Hillsdale, New Jersey: Lawrence Erlbaum Associates.