Understanding Classroom Lectures

ADRIENNE LEHRER
University of Arizona

After hearing three lectures as part of a regular university course, students were asked to summarize the lectures. The lectures were graded for accuracy and completeness. Each one was taped, transcribed, and coded, using a coding system based on work by van Dijk and Kintsch (1983) and Meyers (1975a, 1975b). The hypotheses to be tested were: (1) The percentage, but not the number of propositions would correlate positively with the accuracy scores in the summaries. (2) Completeness scores would correlate positively with the total number of propositions, but the number and percentage of propositions coded as macropropositions would correlate even more highly. (3) Signaled propositions would be better remembered than nonsignaled ones. Hypothesis 1 was confirmed in the summaries of all three lectures; Hypotheses 2 and 3 were confirmed only in two of the lectures. Analyses of the errors, lexical substitutions, discrepancies in the grades assigned by the two coders, and differences in the students' scores for the different lectures revealed interesting (and in some cases unexpected) results concerning lexical–semantic networks and grading reliability. Finally, implications for classroom lecturing are discussed.

A common occurrence in a university setting is that of an instructor giving some lectures, subsequently testing the students on material in the lectures, often by requiring the students to write essays, and finally grading the students’ papers. Sometimes the grading is done by the instructor and sometimes by someone else, for example, by a teaching assistant. In any case, there are many steps to the process: What do the students comprehend, remember, consider important, and report? What is the evaluator (the individual assigning grades) responding to? When the evaluator is someone other than the instructor, there is an additional step in the process, because the grader has to comprehend, remember, and assess the relevance of the lectures and then access the performance of the students’ essays.

Studying the cognitive processes involved in such common experiences is extremely difficult because of the large number of variables, each of which is

This research was supported in part by a grant from the Social and Behavioral Sciences Institute of the University of Arizona. I wish to thank Thomas Bohman for many of the statistical analyses and for comments and suggestions on the first draft, and Lee Fulmer for transcribing and coding. I am also indebted to Duane Roen, Diane Meador, Sue Lorenson, Keith Lehrer, and members of the University of Arizona Cognitive Science audience, the Department of Linguistics at Monash University, and the Department of Linguistics at Macquarie University, who provided many valuable comments and suggestions.

Correspondence and requests for reprints should be sent to Adrienne Lehrer, Department of Linguistics, University of Arizona, Douglas 200 East, Tucson, AZ 85721.
itself hard to study, for example, memory, the interest and motivation of the
students, the background knowledge and beliefs that the students bring to the
lectures, the content and complexity of the lectures, the rhetorical skills of
the lecturer, and the knowledge, beliefs, and standards of the evaluators.

Although these problems are formidable, I believe that it is possible, indeed
necessary, to study the problems of lecture comprehension in naturalistic set-
tings, by building on the extensive research that has been done in carefully
controlled experimental settings—research on remembering prose (Alba &
Hasher, 1983; R.E. Johnson, 1970; Meyer, 1975a), text and story comprehension
(Black, 1984; Reyna & Brainerd, 1991; Reyna & Kiernan, 1994; Trabasso & van
den Broek, 1985), text and discourse analysis (Roy, 1989; van Dijk, 1980; van
Dijk & Kintsch, 1983), and protocol analysis (N.S. Johnson, 1985). Many of the
previous studies on text comprehension (Kintsch & Bates, 1977) have tested for
recognition rather than recall, but Fletcher (1992) and Stevenson (1988) have
criticized the validity of these experiments. Stevenson found that subjects who
had not heard a lecture were as good at guessing which of two sentences occurred
in that lecture as were the subjects who actually heard the lecture and remem-
bered which of the two sentences was presented. Other researchers have used
very short texts, for example, a story of four lines (Black, 1984; Reyna &
Kiernan, 1992) or medium-length written texts (Roen, 1984; Roen & Piche,
1984).

The focus of this article will be aspects of classroom lectures that contribute to
what students comprehend and remember about a lecture, and how students’
explores only some of the variables that must be controlled for—the ultimate goal is to
summaries about lectures are evaluated. This study is a preliminary one and studies of this sort
contribute to a theory of language understanding. In addition, studies of this sort
should have implications for classroom teaching and learning that can point the
way to better use of lectures for teaching.

SUMMARIZING LECTURES AND EVALUATION
SUMMARIES: METHOD

During a semester in which I taught an introductory linguistics course, three of
the lectures for the course were taped: The first one was on parts of speech, the
second was on norms of correctness, and the third was on historical linguistics
and language change. After about 40 min, I ended my lecture and asked the
students to summarize it. Students had not been told earlier that they would be
asked to write a summary, and the instructions were fairly vague, thereby allow-
ing them considerable choice in what and how much to write. The instructions
after the first lecture were as follows:

Now I'd like to do something for the rest of the hour which is a little unusual. I
would like you to write down what you think I said. I'm not going to grade these
papers]; it isn’t a test in the sense of me testing you. In some sense I’m testing myself. I’ve found in the past that there are certain points that many people have a lot of trouble getting, and I’d like to find out now if this is one of them. So take a piece of paper and put your name on it and just give a summary of what I said.

The students’ summaries were read by two graduate students in linguistics who assigned two grades to each student protocol: a grade for accuracy and a separate grade for completeness. One of the evaluators (M.L.) was the current teaching assistant for the course and was present at all the lectures. He had made his own notes during the lectures and also had access to the notes of a professional note-taker for the course. The other evaluator (S.F.) had been a teaching assistant for the same course the year before. She transcribed the tapes for the three lectures in this study, and she also served as a coder in the analyses, so she was very familiar with these texts.

The two evaluators assigned letter grades, including pluses and minuses, and these grades were converted to nine numerical categories: \( A = 9; A-/B+ = 8; B = 7; B-/C+ = 6; C = 5; C-/D+ = 4; D = 3; D-/E+ = 2; E = 1. \) On each of the six correlations, \( t \) tests for dependent samples were performed, and in all except one case (completeness scores for the third lecture), all differences in grading between the two graders were nonsignificant at the .05 level. Although there were considerable differences on some of the protocols, the average of the two grades seemed to be a reliable guide, and averages were used for the calculations that follow. In addition, the protocols which received rather different evaluations (a difference of 3 or more points) were studied carefully to see why the scores were so different. That is, the difference in scores between the evaluators was itself a source of data. The correlations of grades assigned by the two evaluators are discussed in the section on reliability of the evaluators.

**Coding the Lectures**

The lectures were taped, the tapes transcribed, and the transcriptions coded. The coding system used was based on the text analyses found in Meyer (1975a) and van Dijk and Kintsch (1983). The unit of analysis that we used—the idea or proposition—was basically the sentence, although compound and complex sentences containing separate ideas were counted as separate propositions.

Two important features of the coding system are that (1) the text is analyzed according to *levels* and (2) specific signaling devices are noted. The notion of *levels*, as discussed in Meyer (1975a), van Dijk (1980), and van Dijk and Kintsch (1983) is that propositions in texts may differ according to their generality. *Macrostructures* are propositions that relate proposition sequences and provide global meanings for more specific or local sequences in a text. Macrostructures can themselves be grouped into more abstract or general propositions. In coding the lectures, four levels seemed sufficient: \( L_1 \) was given to the most general propositions and \( L_4 \) to the most specific. In addition, \( L_0 \) was used for statements
that merely introduced topics but which contained no information on a topic, such as, "I will discuss parts of speech" and for questions that were used to introduce content, such as "Who decides what is correct?" (Such sentences were counted as markers, however, as will be described.) This method of coding worked well for the first and third lectures but was somewhat problematic for the second, and revisions were made, as described below.

Alternative coding systems have been proposed, such as that of analyzing causal relationships (van den Broek & Trabasso, 1986), but many studies show that subjects' summaries and recall reflect both a statement's hierarchical position and its causal links. Varnhagen (1991), who investigated three analytical systems for recall of prose texts, including those of Meyer (1975b) and Trabasso and van den Broek (1985), found that both were equally sensitive to differences across age groups and recall differences. However, as in much research of this type, different analysis systems are better for recall of some texts than for others.

Signaling devices are words or phrases in the text that call attention to some aspect of the text. Various investigators have interpreted this concept differently, so that in some cases simple adverbials or conjunctions have been counted as signaling devices. However, in the three lectures the items counted as signals were fairly strong and explicit markers such as "And let me try to explain why," "what this shows us is," or "one of the most important discoveries is." Also counted as signaling devices were phrases and clauses that introduced new topics, for example, "okay, now let's turn to adjectives." As noted above, many of the \( L_0 \) clauses served as markers, as in "now I'd like to give you a very brief overview of the history of English and why English is the way it is." Van den Broek and Trabasso (1986) observe that very general statements—those at the highest level of superordination—are often omitted from summaries, especially if they have few connections to the more specific details of the text as a whole.

Each lecture was independently coded by two coders (SF and myself), after which we met to compare the coding, discuss any differences, and agree on the most reasonable choice. Tables 1 and 2 summarize the salient statistics on the lectures: the number of propositions, the number of markers, and the levels of generality of the propositions.

Some of the previous work on the recall of prose took the percentage of (correctly) remembered sentences as the measure for performance. Several of these studies (Meyer, 1975a; Roen & Piche, 1984) showed that whatever was

<table>
<thead>
<tr>
<th>Organization of the Lectures</th>
<th>Propositions</th>
<th>Markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parts of Speech</td>
<td>139</td>
<td>29</td>
</tr>
<tr>
<td>Norms of Correctness</td>
<td>270</td>
<td>28</td>
</tr>
<tr>
<td>Language Change</td>
<td>203</td>
<td>11</td>
</tr>
</tbody>
</table>
recalled tended to be correct, so that the percentage of material recalled was a
reliable indicator of performance. However, the summary of lectures in this
study contained false propositions as well as correct ones, and therefore, it
seemed reasonable to separate accuracy and completeness.

The following hypotheses were to be tested:

*Hypothesis 1:* Although the number of propositions per se would not correlate
with accuracy scores, the percentage of correct propositions would correlate pos-
itively.

*Hypothesis 2:* Completeness scores would correlate positively with the total num-
ber of propositions, but they would correlate even more highly with number
(and/or) percentage of propositions at $L_1$.

*Hypothesis 3:* Signaled propositions would be better remembered than nonsig-
naled ones.

**Coding the Student’s Summaries**
The students’ protocols were first divided into propositions, using the same
criteria that were used in the analysis of the lectures. Each proposition was coded
according to whether it occurred in the lecture, whether it was correct, and what
level it occurred at. Propositions corresponding to signaled sentences were also
noted, along with their levels. Propositions not corresponding to any in the
lecture were coded as

*Follows:* not said, but a correct inference from what was said. This category also
covers correct examples offered by the student but not in the lecture.

*Vague:* the proposition was so vague the coders could not decide if it was com-
pletely accurate or merely on the right track.

*Not in:* the proposition was not in the lecture and was not a reasonable inference.

Accuracy was coded in the following way:

*True and accurate:* this included propositions that were marked as “follows” as
well as those corresponding to propositions in the lectures.
Partial: the proposition was partially correct.

Wrong polarity: the proposition was false because of an incorrectly placed or missing negative. That is, “X” was reported as “not X” or vice versa.

Wrong implication: the proposition was partially correct but the inference was wrong or the proposition was wrong for that context.

False: for other cases of incorrect propositions.

In addition, specific words were marked in the protocols when a proposition was not completely acceptable and where the problem could be localized in a particular lexical item. These were marked as

Close: a semantically close but inappropriate word was used.

Context: the word had been used in the lecture, but its use was inappropriate in the protocol.

Full credit was given to propositions coded as “accurate” and “follows”; half credit was given to those coded as “vague,” “partial,” and “close”; no credit was given to those coded as “false” or “wrong polarity.” Propositions marked “awkward” or “context” were noted, but no credit was taken off. However, these categories may have played a role in the qualitative evaluation carried out by the graders.

Two coders (S.F. and myself) independently coded the first 20% to 30% of each lecture. Our independent scoring produced the following agreements: breaking the text into proposition, 93% agreement; level or height of information, 76%; and other (accurate, follows, vague, etc.), 90%. We compared our scoring and discussed the differences. When we were satisfied that we agreed sufficiently, coding was completed by S.F.

RESULTS

Lecture 1: Parts of Speech
Table 3 summarizes the quantitative results of the Parts of Speech lecture.

<table>
<thead>
<tr>
<th></th>
<th>Total Propositions</th>
<th>Total Correct</th>
<th>% Correct</th>
<th>Total at $L_1$</th>
<th>% at $L_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>.20</td>
<td>.38*</td>
<td>.71***</td>
<td>.41**</td>
<td>.40**</td>
</tr>
<tr>
<td>Completeness</td>
<td>.71***</td>
<td>.80***</td>
<td>.56***</td>
<td>.59***</td>
<td>.24</td>
</tr>
<tr>
<td>Accuracy and</td>
<td>.51**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completeness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < .05.  **p < .01.  ***p < .001.
Accuracy. Hypothesis 1, that accuracy scores would correlate with the number and percentage of correct propositions but not with the number of total propositions, was supported. The Pearson correlation between the number of correct propositions was .38 ($p < .05$, two-tailed) and the percentage of correct propositions was .71 ($p < .001$). The correlation between accuracy scores and the total number of propositions was not significant ($r = .20$), as predicted. An examination of individual protocols showed that all of the protocols with 100% accuracy received (averaged) grades of 9 (A) or 8 (A−/B+), with one exception.\(^1\) Therefore, we see that Hypothesis 1 was confirmed by these data.

Completeness. Significant correlations were found between completeness scores and the total number of propositions ($r = .71$, $p < .001$), the total number of correct propositions, which was even higher ($r = .80$), and the percentage of correct propositions ($r = .56$, $p < .001$). The correlation between completeness and the number of propositions at $L_1$ was also significant ($r = .59$, $p < .001$). Therefore, Hypothesis 2 was confirmed.

However, some minimum number of propositions seemed necessary for a protocol to receive a high completeness score. For example, one protocol consisted of only two sentences, both at $L_1$. Although I considered this summary to capture the essence of the lecture succinctly, the completeness scores were only 2 (D−) and 6 (B−/C+). The large discrepancy between the two evaluators probably reflects a difference in attitude toward such extreme economy. What this shows is that to achieve a high completeness score, it is not sufficient to summarize the lecture succinctly; it is necessary to elaborate and add examples.

The protocols that received the highest and the lowest evaluations on completeness were studied more carefully to see what specific features they contained. The worst protocols had a small number of propositions (3 to 5), some of which were incorrect. But more telling, these protocols had many propositions coded as “vague” and a high proportion of $L_0$ scores. By contrast, the protocols rated as the best contained between 7 and 18 propositions, all of which or almost all of which were correct, and which contained either a large proportion of propositions at $L_1$ or contained a large absolute number of propositions at any level. For example, two protocols receiving completeness scores of 8 or 9 by the evaluators contained few $L_1$ propositions (1/18 and 2/12), but compensated with the large total number of propositions. It is also noteworthy that the items mentioned in the summaries contained material spread throughout the lecture.

The correlation between accuracy and completeness scores was .52 ($p < .005$), a correlation which is not surprising, because presumably only correct propositions would contribute to a completeness score. In other words, we would

\(^1\)The one exception involved a proposition with a not omitted. Since everything else in the summary was correct and intelligent, the evaluators may have missed the omission or may have noticed it but judged it to be an oversight. However, in scoring the protocol, it was scored as “wrong polarity.”
not expect an evaluator to give a high completeness score to a protocol that contained a large number of incorrect propositions.

**Markers.** Meyer (1975a) found that signaled content was better remembered than nonsignaled content, especially when the propositions were high in the hierarchy.

The percentage of marked propositions, that is, those signaled by an expression such as "what is important" or "let me tell you why" or a preparatory rhetorical question like "Do you know why?" in the lecture was 21% (29/139). The percentage of the propositions from the student summaries of such signaled sentences was 23% (40/169). A goodness-of-fit \( \chi^2 \) test was used to compare the number of propositions at each level in the lecture and in the protocols. Figures are given in Table 4; \( \chi^2 = 45.26, p < .01 \). What we see is that \( L_1 \) propositions that are signaled tend to be well remembered.

**Lecture 2: Norms of Correctness**

**First Calculations.** Pearson correlations were carried out between the accuracy scores, completeness scores, numbers of propositions, number of correct propositions, percentage of correct propositions, and the number and percentage of propositions at \( L_1 \). Table 5 summarizes the results.

**Accuracy.** The correlations between the accuracy score, the number of sentences (\( r = .38, p < .01 \)), accuracy and the number of correct propositions (\( r = .56, p < .001 \)), and accuracy and the percentage of correct propositions (\( r = .46, p < .01 \)) were lower than in the Parts of Speech lecture. Three of the Norms of Correctness protocols that showed a great discrepancy between the score and percent correct were examined. In one case the accuracy score was high although the percentage of correct propositions was lower than it should have been. It

<table>
<thead>
<tr>
<th>TABLE 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interaction of Signaling and Level of Propositions</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>( L_0 )</td>
</tr>
<tr>
<td>( L_1 )</td>
</tr>
<tr>
<td>( L_2 )</td>
</tr>
<tr>
<td>( L_3 )</td>
</tr>
<tr>
<td>( L_4 )</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

*Note. \( \chi^2(3) = 45.26, p < .001 \). Because the values for \( L_4 \) are 0, \( L_4 \) was omitted from the computation.*
seems that a superficial reading along with a generous interpretation might win a high score. However, a more careful reading revealed difficulties with the summary. In the other two cases, the percentage of accurate propositions was 100%, although the average of the evaluators' scores was 6.5. In one of the cases, the summary consisted of personal statements about what the student liked about the lecture and agreed with, and the low accuracy score may have been a judgment that the protocol was not really a proper summary of the lecture. In the third case, only a few items were mentioned, but this fact should have been reflected only in completeness scores.

The part of Hypothesis 1 predicting a positive correlation between accuracy and percentage correct was confirmed, but in addition, the positive correlation was found between the number of propositions as well.

**Completeness.** Completeness scores were correlated with the total number of propositions \( r = .56, p < .002, \) two-tailed and the number of correct propositions \( r = .68, p < .0005 \); the number of propositions at \( L_1 \) \( r = .41, p < .01 \) and the percentage of propositions at \( L_1 \) \( r = .20, \) n.s.). The correlation between accuracy and completeness scores yielded \( r = .58, p < .001 \).

Many of the protocols that received high completeness scores from the two evaluators contained relatively few propositions at \( L_1 \). Moreover, they frequently contained propositions that the lecturer considered important but which were coded as \( L_2 \). Therefore, I reanalyzed Lecture 2 carefully, trying to explain the discrepancy. One section was coded in the following way:

1. "There is an important distinction between covert and overt values." This was coded as \( L_1 \).
2. "Covert values are..." (definition given) \( (L_2) \)
3. "Overt values are..." (definition given) \( (L_2) \)

It seemed that the statements in (2) and (3) were more important than in (1), although (1) is more general and would be at a higher level. Therefore, I recoded
this section and one other like this, assigning sentence (1) $L_0$ and (2) and (3) $L_1$. One of the problems of simply counting the number of propositions at a particular level, especially when working with spontaneous speech, is that there is no guarantee that all the propositions at the highest level will be equivalent. If I had simply omitted the statement in (1) and had begun by defining overt and covert values, then sentences (2) and (3) would presumably have been coded as $L_1$. This modified coding utilizes a method similar to that used by Johnson (1970), in which sentences in texts were rated as more or less important. Alternative analysis systems, such as that of Trabasso and van den Broek (1985) might also be appropriate. Revised correlations are also given in Table 5.

The correlations based on revised coding are as follows: completeness score and number of propositions at $L_1$ yielded $r = .61, p < .001$; completeness score and number of propositions at $L_1$ yielded $r = .13$, n.s.

Hypothesis 2 was confirmed, but not exactly as stated. Completeness was highly correlated with both the number of propositions and the number of propositions at $L_1$, but the correlation was higher between the total number of propositions than the number of $L_1$ propositions, the opposite of what was predicted and the opposite of what was found in Lecture 1.

Inspection of the protocols receiving the highest and lowest scores for accuracy and completeness showed patterns like those in the parts-of-speech summaries. Those receiving high scores on accuracy had all or almost all of the propositions coded as "accurate" or "follows," and those with the highest scores for completeness either had a large number of propositions (at least 10) or if fewer than 10, they had a high number and/or percentage of propositions at $L_1$.

**Markers.** The propositions appearing in the protocols which corresponded to signaled sentences in the lecture were noted and tabulated. Only 21 such propositions were found, and none was in propositions whose level was recoded. A $\chi^2$ calculation was performed as described, where $\chi^2(4) = 9.61, p < .05$. Results appear in Table 6. Hypothesis 3, that signaled sentences at $L_1$ would be better remembered than nonsignaled ones, was confirmed.

### Lecture 3: Historical Linguistics

Of the 31 protocols, two were eliminated, one of which consisted of two sentences: the first correctly identified the topic of the lecture as historical change, and the second stated that the student had not been able to understand the lecture. In the second protocol, one evaluator did not assign any accuracy score, and the differences between the two scores for completeness were greater than for any other protocol in the three sets, with one evaluator assigning 2 (D−) and the other 8 (A+/B−). In this protocol, the student listed some of the topics covered, but said very little about any of them, such as, "Language families were discussed along with Grimm’s law." Table 7 summarizes the resulting correlations.
TABLE 6
Interaction of Signaling and Level of Propositions

<table>
<thead>
<tr>
<th>Norms of Correctness</th>
<th>No. in Lecture</th>
<th>No. in Protocols</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L_0 )</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>( L_1 )</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>( L_2 )</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>( L_3 )</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>( L_4 )</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>21</td>
</tr>
</tbody>
</table>

*Note. \( \chi^2(4) = 9.61, p < .05 \) (one-tailed).

**Accuracy.** Accuracy did not correlate significantly with the total number of propositions \( (r = .21, \text{n.s.}) \) or with the number of correct propositions \( (r = .19) \), but it did with the percentage of correct propositions \( (r = .56, p < .001) \), thus confirming Hypothesis 1. However, all correlations were lower than for Lectures 1 and 2.

An examination of the protocols receiving the best accuracy ratings showed that the summaries contained no obvious errors, although on careful reading a few propositions had been coded as "vague" or "partial," whereas those with the lowest accuracy scores received ratings of 50% or less.

**Completeness.** Completeness scores did not correlate significantly with accuracy scores in this set \( (r = .08) \). Correlations between completeness scores and both the total number of propositions and the number of correct propositions were high: \( r = .58 (p < .001) \) and \( r = .66 \), respectively. However, correlations between propositions at \( L_1 \) and other values were low \( (r = .23, \text{n.s.}) \) for completeness scores and propositions at \( L_1 \) \( (r = 14) \). Thus, Hypothesis 2 is only partially confirmed, in that there was a significant correlation between the com-

TABLE 7
Pearson Correlations for Lecture 3: Historical Linguistics

<table>
<thead>
<tr>
<th></th>
<th>Total Propositions</th>
<th>Total Correct</th>
<th>% Correct</th>
<th>Total at ( L_1 )</th>
<th>% at ( L_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>.21</td>
<td>.19</td>
<td>.56*</td>
<td>.05</td>
<td>-.07</td>
</tr>
<tr>
<td>Completeness</td>
<td>.58*</td>
<td>.66*</td>
<td>.34</td>
<td>.23</td>
<td>.14</td>
</tr>
<tr>
<td>Accuracy and</td>
<td>.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completeness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < .001.
completeness scores and the number of propositions, but a very low one between that and the number of propositions at \( L_1 \).

The protocols with the highest scores for completeness contained quite a few propositions (at least 7), covering points from all parts of the lecture, whereas those receiving the lowest completeness scores had few or no \( L_1 \) statements or had all or most of the propositions on one topic in the lecture. One protocol had 8 propositions, but 6 were only partially correct.

**Markers.** The interaction of signaling markers and level of the hierarchy are given in Table 8. For this set, \( \chi^2 = .8 \), which is not significant. In fact, the distribution of marked propositions in the student protocols and in the lecture was very similar. The data from Lecture 3 do not confirm Hypothesis 3, although it should be noted that this lecture contained only 11 markers.

**Discussion**

The results of the students' summaries of classroom lectures correspond to results obtained on studies using written texts, although the task in the present study was somewhat different, since subjects were not asked to write down everything they could remember but rather to summarize what they learned. Therefore, their responses were selective. However, the hypothesis by Meyer (1975a), that signaled propositions at \( L_1 \) are better remembered than nonsignaled ones, was found in only two of the lectures.

The scores assigned by the two evaluators were very much like grades on an examination. This study attempted to look at the protocols very carefully to see what features of the summaries contributed to high or low scores. In this study accuracy and completeness were separated, whereas in the normal classroom situation both features would be combined. Not surprisingly, accuracy correlated

---

**TABLE 8**
Interaction of Signaling and Level of Propositions

<table>
<thead>
<tr>
<th>Historical Linguistics</th>
<th>No. in Lecture</th>
<th>No. in Protocols</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L_0 )</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>( L_1 )</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>( L_2 )</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>( L_3 )</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>( L_4 )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>30</td>
</tr>
</tbody>
</table>

*Note. \( \chi^2(3) = 0.8 \), n.s. Because the values for \( L_4 \) are 0, \( L_4 \) was omitted from the computation.*
significantly with the percentage of correct propositions, ranging from .56 to .71, thus confirming Hypothesis 1.

As for correlations with completeness scores, there were differences in the three lectures. In all the lectures, the highest correlations were between the completeness score and the number of correct propositions, and secondly the total number of propositions. Significant correlations were found between the completeness scores and the number of propositions at \( L_1 \) only for Lectures 1 and 2. Thus, Hypothesis 2 was partially confirmed in that the correlations predicted were generally found, although the hypothesis incorrectly predicted that the number of propositions at \( L_1 \) would correlate better than the total number of correct propositions. It appears that completeness cannot be reduced to a simple measure; it is a function both of the number of propositions and/or the generality of the propositions.

Signaled propositions at \( L_1 \) showed up in the protocols more often than would be expected by chance in Lectures 1 and 2. The small number of markers in Lecture 3 may have been a factor in the failure to find the effect of marker—\( L_1 \) interaction in that text.

One variable not considered, but which should have been, concerns the use of lecture notes. No instructions were given as to whether students could look at their notes. When the task was announced, students immediately began their summaries, but I do not know whether any glanced back at their notes to refresh their memories. The extensive research on note-taking (e.g., Dunkel, 1988; Dunkel, Shitala, & Berliner, 1989, and references therein) suggests that note-taking without the opportunity to review and rehearse the notes may be of limited use. Dunkel et al. (1989) found a nonsignificant effect for note-taking on immediate recall. Aiken, Thomas, and Shennan (1975) found that note-takers who had no opportunity to review their notes did no better than those who took no notes at all, whereas those who reviewed their notes performed better. Since subjects in the current study were not given an opportunity for review, note-taking was not likely to have been a significant variable.

AN ANALYSIS OF ERRORS

This section deals with some of the features in the protocols that are less amenable to quantitative analysis (namely, the errors, misunderstandings, and infelicitous sentences in the protocols) to see what clues they can provide for language comprehension.

The study of speech errors has been a fruitful data source for helping us understand language mechanisms (e.g., Cutler, 1982; Fromkin, 1973, 1980; Garrett, 1980, 1992). In many of these studies the speech error consisted of a target word or phrase being replaced by something else. In the present study, I have looked at some of the false propositions and lexical "errors," that is, inappropriately used words in the summaries, but words which are close in
TABLE 9
Kind of Errors and Incorrect Words

<table>
<thead>
<tr>
<th></th>
<th>Too Vague to Evaluate</th>
<th>False</th>
<th>Wrong Word</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protocols</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Instances</td>
<td>4</td>
<td>4</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>Lecture 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protocols</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Instances</td>
<td></td>
<td></td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Lecture 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protocols</td>
<td>2</td>
<td>11</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Instances</td>
<td>2</td>
<td>19</td>
<td>15</td>
<td>1</td>
</tr>
</tbody>
</table>

meaning. The use of such words or phrases is open to many interpretations, for example, a temporary lapse in word retrieval, but it seems that in at least some cases the “wrong word” may provide clues about how prose is remembered and represented. Table 9 summarizes the categories of errors.

First, in looking at the incorrect propositions in Lecture 1 there was only one proposition in the protocol that was false and for which nothing in the lecture could be found that might explain the cause. In another case, the student misunderstood the point of an example. Three protocols had at least one proposition which was so vague that its accuracy could not be evaluated. Four protocols contained one or more propositions whose polarity was incorrect; that is, the lecturer said $P$ but the students wrote “not $P$.” A plausible explanation in one case is that the student simply omitted the word not.

In most cases, however, recurrent patterns could be found. For example, in three protocols, the students said that English has eight parts of speech. What the lecturer said was:

How many parts of speech does English have? In the traditional classification there were eight, but then you have to fit everything into these eight categories, and that’s how adverbs got to be such a garbage can; they didn’t fit into anything else. I don’t really know how many parts of speech English has; it depends on whether when you make your subdivisions you want to call them different parts of speech or subcategories. Oh, my guess would be anywhere between 15 and 25.

What we see here is that the incorrect proposition in the protocols was taken from the lecture, but either the students did not listen to the whole section, or their previous beliefs prevented them from accepting and remembering new, conflicting material.

The most common errors involved the use of a term that partially overlaps in meaning with a synonym or near-synonym in some contexts, but which was not quite right in the summaries. These occurred in seven protocols, some of which
had more than one inappropriate word. For example, one of the main points in Lecture 1 was that parts of speech within a language are determined by the distribution of items. Although the students must have known the word distribution, it was apparently used in a new way for some students, and in their summaries where distribution would have been appropriate, they used terms like location or placement: “We call things different because of the location of the things we can do with the word”; “The placement of words determines the parts of speech.”

Another concept that was inappropriately lexicalized in Lecture 1 was expressed several times in the lecture by the word property. This concept was expressed by characteristic or quality: “One big purpose for linguistic research is to identify universal qualities”; “We recognize parts of speech by certain learned characteristics.”

A third word that was inappropriately used was rule. In the lecture, determining the part of speech was described by tests or criteria. The word rule was never used in this context. In fact, rule was used only at the very beginning of the lecture in the following context: “On Wednesday we were talking about phrase structure rules.” Among the student protocols we find the following: “It is not always the easiest case in identifying the parts of speech as the rules are no [sic] always the easiest to point out.” “The way of telling the difference [among parts of speech] is by applying simple rules to the word.”

Some incorrect or inappropriate words may have been misunderstood as well as misremembered. For example, one student wrote, “We also talked about some of the unique features in various languages such as Indian, Japanese, and Chinese.” The lecturer said, “Languages like Chinese, Japanese, and Navajo . . .”

In Lecture 2 there were fewer categories of inaccuracies as well as fewer errors. In one protocol the student apparently misunderstood and made false statements. However, most of the inaccuracies were the result of inappropriate word choices. (There were additional infelicities of style that are probably general problems of poor writing, for example, inappropriate metonymy, but these were not counted as errors.)

One major point in Lecture 2, on the norms of correctness, was that when there are lexical or syntactic variants that correlate with socioeconomic classes, the form used by the upper and middle classes is considered the “correct” one. Words such as influence, determine, and persuade occurred in the lecture, but not govern, dominate, or control, yet these latter two words appeared in protocols more than once: “One group (UMC) seems to dominate the other classes”; “It’s difficult to say who should govern the language”; “Language in the United States is controlled by the upper classes”; “How we talk and what we say is governed by social and political control.” In the last sentences, control is also inappropriate. This word did appear in the lecture, but with a different meaning: “In this society if you don’t control English, doors may be shut.”

Of the three lectures, Lecture 3 contained the largest number of false statements, undoubtedly because of the large amount of new material. Hence, the
students’ ability to remember and summarize the lecture was quite limited. (By contrast the material in Lectures 2 and 3 built on what they already knew or challenged what they previously believed.) Most false statements involved mis-
understanding of new material, but no clear patterns emerged.

In looking at wrong words, there were four clusters of expressions that ap-
peared inappropriately in the protocols using words that were not used in the
lectures. The first was broken down to describe the separation of a language
into dialects and then distinct languages: “The Indo-European is one group of
languages which was broken down into many different languages and which in
turn broke down even further.” The lecturer used split or broken up, as in “Indo-
Iranian split into the Indic languages and the Iranian languages” and “Germanic
would be broken up into West Germanic. . . .”

A second wrong word in two protocols involved using letter instead of sound
or phoneme: “Grimm’s law was discussed and how it applied to Into-European
words changing the letters and all.” The term letter was not used at all during the
lecture, and students had already completed the phonology section and should
have been clear on the distinction between letters and sounds.

Three protocols incorrectly mentioned the origin or beginning of a language
or word: “We discussed the beginning of language”; “We can trace the origin of
our words by tracing the sound changes which occurred”; “Historical reconstruc-
tion is a way to try and see if the origins of a language belong to another group.”
In the lecture there were no propositions that described earlier stages of a lan-
guage as the beginning or the origin. Earlier stage was the most frequent expres-
sion used in the lecture. The only context in which origin was used was when I
discussed semantic pairs of words such as mutton and sheep; “The words that are
of French origin are for what you put on the table.” Because it is common to
speak of earlier stages and earlier periods of a language as original, and to say
things like, “originally, this word meant,” it is easy to understand why the
students’ summaries contained these words.

The last cluster of wrong words involved the concept of language development
or of a daughter language being derived. Both of these words were used at the
beginning of the lecture. In the protocols the following descriptions occurred:
“They [linguists] are tracing language back to see how languages compare and
correspond in order to get a better understanding from where certain languages
formed and how they came about”; “We seem to get our words from Indo-
European which turned into Germanic”; “First you chatted about the 10 language
families and how they all came off of Indo-European.” In one instance derivation
was used, but not appropriately. “There are 10 major derivations of this [Indo-
European] language.”

Discussion
Looking at the kinds of errors and infelicities has shown that there were more
false statements in the protocols on the lecture with the most new material, a
result which is not unexpected. In fact, these kinds of data have been used as evidence for schema theories (see Alba & Hasher, 1983, pp. 207ff. and references therein). It was not possible to further analyze the propositions which were coded as “too vague to evaluate,” because it would have been necessary to have interviewed the students immediately afterward to see whether or not unclear writing was caused by conceptual confusion or simply by insufficiently developed writing skills.

The inappropriate word choices showed that although students may not have remembered the correct or appropriate word, what they remembered was attached to and retrieved from a lexical network in which synonymous and near-synonymous words are stored. If we can use a storage bin metaphor, as Forster (1976) does, and if we suppose that semantically related words are in the same bin, then in these cases the students correctly recalled the bin but not the right item in the bin. These results support the view that words are organized into semantic fields in the mental lexicon, as proposed in much other psycholinguistic literature, such as in the literature on word association and lexical substitution errors (see Deese, 1970; Garrett, 1992; Kittay & Lehrer, 1992; Lehrer, 1974).

These data are also compatible with and possibly support the model of the mental lexicon proposed by Levelt (1989), Levelt and Schriefers (1987), Schreuder and Flores D’Arcais (1989), and Kempen and Huijbers (1983), in which a distinction is made between a lemma, which is the meaning of a lexical entry, and the morphophonological form. Although subjects were able to remember the meaning they wanted to express, they did not remember the form. There are similarities to the “tip of the tongue” phenomenon, but no data are available as to whether subjects would have been able to recognize the correct form.²

Jean Kazez suggested (and other members of the Cognitive Science audience concurred) that when students summarize material, instead of using the same words, they select different words on purpose to show that they understand the material. This was a possibility that did not occur to me, because I was never given such advice as a student. Because the instructions were open on this point I cannot evaluate this variable. However, future work should certainly control for it. In any case, many of the students’ choices were not completely appropriate.

The frequencies of the expressions from the lectures and the inappropriate substitutions in the student protocols were compared where possible, using Carroll, Davies, and Rickman, (1971) and Kučera and Francis (1971). However, these figures are of limited value in that only word tokens are counted without concern for the part of speech or the relevant meaning. The word sound functions as both noun and verb, and as a verb has a wide syntactic range. Property is more frequent than any of the words substituted, but the relevant meaning “characteris-

² Whether these data would support a representation of meaning in the form of abstract meaning units rather than a representation of words is not clear to me. Hayes-Roth and Hayes-Roth (1977), arguing for a word representation, outline the two positions.
TABLE 10
Frequency

<table>
<thead>
<tr>
<th>Expression From Lecture</th>
<th>Expression in Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AHWF</td>
</tr>
<tr>
<td>Distribution</td>
<td>56</td>
</tr>
<tr>
<td>Property</td>
<td>636</td>
</tr>
<tr>
<td>Test (N,V) Criteriaa</td>
<td>1012</td>
</tr>
<tr>
<td>Determine</td>
<td>548</td>
</tr>
<tr>
<td>Persuade</td>
<td>97</td>
</tr>
<tr>
<td>Influence (N,V)</td>
<td>342</td>
</tr>
<tr>
<td>Split (N,V)</td>
<td>217</td>
</tr>
<tr>
<td>Soundb</td>
<td>6373</td>
</tr>
<tr>
<td>Phoneme</td>
<td>58</td>
</tr>
<tr>
<td>Develop</td>
<td>1168</td>
</tr>
<tr>
<td>Derive</td>
<td>105</td>
</tr>
<tr>
<td>Placement</td>
<td>15</td>
</tr>
<tr>
<td>Location</td>
<td>219</td>
</tr>
<tr>
<td>Characteristic</td>
<td>336</td>
</tr>
<tr>
<td>Quality</td>
<td>367</td>
</tr>
<tr>
<td>Rule (N,V)</td>
<td>1062</td>
</tr>
<tr>
<td>Dominate</td>
<td>63</td>
</tr>
<tr>
<td>Control (N,V)</td>
<td>303</td>
</tr>
<tr>
<td>Govern</td>
<td>112</td>
</tr>
<tr>
<td>Broken down</td>
<td>6</td>
</tr>
<tr>
<td>Letterb</td>
<td>3536</td>
</tr>
<tr>
<td>Beginning</td>
<td>1643</td>
</tr>
<tr>
<td>Origin</td>
<td>233</td>
</tr>
<tr>
<td>Form (N,V)</td>
<td>4523</td>
</tr>
<tr>
<td>(Turn into)</td>
<td></td>
</tr>
</tbody>
</table>

Note. Word frequencies are from Carroll et al. (1971), The American Heritage Word Frequency Book and Kučera and Francis (1967), Computational Analysis of Present-Day American English. Verb forms combine all inflected forms (base, -ing, -s, -ed); noun forms include plurals. 
aIncludes one instance of criterion. bOnly base and base + -s are included.

tic” is probably quite rare compared with its more salient meaning of “land holdings and personal possessions.” Moreover, no frequencies are provided for most of the phrases.
Table 10 provides information on the relative frequencies. No clear patterns emerge. In some cases, the word in the lecture is more frequent, in other cases the words in the protocols are. This result is similar to Garrett’s findings (1992) concerning lexical errors involving semantics, where the word substituted for the intended item was not necessarily more frequent.

RELIABILITY OF THE EVALUATORS

One important factor in studies involving judgment has to do with the reliability of evaluators—in this case, of the two teaching assistants who assigned grades to the student protocols. Pearson correlations between the evaluators on accuracy and correctness were disappointingly low, especially on accuracy. However, Pearson correlations may not have been the most appropriate statistic, and subse-
quently $t$ tests for dependent samples were calculated. No significant differences were found, except on the completeness scores for Lecture 3.

In cases where the evaluators differed by more than 4 points (recall that scores ranged from 1 to 9), I read all the protocols carefully and made my own assessment, which in almost all cases was within 0.5 points of the mean of the two evaluators. Therefore, it seemed that using the mean score was reasonable in carrying out the other statistical measures.

Protocols in which there was a disagreement of 3 or more points between the evaluators were carefully studied to better understand what factors were involved. In some cases, systematic differences in evaluating strategies appeared. For example, M.L. gave consistently lower ratings than S.F. on completeness to protocols containing a small number of propositions, even though the propositions were precise, succinct, $L_1$ statements. In addition, S.F. was more generous than M.L. in judging the accuracy of propositions that were almost correct, but not well expressed. This experiment involved two sets of coding: One was the analysis of student protocols, in which two judges used clearly defined criteria, resulting in a high reliability coefficient. The second coding consisted of the two grades, one for accuracy and one for completeness, assigned to each student summary by two teaching assistants. The reliability between these two coders was much lower, but this phenomenon replicates normal classroom procedures and underscores the intrinsic subjectivity of evaluation. This problem is not necessarily insoluble, however. Graders could be presented with clear criteria for evaluation, and they could review a sample of the others' grades until a reasonable consensus could be achieved. Furthermore, if explicit grading procedures could be communicated to students at the beginning of their course, students might be able to utilize such information and produce more satisfactory work.

**STUDENTS’ PERFORMANCE IN THE EXPERIMENTS AND OTHER FACTORS**

Finally the scores that subjects received on their protocols were correlated with their final grade in the course and their rank (freshman, sophomore, junior, senior). In addition, the final examination contained questions eliciting information on norms of correctness and historical linguistics, and correlations were performed comparing students' scores on their protocols with their performance on those two questions.

ANOVA's analyzing protocol scores and final grade and class rank were calculated. A significance level of .05 was reached for grades, with the expected ranking: $A < B < C < D/E$. However, this level of significance was not reached in the class rank, where the order was junior $<$ senior $<$ sophomore $<$ freshman.

More interesting was the fact that there was no correlation between a student's score on the protocol and his or her performance on the question on the final
examination. Pearson correlations for the norm of correctness lecture and final exam question was .08 (n = 30); for the historical linguistics lecture \( r \) was .19 (n = 28).

**SUMMARY AND CONCLUSION**

We have seen that much of the previous research on memory for prose, story comprehension, written texts, macrostructures, signaling devices, and protocol coding can be applied to lectures in a natural classroom setting. This study was somewhat different from many earlier ones in that the subjects were not asked to write down everything that they could remember but rather to summarize what the lecture was about. Therefore, subjects were making an assessment of what they thought was important and/or what they thought the lecturer thought was important. The three lectures also had different overall structures. Lecture 1 on the Parts of Speech had a classical hierarchical structure, and the analysis of the protocols for this lecture showed that macropropositions at the \( L_1 \) level were most frequently mentioned, especially when marked by a signaled word or phrase.

The errors and inappropriate words in the summaries were compared with the lecture’s text to see what features of the lectures’ words may have misled or confused the students. One mistake occurred because the lecturer asserted something and then went on to refute it, but the students did not remember (or understand) the refutation. Most of the other factual errors appeared in the summary of Lecture 3, which contained a great deal of new information, and apparently many students were not able to learn and remember so much at once. Many of the inappropriate words in the protocols from all lectures were apparently caused by the fact that the correct word was used in a new context for the students, and although they could not remember the correct word, they understood and remembered the concept well enough to store it and retrieve a near synonym. Data of this sort—lexical errors, paraphrases, and so forth—provide a wealth of information for helping us understand how the mental lexicon is organized. These results confirm the view that words are organized according to semantic fields. It is also reasonable to assume a distinction between lemmas and morphophonological forms in the representation of a lexical entry.

**What Are the Implications for Improving Classroom Lectures?**

Research of this kind has implications for teaching using lectures. Probably the most important point is that students must not be overloaded with too much new information. Most factual errors occurred in Lecture 3. The instructor should have gone much slower, repeated the major points, and presented much less new material.

Since macropropositions at \( L_1 \) which are signaled are best remembered, in-
structors should make use of this fact by deciding which points they want to stress, formulate them at the appropriate level of abstraction, and explicitly call attention to their importance by a marker.

When introducing students to new words which are technical terms for that subject field, but which are also used in ordinary English, the instructor must stress that it is important to use those specific terms and not synonyms.

When the instructor introduces material which contradicts commonly held beliefs (such as a statement like *English has eight parts of speech*), greater care is needed than when simply presenting confirmatory or new material. The teacher should repeat and call attention to the fact that the common belief is wrong or arguable.

In evaluating students’ examinations, clear and explicit criteria should be provided to students and to graders (in cases where the grading is not done by the lecturer). If examples and details are required in addition to precise generalizations, this should be communicated.

It is necessary to review material from the whole semester for students at the end (or get them to review for themselves), because material learned early is easily forgotten weeks or months later.

**Future Studies**

Because this type of study involves many variables, more work is necessary to control for them. For example, instructions to the subjects can be varied, using four sets of instructions; (1) subjects should try to remember the “correct technical terms”; (2) subjects should try to summarize the gist, but should make no effort to find different words; (3) subjects should try to paraphrase; and (4) subjects should be provided with no special instructions. In the latter case subjects could be asked later whether they tried to find different words.

Another variable would be the use or nonuse of notes. A third variable involves content and the structure of different topics. One possible dependent variable to study would be the effect of the amount of new information presented in a single lecture. If the information covered in Lecture 3 on Historical Linguistics had been stretched to two lectures, would comprehension have been better? Fourth, the signaled propositions emerged spontaneously in the lectures. An experimental design in which the lecturer intentionally planned to use certain markers would provide a more suitable way of measuring the effects of signaled propositions. Fifth, the role of motivation can be explored. Students in my experiments were told that they would not be graded. Perhaps their performance would have improved if they believed that a good summary would contribute to their course grade.

Finally, it would be desirable to design experiments that could help select between different models of semantic networks—in particular whether the semantic networks are lexical links or more abstract semantic feature links.
REFERENCES


