Interactive Learning Environments. (2000) Vol. 8, No. 2, pp. 87-109.

Developing Summarization Skills through the Use of LSA-Based Feedback*

Eileen Kintsch, Dave Steinhart, Gerry Stahl, and

LSA Research Group**

University of Colorado

Cindy Matthews and Ronald Lamb

Platt Middle School, Boulder Colorado

Abstract

This paper describes a series of classroom trials during which we developed *Summary Street*, an educational software system that uses Latent Semantic Analysis to support writing and revision activities. *Summary Street* provides various kinds of feedback, primarily about whether a student summary adequately covers important source content and fulfills other requirements, such as length. The feedback allows students to engage in extensive, independent practice in writing and revising without placing excessive demands on teachers for feedback. We first discuss the underlying educational rationale, then present some results of the trials conducted with the system. We describe the collaborative process among researchers and teachers which enabled the development of a viable and supportive educational tool and its integration into classroom instruction.

Summary Street is an educational software system that uses Latent Semantic Analysis (LSA) to support the reading and writing activities by which students develop and expand their knowledge in new topic areas. *Summary Street* determines the degree to which a student summary covers important source content and conforms to requirements, such as length. It tells the student what information in the source is missing, provides comments on redundancy, extraneous content and certain aspects of mechanics. Its current operation is described in more detail later. First, however, we discuss the underlying educational rationale and review the course of its research and development.

Text-based activities are indisputably a major vehicle for acquiring basic content knowledge in most school settings, across a range of pedagogical models, from those that emphasize traditionally structured classrooms to those in which students direct their own paths of inquiry. One form of computer support for comprehension and learning that our team has developed uses LSA to provide students with immediate feedback on how well their summaries of informative, expository texts cover the topic they are working on. We intend for this tool to be used by students independently, though still within a classroom setting, so that they can assess their own initial attempts to compose and revise their summaries. We hope thereby to provide students with more experience in extended writing and revising, while leaving teachers more time for other kinds of educational activities, such as coaching and modeling writing and summarization techniques, providing individual help, planning and delivering instruction, evaluating final versions of students' writing and other projects. Thus, in no sense is the tool intended to replace the teachers' role, for it is they who must teach the skills and, at least in our implementation, evaluate the final products of students' writing. Even though students are able to use the summarization tool on their own, we want to emphasize that it is a system that seeks to complement classroom instruction, rather than existing as a stand-alone system. Its purpose is to reinforce what is being taught rather than just provide an adjunct learning activity. Thus, in designing our first prototype, called *State the Essence*, we began with the premise that this would take place in collaboration with teachers who were the intended users. The summarization tool in its many transformations and its integration into the instructional curriculum represents a collaborative effort of researchers and teachers.

The current system evaluates only the completeness of the content, for the most part, leaving other important aspects of writing, such as sentence structure, organization and style, for traditional instructional methods. Nonetheless, we believe that in addition to improving their writing skill, students will benefit metacognitively from working independently, guided by the immediate feedback they receive. With frequent practice in assessing and revising the content of their summaries, we believe that students will

also become more attuned to their own thinking and writing processes; they will be more likely to realize what they do and do not understand and better able to express what they mean in writing.

Importance of Summarization as a Learning Skill

Our initial discussions quickly converged on summarization as the kind of learning activity that LSA technology could effectively support and that conformed well with the teachers' instructional goals. The sixth-grade classrooms in which the tool is being tested employs a problem-based learning approach for instructing the district mandated curriculum. Learning how to summarize text is emphasized throughout the school year as a crucial study skill that helps students acquire a basic understanding of difficult and novel subject matter which they can then apply to solving problems or developing a project. Summarizing is more constrained than an open-ended writing task, with which young students often flounder, and it has a number of advantages over simply reading text and answering "comprehension questions", including the following:

- Summarizing not only provides practice in extended expository writing, it also teaches important study skills, such as identifying important content and separating main ideas from details. The fact that students at this age tend to highlight everything in a text – creating a "sea of yellow" – is symptomatic of their inability to do this. This happens especially when students are dealing with content that is completely new to them.
- Summarizing for a given purpose (e.g., to write a report on Mayan religious beliefs) requires even deeper thinking and analysis to select the relevant information.
- Summarizing is a way to develop solid understanding of complex material and also to articulate one's understanding so that it can be shared with others. The teachers with whom we work have noted clear differences in depth of understanding of topics that students have summarized as opposed to those they have only read about. Students appear to retain appreciably more information over longer periods of time if they have summarized it, and in classroom discussion they display an ownership of those topics, which shows up in their ability to contribute detailed and well reasoned ideas.
- Having to express content adequately yet concisely makes students aware of the need to learn summarization strategies that go beyond just adding and deleting single words, phrases or sentences. This awareness becomes a starting point for introducing students to higher-level strategies, such as how to reformulate text content by combining several ideas in a single sentence and generalizing across details.

- Summarizing requires active meaning construction to a much greater degree than choosing a response on a multiple-choice recognition test, or even than writing short answers to isolated questions. Thus, not only is summary writing an effective means to construct and integrate new knowledge, it is also a more authentic method for assessing what students do and do not understand than traditional comprehension tests.

The Use of LSA to Provide Writers with Content Feedback

As the rationale as well as technical details about LSA can be found in various other publications, we will not review them here (please see Landauer & Dumais (1997), Landauer (1998), Landauer, Foltz & Laham (1998), as well as the introductory article by Landauer & Psotka in this volume). Essentially, LSA is an automatic statistical method for representing the meaning of words and text passages based on the analysis of a large amount of textual input. A semantic space is generated in which words, sentences, and whole texts can be represented as vectors. How closely related these vectors are to each other is measured by the cosine between them. We use this cosine measure to calculate what feedback to provide writers.

The most general LSA space available today is based on an input of about 11M words from carefully selected texts that form a representative sample of what a single student finishing high school might have read during his or her school years. This space is sufficient for our analysis, except for technical topics. Thus, for students writing on the functioning of the pulmonary and cardiac systems, or students writing on Meso-American civilizations, the general space does not have enough information to make the fine distinctions required. It has some basic information about the Inca and Maya cultures, for instance, but not enough to tell apart details of their religion or agricultural practices. Therefore, a specialized space must be constructed in order to use LSA. For instance, the Heart space discussed below was constructed from an input of 830 documents comprising about 17,688 words describing the function of the heart. The Meso-American space was based on 530 documents, comprising 46,951 words dealing with this topic. At the moment we do not yet have a good understanding when specialized spaces are required and when the general space suffices. Thus, ad hoc decisions must be made based on the performance of the system.

Because misspelled words are not considered words by LSA, we first have to correct spelling. For this purpose, all misspelled words (or rather, all strings LSA does not recognize) are flagged with asterisks, and the student is asked to make sure that they are spelled correctly. In principle, although this is not done in the present system, a standard spell checker can provide the student with alternatives, and LSA can select the most promising alternative(s) by looking at the cosine between each alternative

identified by the spell checker and the immediate neighborhood of the word. Most likely, words with a higher cosine to the context are the right choice.

Content feedback is provided in the following manner. Suppose students are asked to summarize a text T containing the sections {T_1, T_2,…., T_k}. The teacher requires that each of these sections be covered in the student's summary. What we do is to compute the cosine, C_i, between the summary a student wrote and each of the sections T_i. If C_i $\square$ t_i, where t is an empirically determined threshold value, the student is told that section T_i is not adequately covered in the summary. The student then has the option to look at the appropriate section of the text on the computer screen and add some material about this section to the summary. If C_i $\square$ t for all sections, the student is told that he or she has now covered all parts of the text.

Since the teachers require summaries to be of a given word length to avoid extensive copying (about one quarter of the source text), students are told how many words they have written so far and whether this is within the allowed limits. If the text is too long, the student is given two kinds of feedback to help shorten it. One the one hand, irrelevant sentences in the summary are identified. The cosine is computed between each sentence in the summary and the text as a whole. If it is below some lower threshold, the sentence is identified as (possibly) irrelevant. This relevancy check tends to pick up sentences that are truly irrelevant (such as "I hope you like the summary I wrote") or sentences that refer to obscure details in the text that are not appropriate for a summary. On the other hand, redundant sentences are identified by computing the cosines among all sentences in the summary. If a cosine is greater than some upper limit, the two sentences are highlighted in the text and the student is told to inspect them for the purpose of combining them or deleting one. Sixth-grade students tend to repeat themselves, so this is a very useful check. Note, however, that both the relevance and the redundancy check occasionally pick up false positive: sentences, for example, with several overlapping words, but distinct meanings. This has the positive result that students must critically evaluate the computer's advice and decide whether they agree with it or not. Upper and lower limits for the relevance and redundancy checks are, once again, set empirically. For example, sentences with a cosine to the text that are below .30 might be termed irrelevant, and sentences with a cosine greater than .80 between themselves might be termed redundant.

The system itself is thus quite simple. However, what was not simple was to determine the best ways to provide this kind of feedback to students and the optimal sequencing of this feedback, as described below.

History of Trials Using *State the Essence*: Fall 1997 – Fall 1998

Instruction

Two team-taught classes participated in trials using an early version of the summarization tool called *State the Essence* during the 1997-1998 school year, and a subsequent trial took place in the fall of the next academic year. The system was designed to support students' summary writing in three curricular units, each lasting about three-to-four weeks: Energy Sources (September, 1997 and September, 1998), Ancient Civilizations of the Western Hemisphere (January, 1997) and The Human Circulatory System (April, 1998). Students first composed their summaries using a word processor or pen and paper in advance. They then pasted or typed them into *State the Essence* in order to receive feedback on how to revise them. For the trial on the circulatory system, we collected summaries that students wrote using traditional means as well as those written with *State the Essence*, which allowed us to make within-subject comparisons. However, our main goal during this initial period was to test the system rather than to collect learning and performance data.

1. Sources of Energy. In addition to teaching students about the new content, during the first unit the teachers' instruction introduced students to the concept of summarization and the appropriate strategies. The teachers' instruction included directly explaining the strategies and their purpose, together with modeling the strategies and class discussion of good and poor examples of summary writing.

Students read 10 brief texts (two to two-and-a half pages) about different sources of energy (nonrenewable: coal, natural gas, nuclear, petroleum, propane; and renewable: biomass, geothermal, hydropower, solar, wind) and wrote one summary (75 - 200 words) of each energy type. Students used this task as the starting point for their projects, which involved becoming an expert in one energy source, organizing a science station and teaching the subject to other students in small groups.

2. Ancient Civilizations. For this unit students were required to summarize three texts (each about two-and-one-half to three pages) about the Maya, Aztec and Inca civilizations, again to develop basic knowledge about the cultures. The summaries were to be between 200 to 300 words long. Each class then divided into three groups, each focussing on one of the cultures, and each member of a group researched one particular aspect of the culture (e.g., history, religion, artistic or scientific contributions, social structure). Finally, each group made a joint presentation with visual props to the class as a whole, each member filling in a piece of the topic in jigsaw fashion. The summarization instruction this time focused on higher-level strategies, such as sentence combining and constructing generalizations to achieve conciseness. Students prepared two of their summaries in the traditional manner, using a word processor or pen and paper, and revised a third summary guided by feedback from the summarization software.

3. Circulatory System. Unlike the preceding units, the instructional focus here was primarily on developing a deep understanding of the content - a challenging topic with a great deal of unfamiliar technical vocabulary and difficult concepts. Summarization of two texts about the lungs and the heart was used to help students integrate this information and to assess their conceptual understanding of the dual-loop circulatory system. The summaries were to be 150-250 words in length, and students used *State the Essence* to work on one of these summaries. They wrote the other summary using traditional means.

Evolution of *State the Essence*

Initial trials with *State the Essence* were beset by technical problems from overloading the system with too many simultaneous submissions. However, these problems were overcome in our later trials. In general, the school trials with the summarization software were a success in terms of student enthusiasm and teacher satisfaction, at least to some degree: the system worked well, was relatively easy to learn, and using *State the Essence* did not interfere with students' learning of the content (there was no significant difference between summarizing conditions in scores received on a short-answer test on the unit of study). However, as mentioned, the purpose of these school trials was not a formal evaluation of the system but rather to further develop and refine it.

There are three classes of changes that we explored:

1. How the student's writing is to be evaluated by LSA: There are several options here; for example, a given essay can be matched against a set of pre-graded essays, or against an expert summary prepared by the teacher or expert writer. In the end we adopted a more practical method that would only require a teacher to submit the text to be summarized, subdivided into topic sections, a method that has been incorporated into the later versions of the system.
2. What feedback to give the student, and in what order: It is easy to overwhelm users and confuse them with the rich feedback the system is able to provide. Over the course of the year we experimented with several different feedback formats before arriving at a system that is somewhat constrained yet still flexible to use. "Less is more" was our take-home message - less feedback and more support.
3. How to embed our system into classroom instruction: Use of the summarization tool as a stand-alone system is rather inefficient for middle-school students. Most students at this level need explicit instruction on how to summarize, and how to revise. Furthermore, available technology has made it difficult to use the system in a classroom without taking too much time away from other instructional activities. Our trials therefore took place over one or two sessions with the entire class – a practical necessity, though not an optimal way to learn revision skills.

Evaluating the summaries

Our initial problem in delivering feedback to the students was to decide what text to use as a basis for comparison. Several different approaches to evaluating college students' essays are described in Landauer, Foltz, and Laham (1998), some of which we also applied to evaluating the students' summaries. One approach is to compare a summary to a corpus of previously graded summaries. The summary which is the closest match in terms of the LSA cosine becomes the basis for assigning a grade of A, B, or C, and so on. Since we had not yet accumulated a set of graded summaries to draw on, this option was not open to us. Hence, we first tried matching the sixth-graders' summaries against a set of four or five summaries written by expert writers (teachers and researchers). Given that even expert writers do not completely agree on what content to include or exclude, the student's overall score was based on the best fit (i.e., the highest LSA cosine) to one of the expert texts. Section scores were based on a comparison of the summary to each section of a "golden" summary that incorporated the main content in all the expert summaries. Although this method worked quite well, putting together a set of expert summaries for each novel text proved too cumbersome in the long run.

An alternative basis of comparison is to use the source text itself. A holistic score can be obtained from the cosine between the student's summary and the original source text. In addition, section scores may be derived by dividing the text into distinct topic sections, approximately equal in length, and comparing the entire summary to each of these sections. As described earlier in this paper, a set of empirically determined thresholds is used as the basis for the feedback given to the student on how adequately each section was covered. The summary "passes" when all sections have met the criterion for each section within the given length constraints. This method underlies all the versions of the summarizing software described here.

Presenting the Feedback

LSA-based feedback goes far beyond other forms of automatic feedback, such as spelling and grammar checks, by evaluating the semantic content of a piece of writing. For essays and summaries, it can tell the writer whether or not all the important subtopics have been covered and what kind of information is missing; it can point out sentences that appear to have too much overlap in content with each other or with the original text; and it can suggest sentences that seem to have little relevance to the topic of the text.

In addition to this content information, in our initial trial on Energy Sources we provided students with feedback on the length of their summaries. Length constraints across all three trials varied between 100 to 300 words for texts that ranged from about 800 to 1450 words. Students received an overall score weighted to reflect appropriateness of length, the adequacy of section coverage and overall content

coverage. In addition, they could request checks for (a) redundancy, (b) relevance (both based on a comparison of sentences in the summary with those in the original text), and (c) repetition (based on a comparison of all sentence pairs in the summary). Our sixth-grade students, although appreciative and highly motivated, seemed confused and floundered in their attempts to revise their summaries. In addition to solving various technical problems, it was clear that we needed to provide better editing tools, a clearer presentation, and more support for summarizing and revising both within the system and through classroom instruction. We especially needed to present the feedback in a way that was easier to understand than the set of numerical scores that were initially presented simultaneously.

In our second trial on Ancient Civilizations the feedback was given in three stages, accessed by the user's request first for general feedback, then successively more. The general feedback included length (*too long, too short*), an overall score, and adequate/inadequate section coverage, as before. Requests for more feedback first displayed irrelevant and relevant sentences (the latter were praised); then, at an advanced level, feedback was provided on redundant sentences (summary sentences with too much overlapping content). In addition, we added an overview of summarization strategies to the Introduction to *State the Essence* and hyperlinks to further hints and examples. Links were also provided to the Maya, Aztec, and Inca source texts and to additional background information.

The results of this classroom trial were both encouraging and revealing of significant weaknesses in the system. Again, the overall point score was a great motivator: students were challenged to try to improve their scores and remained focused on the task. However, the scores were not always reliable, tending to be inflated and too sensitive to small local variations. Sentence level feedback was especially problematic, with too many inappropriate flags (both good and bad), and difficult to use because problematic sentences were presented in a list, out of context and on a separate screen from the writer's textbox. Presenting misspelled words as a list posed similar difficulties for making corrections. Even though presented in stages, or at different levels, students were still overwhelmed by the amount of feedback they received and often dismayed at the multiplicity of problems to deal with. Further, many students needed extensive and quite explicit guidance on how to make meaningful changes in revising their summaries; in particular, they needed to be shown how to generalize across sentences or how to combine ideas from several places into a single sentence in the context of their own work. This need clearly goes beyond what LSA-based feedback provides, but highlights an area where the teacher's classroom intervention can be helpful.

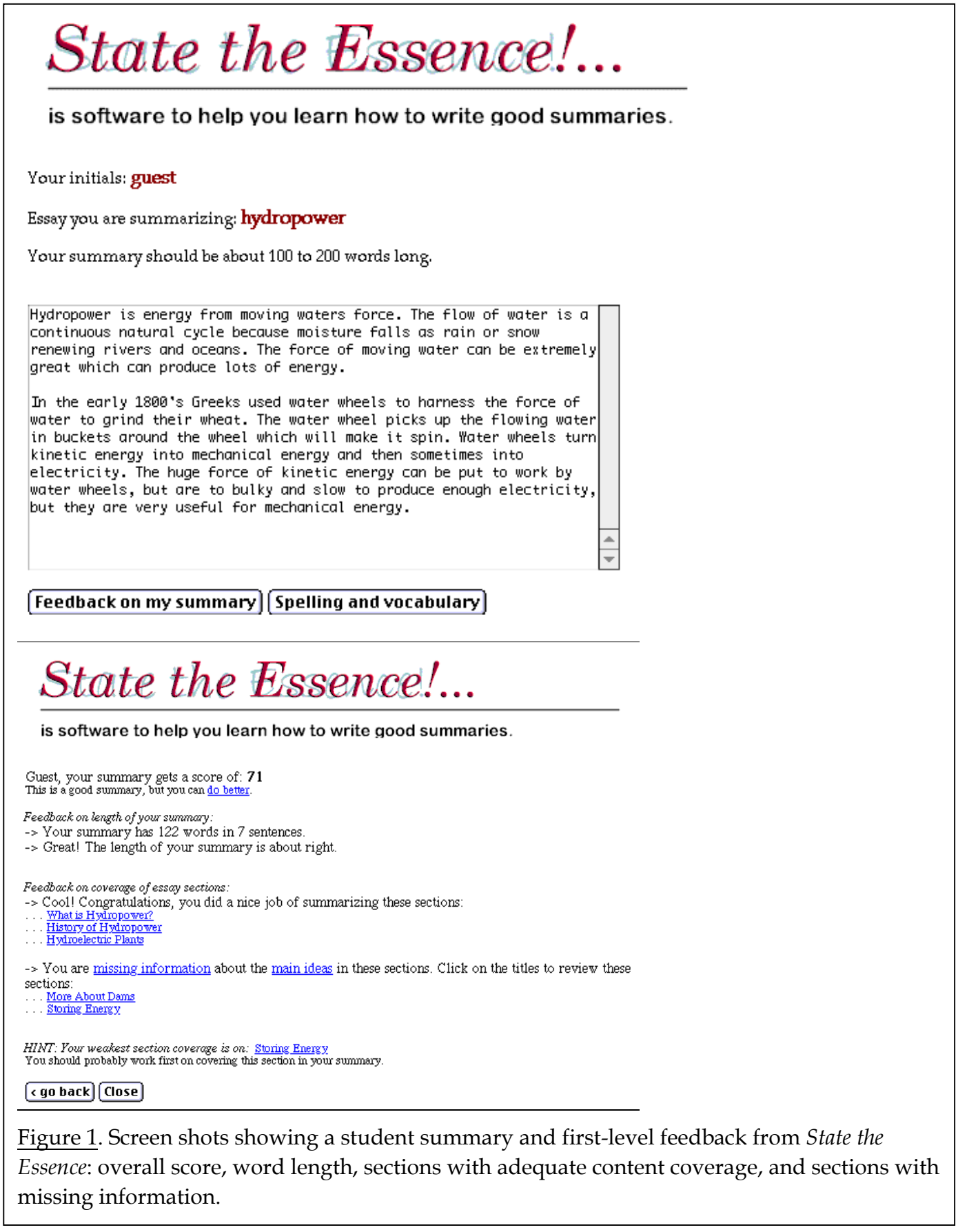


Figure 1. Screen shots showing a student summary and first-level feedback from *State the Essence*: overall score, word length, sections with adequate content coverage, and sections with missing information.

Our next attempt to improve the system consisted in greatly simplifying the feedback, both what was provided and how it was presented. Thus, for the unit on the Circulatory System, feedback consisted only of a point score (0-100 points); length (*too short*, *too long*, or *about right*); an evaluation of the content of each section (*good*, *ok*, *needs improving*, or *missing*); and listing the weakest section with a hyperlink to that section of the source text. The same version of the system was used again with minor changes for the fall 1998 unit on Energy. A screenshot showing the first feedback page for a summary on hydropower is shown in Figure 1.

This version was easier to use, although the overall content scores were still not sufficiently reliable, and interactions were sometimes confusing. For example, students were frequently frustrated to see large decreases in their point score when content was cut to stay within the length constraints. Indeed, the difficulty of balancing the need to be complete and the need to be concise revealed again a need for more explicit instruction of higher-level summarization strategies in the classroom. It is still impossible to provide automatically the kind of concrete help often needed in the context of a student's own summary without, so to speak, giving the answer away. In other words, our computer tool cannot yet interact with a student about their writing like a human tutor, eliciting an appropriate response through carefully calibrated questions. However, it can make both teachers and students aware of where the gaps lie in skill and understanding, which can then be addressed individually or in general class discussion.

To summarize, the goals for our summarization tool consisted of the following:

- to provide support for a challenging activity that fosters both deep learning of difficult new content and promotes writing skills;
- to give students extended practice in writing and revising summaries while relieving teachers from the burden of reviewing and grading successive drafts;
- to motivate students to work hard and independently by providing immediate and individualized feedback on how to revise their writing.

How well did the system fulfill these expectations? As stated previously, we have a modest amount of formal empirical results at this point and quite a lot in the way of informal observations and feedback from both teachers and students.

Empirical results with *State the Essence*

During our classroom trials with *State the Essence* in the spring of 1998 we collected data comparing the summary scores awarded by the system with those of human scorers (Ancient Civilizations). In addition, we examined the effects on learning of summarizing with *State the Essence* vs. using traditional means in a within-subjects

summary street

Your name: **Guest**
Text you are summarizing: **hydropower**
Your summary should be 100 to 250 words long.

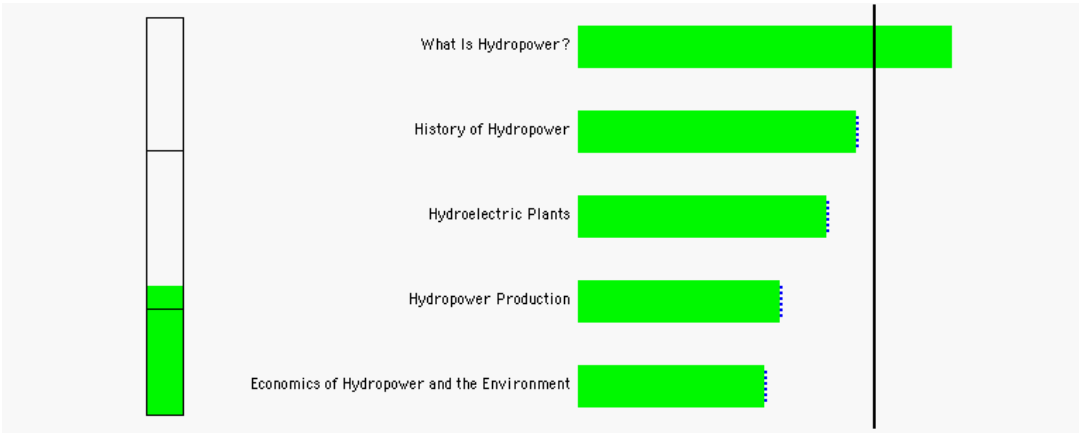
Hydropower is energy from moving waters force. The flow of water is a continuous natural cycle because moisture falls as rain or snow renewing rivers and oceans. The force of moving water can be extremely great which can produce lots of energy.

In the early 1800's Greeks used water wheels to harness the force of water to grind their wheat. The water wheel picks up the flowing water in buckets around the wheel which will make it spin. Water wheels turn kinetic energy into mechanical energy and then sometimes into electricity. The huge force of kinetic energy can be put to work by water wheels, but are to bulky and slow to produce enough electricity, but they are very useful for mechanical energy.

Save changes/check my spelling

Feedback on my summary **Format for Printing**

Close



Guest, these bars show how well your summary covered the sections of the text you read. If the bar passes the black line, then you've written enough information about that section. When your summary contains enough information about every section, you can advance to the next level and will receive more advice on how to improve your summary.

The blue dashed lines show how much information your **previous** summary contained, so you can see if you are improving or not.

Guest, the following sections still need work (you may click on any one of them to see that section of the text in a separate window):

History of Hydropower
Hydroelectric Plants

design in the unit on the Circulatory System. And in the fall of 1998 we again compared teacher assigned scores with LSA scores for summaries of two different energy topics.

1. Comparison of LSA scores with human graders: Ancient Civilizations. For this unit students summarized texts on all three cultures, one using *State the Essence* for feedback on revision. Students were allowed to choose the culture they wished to specialize in. Most students chose to summarize the Inca text. For the first comparison we derived the LSA cosine between the student summary and the text the students had read. We then compared this cosine with the overall content grade assigned by an outside teacher (LK) to the students' Inca and Aztec summaries. For these 50 summaries, the correlation between the teacher grade and the LSA cosine was r = 0.64. The correlation between a second scorer (EK) and the teacher was r = 0.69. Therefore, LSA scores are quite comparable to how an experienced teacher rates these summaries.

The second comparison was intended to assess whether LSA could match a given sentence to a particular section of the source text as well as human graders. Thus, the same two expert graders (LK & EK) were shown 119 randomly chosen sentences and asked to choose which of the five sections of the Aztec text the sentence was describing. For each sentence the LSA cosine was computed against each of the five sections, and the section with the highest cosine was considered to be LSA's "choice." The two expert graders were in agreement for 109 of the sentences (91.6%). LSA agreed with the first grader on 101 of the sentences (84.9%) and with the second grader on 99 of the sentences (83.2%). Therefore, LSA does almost as well as humans at determining the source of knowledge for a given sentence, a fact that we hope can be useful in designing future versions of the system.

2. Posttest scores from summarizing with *State the Essence* versus with traditional means: Human Circulatory System. During the unit on the Human Circulatory System, 39 students from two classes summarized two texts, one on how the lungs function and one on the heart and the circulation of blood in the body. Each student wrote and revised a summary on one of the topics by conventional means (using pen and paper or word processor) and one using the *State the Essence* software. To see if there were differences in how well students had learned the material about each topic, we compared their scores on an end-of-unit, short-answer test on the human circulatory system with respect to these two topics. We found no difference in students' understanding of the two topics related to how the summaries were written, although one class performed consistently better than the other on all questions. A comparison of the average grades (0-10 points) given by two outside teachers (LK & AW) likewise showed no difference in quality of the students' summaries related to condition. The average grades were 6.80 for traditionally written summaries and 6.74 for the summaries written with LSA-based feedback. The agreement between the two human

graders was r = .59. Thus, based on evidence from a single trial, the summarization software did not appear either to benefit or to harm students' learning or writing.

3. Comparison of LSA and human graders: Energy unit. Fifty-six students wrote their two required summaries on chosen topics as homework and used *State the Essence* to revise them. The average correlation based on the grades of the two classroom teachers (CM & RL) with LSA scores for four of the texts (biomass, hydropower, petroleum, & propane) was quite high: r = 0.88. However, the average correlation between teacher and LSA grades was quite low for the remaining six topics: r = 0.32. In part this low correlation between LSA and teacher grades is due to missing data (the summary topics were unequally distributed among the two classes). Additionally it results from the fact that *State the Essence* used a single threshold for all topics which, however, are not equal in terms of their conceptual difficulty. Hence, for some topics students' summaries received higher scores from LSA than the teachers thought they deserved.

Evaluation of the system based on classroom observations

Despite overall encouraging comments on *State the Essence* by both teachers at the end of the year, our concern about the unreliability of the overall score remained. In order to avoid frustrating the students, we wanted to make it possible to obtain 100 points. However, this often made it too easy to reach a high score, which then discouraged students from continuing to revise their work. As mentioned earlier, minor changes often resulted in unreasonable jumps in the score. The students tended to regard their scores as an overall measure of writing quality. Hence, once they had reached 100 points or were close enough, they often did not review what they had written and consequently were upset when they received poor grades from the teachers due to lack of organization and poor writing style. These observations led us to question whether an overall score was a good kind of feedback to provide. Many students treated the score as an end in itself, trying to increase it by the cheapest means possible, rather than focusing on improving their writing. These problems suggested that the feedback should be displayed in a form that was more concrete and easier for students to use than the 100-point score combined with textual pointers in *State the Essence (good, ok, needs more work*, etc.).

History of Trials Using *Summary Street*: Spring 1999

In 1999 we changed the name and modified the interface to reflect a major revision in our approach to providing students feedback on their summaries. After typing their summaries into the textbox, students now receive the following kinds of feedback:

1. Misspelled words are highlighted and can be corrected in the textbox; the student's summary is automatically saved by this operation.

2. The request for feedback returns a graphic display indicating the length of the summary and how well the content of each section of the original text has been covered (see Fig. 2). The display for content coverage consists of horizontal green bars extending out to a vertical line symbolizing the threshold. The weakest section is indicated, and a hyperlink is provided that the writer can use to access that topic section in the original text. Instead of an actual word count, length is shown by a vertical bar on the left, with bisecting lines indicating the prescribed minimum and maximum. A green bar is displayed if the summary length is within these limits, while a red bar is shown if the summary is either too short or too long. Praise is given once the summary has passed the criterion for content coverage for all sections.
3. Further help for revising is available at this point, for example, if the summary exceeds the prescribed length, in the form of a redundancy check and a relevance check. These tools help students locate sentences that have overlapping content or seem not very related to the topic being summarized and that would be possible candidates for deleting or collapsing together.
4. Finally, a "Format for Printing" button allows the student to obtain a double-spaced version to print out, review and hand in to the teacher.

Empirical results with *Summary Street*

The new system functioned quite well during two classroom trials in spring 1999. Students were better able to deal with the feedback on their own and this helped them to stay focussed on their writing for extended periods of time. Although we still see ways to improve the system, future changes should be fairly minor ones. With a stable system now in place we have begun more formal testing of the system than was possible until now.

In trials that took place during spring 1999 two classes of the sixth-grade students used *Summary Street* to compose or revise some of their summaries on-line, guided by the feedback, and other summaries using a word processor or pen and paper. Fifty-two sixth-grade students participated in both the trial on Ancient American Civilizations and the trial on the Human Circulatory System. Classroom instruction which incorporated use of the *Summary Street* software followed the same procedure as in the trials the previous spring with *State the Essence*. Thus, as part of their learning activities during the unit on Ancient Civilizations, students were required to write summaries on three cultures - Maya, Aztec and Inca – one using *Summary Stree*t, the others by hand or on a word processor. The students prepared rough drafts of their summaries as homework and revised them in class. For the Circulatory System unit, students composed and revised both of their summaries of texts about the heart and lungs on-line, either using the summarization software or a word processor, with the experimental conditions *(Summary Street* or word processor) counterbalanced.

1. Ancient Civilizations summaries. The two classroom teachers each scored half of the traditionally written summaries and half of the *Summary Street* summaries on a 10-point scale, but they were not blind to experimental condition since the main purpose of this trial was to try out the new system. The results, which are presented in Figure 3, show for the first time a significant advantage for the *Summary Street* condition: Grades assigned by the teachers are significantly higher for the Inca summaries written with *Summary Street* than those written by hand or a word processor ($t(50) = 2.47$, $p = .02$). Interestingly, both teachers and students considered the Inca text the more difficult of the three, which is confirmed by the lower mean grades these summaries received (8.02 for Inca summaries vs. 8.94 for Aztec and 8.79 for Maya).

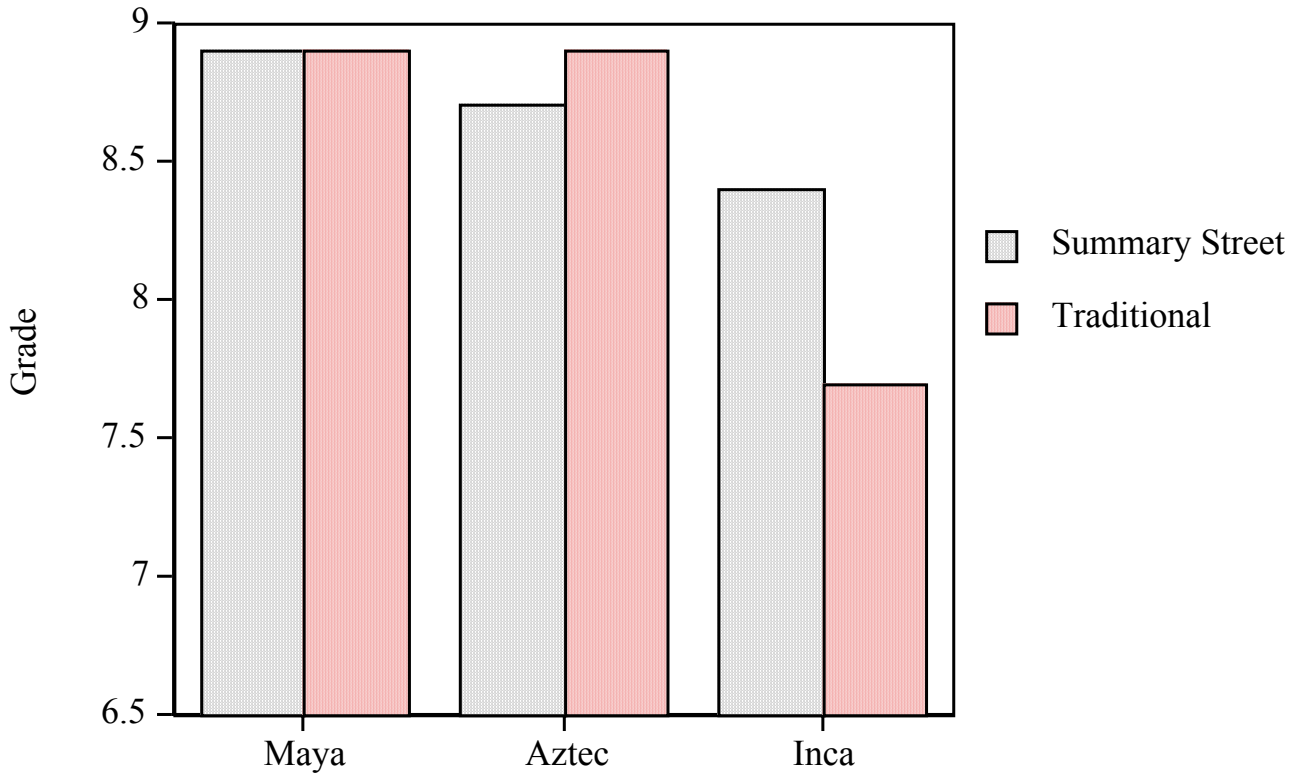


Figure 3. Mean grades of Ancient Civilizations summaries by condition.

2. Circulatory System trial. The results obtained from this trial consist of students' Heart and Lungs summaries and their scores on a posttest at the end of the unit. Summaries of the two texts were scored on a 10-point scale by the two classroom teachers who were not aware of students' identity and experimental condition. Overall, the Lungs summaries received lower grades than those of the Heart texts (mean grade = 3.39 vs. 4.00, respectively), confirming our impression that this text, although shorter, was the more difficult of the two. Quite stringent length constraints also added to the difficulty of summarizing this text. It is noteworthy, therefore, that the Lungs summaries composed with *Summary Street* received significantly higher grades than those composed on a word processor ($t(50) = 2.32$, $p = 0.02$), thus confirming the "hard text effect" we had found with the Inca text in the previous trial. There was no difference in summaries of the Heart text, regardless of how they were written. These results are shown in Figure 4. The posttest grades did not differ significantly across text or condition.

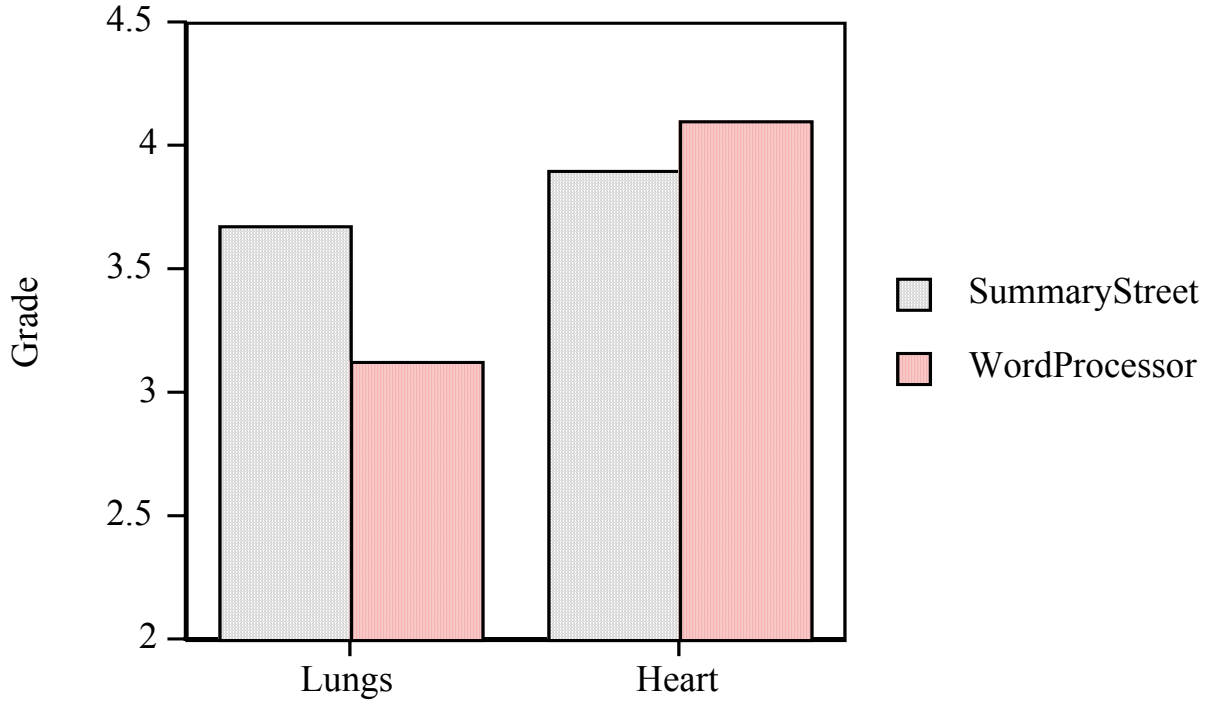


Figure 4. Mean grades of Heart and Lungs summaries by condition.

The results of these two preliminary trials suggest that the content feedback delivered by *Summary Street* is especially helpful when students are faced with more difficult task demands or a harder text. In a recent trial we have further investigated the

effectiveness of the *Summary Street* with hard text in more tightly controlled circumstances. More specifically, we compared the revision process, and its time course, when students are guided by content feedback versus when they only receive feedback on length, such as they would get from a word processor. Preliminary results are very encouraging, but since the analyses are not yet complete, they must be reported elsewhere. In future projects, we hope to replicate the results we obtained with difficult text in an older population of college students. Finally, we would like to investigate the use of this software in collaborative work sessions as opposed to individual work sessions. Our informal observations during the Spring 1999 trials suggest that *Summary Street* could provide an especially rich context for collaborative learning.

Informal observations of educational benefits: 1998-1999 trials

Based on both our observations and what the teachers have reported to us we believe that the new version of the LSA-based summarization tool has the potential to effectively enhance what teachers teach in a meaningful way. Providing students with many opportunities to do extended writing with feedback about their writing is one of the most effective tools for helping them improve their writing. Such opportunities are necessarily reduced if students have to rely on teachers' feedback for every piece of writing. Furthermore, as some sixth graders have pointed out in their evaluation of the software, getting feedback from the teacher just takes too long. Therefore, if we can develop a system that gives students immediate feedback on their writing in a form they can use, teachers will not hesitate to give more writing assignments.

It may be difficult to attribute direct benefits to students' learning from use of the software by any formal means, given the difficulty of controlling instructional, teacher, and a number of other variables in a classroom setting and with a fairly small sample size. Yet we can assume some positive effects on learning from the kind of behaviors we have witnessed in the course of classroom trials, especially those that took place in spring 1999 with the new system:

- Students are able to assume more responsibility for writing and revising on their own, using the feedback on how well their summary covers the essential content. Moreover, they seem to enjoy the challenge.
- By focussing their attention on specific content that is not adequately covered students find it easier to identify important information – certainly a good beginning in learning how to summarize!
- Students who used the summarization tool in the classroom worked hard and long to satisfy the criteria set for them. Being able to track their progress motivated them to work through many cycles of revision guided by the immediate feedback: they kept trying to get it right and in so doing they interacted with the text content for a longer period of time and at a deep and analytic level. Without feedback marking

their progress, students typically make few changes in their writing and are likely to be satisfied with a first draft, nicely formatted with a word processor.

- The summarization tool helps students locate the specific problems with their summaries and makes them aware of the task demands involved in summary writing. It's hard to balance the conflicting demands of topic coverage and conciseness. And it is frustrating to see the coverage indicator go backwards when you delete material to satisfy the length constraint. Thus, in working with the system students feel a need for the strategies being taught. They eventually discover that adding or deleting a word or a sentence is not enough. And besides, which sentence? As they try to revise their summaries students come to realize that they need better methods. These are valuable metacognitive insights that make students more receptive to instruction.

Integrating Summarization Software into Classroom Instruction

It has become clear to us in observing students work on their summaries, students are not always able to apply the feedback they receive from either *State the Essence* or *Summary Street*. For it is one thing to know what the problems are with your writing and quite another to know how to fix them. Nor is it enough to be told what the processes are. Most young writers still need more explicit instruction on how to make appropriate changes, on how to apply strategies like combining sentences or finding a generalizing term in their own summaries. Taking these lessons back to the classroom, the teachers (CM & RL) revised the way they were teaching summarizing, devoting more time to explicit instruction of the skill at a more concrete level than previously. Their new approach involves discussing all phases of planning and writing a summary using a common text:

- discussing the purpose for summarizing information: to develop the ability to share knowledge with others and to come to a deep understanding of complex information;
- modeling how to differentiate main ideas from details in highlighting, to counteract students' tendency to select all;
- developing together an outline of the most important information;
- developing an understanding of what makes a good summary: a focus on main ideas, good organization, low redundancy and making it interesting for other readers;
- showing models of good and poor summaries and jointly identifying the specific properties of each;

- modeling strategies for identifying appropriate content to include, for collapsing and reorganizing information;
- providing more opportunities for students to collaborate in planning, writing and revising their summaries.

The goal of this instruction is to develop a shared vision among the classroom community of the steps involved in summarizing and a common language to talk about and improve writing. This shared knowledge becomes the basis for rich discussion of matters involving both content and style, both in the classroom and between individual students. We find that the summarization tool provides a natural setting for collaborative problem solving by helping students identify specific problems with their writing on their own. Teachers and researchers observed many instances of students spontaneously discussing their work - about whether a given idea really was important enough to include, how information could be combined and collapsed - evidence that students were thinking about the content of their summaries at a deep and critical level. We believe that such discussions came about through the shared learning experience that included more practice in writing, discussions and modeling of summarization strategies in the classroom, collaboration, and the use of the new summarization tool.

Factors affecting use of the system

The system may have potential not only for giving students more opportunities to write, but also for lowering class size for limited times. Once students are familiar with the system, small groups, supervised by a paraprofessional educator or adult volunteer, could work independently on their summaries in the computer lab, while the others work with the teacher in the classroom. Rotating students in the lab and classroom has the dual advantage of giving them more practice at writing – with more feedback than they normally would receive – and more opportunity to interact with the teacher on an individual or small group basis. However, this optimistic scenario depends on factors that are not entirely under the control of either teachers or the researchers. Namely, a person with technical skills must be available at the school to deal with not infrequent interruptions in network connections, through which *Summary Street* operates. The lack of a technology expert in a school places the burden of maintaining the computers on teachers, whose highest priority must be on planning and executing a rich curriculum and mentoring students. In fact this may be one of the major obstacles to adopting the summarization tool as broadly as intended. It is essential, however, that any components of the system that make it unreliable or inconvenient to use (e.g., occasional problems with log-in failures or losing student work) must be corrected.

Suggestions for using the system

The amount of time required for teachers and students to learn the system is not unreasonable, but the time commitment must be made. Classroom instruction on

summarization must precede introduction of the software. It takes about 20-30 min. to demonstrate *Summary Street* to first-time users, which should be followed by adequate time for students to try it out. Thus, teachers should plan one-to-two class periods to familiarize students with the software, because they will need guidance initially to understand the software and how to use the feedback. However, once students have this framework, they can use the feedback from *Summary Street* on their own. If the software is introduced at the beginning of the school year, middle-school students could continue to use it with minimal assistance from a para-educator or volunteer in all content areas. Our broadly stated goal for this summarization tool was to provide students with more opportunities to write. Yet there are a variety of ways it can be used to enhance the effectiveness of teachers' limited resources. Writing a summary need not be viewed as an end in itself, but rather as a step towards another task, such as preparing a report or presentation, or taking a test. The goal of a summarization assignment does not have to be to provide a finished product every time; instead teachers might use a printout of *Summary Street* feedback as a quick, yet authentic way to check students' current understanding of a topic.

In sum, although we aren't quite there yet, we appear to be on the right track toward meeting our goals. Teachers cannot be the sole dispensers of knowledge and feedback. Their job is to give students tools that allow them to evaluate critically the content they learn, and to communicate proficiently with others. *Summary Street* is one more tool that can help accomplish this goal.

References

Foltz, P. W. (Guest Ed.) (1998). Quantitative approaches to semantic knowledge representations [Special Issue]. *Discourse Processes*, *25,* 127-363.

Landauer, T. K. (1998). Learning and representing verbal meaning: The Latent Semantic analysis theory. *Current Directions in Psychological Science*, *7*, 161-164.

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An Introduction to Latent Semantic Analysis [Special Issue]. *Discourse Processes*, *25*, 259-284.

Landauer, T. K., and Psotka, J. (in press). Simulating text understanding for educational applications with Latent Semantic Analysis: Introduction to LSA. *Interactive Learning Environments*.

Author Note

*The authors would like to acknowledge the support of the CSEP Program of the McDonnell Foundation. We also thank the sixth-grade students from Platt Middle School of Boulder Colorado for their participation in this work.

We dedicate this paper to the memory of Ann Brown, in gratitude and profound admiration.

** The members of the LSA Research Group are Walter Kintsch, Thomas Landauer, Rogerio De Paula, Eileen Kintsch, Darrell Laham, Maureen Schreiner, Gerry Stahl, and Dave Steinhart.