Proceedings of the Second European Conference on Computer-Supported Cooperative Work
Bannon, L., Robinson, M. & Schmidt, K. (Editors)
September 25-27, 1991, Amsterdam, The Netherlands

Boosting Connectivity in a Student Generated Collaborative Database

Douglas R. Ward
Centre for Applied Cognitive Science
Ontario Institute for Studies in Education, Toronto, Canada

CSILE is an educational knowledge-media system with which students collaborate to produce a database that encompasses most of their academic work. It is intended to promote the building and exploration of connections between ideas. Keywords are used as the basis for connecting students' notes. A CSILE database constructed by grade 5-6 students was analysed. Although students tend to use few keywords per note, the texts of their notes contain substantial and consistent domain vocabularies. A simulation was conducted of a form of procedural facilitation which would expose this phenomenon to students and significantly boost connectivity in the database.

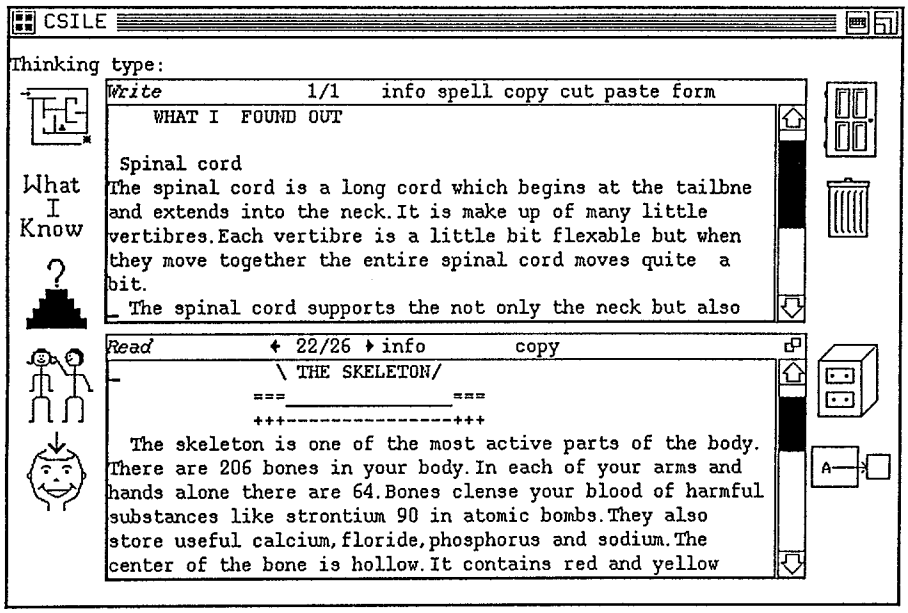
Introduction

Trends in the development of educational computer applications are shifting from the use of computers as surrogate teachers, toward systems which facilitate student-centered "knowledge-building" through the construction and exploration of computerized knowledge bases, often through co-operative or group study efforts. One such application, CSILE (pronounced see-sil), has been studied in a school setting for five years. This *Computer Supported Intentional Learning Environment* is an educational knowledge-media system with which groups of students collaborate to construct a database of text and graphical "notes" about the topics they are studying (Scardamalia, Bereiter, McLean, Swallow, & Woodruff, 1989; see Figure 1).

CSILE is intended to facilitate deep learning and knowledge structuring by its student users as they conscientiously contribute to the database's content and structure, as well as explore the contributions of others. The contributions of

groups of students working on related problems or topics need to become linked in some way so that the students can become aware of and come to understand and utilize each other's ideas in their knowledge-building endeavours. Keywords are a simple mechanism for achieving this integration, because they can provide both a simple summary of notes' contents as well as an index for searching to retrieve related notes.

Figure 1. Screen from the CSILE program showing sample notes that the user is writing (top) and reading (bottom) following a search of the communal database.



CSILE attempts to support communal knowledge integration by allowing students to make and explore connections between notes. Notes can be considered *connected* if they share keywords; a search for all notes with keyword "X" retrieves a collection of notes which presumably have something important in common (i.e. they are all about "X"). It is desirable to have a substantial amount of connectivity in the database, and for the user-interface to contribute to the user's sense of connectivity between notes. Efforts to retrieve connected notes should be rewarded by the return of notes which show what users' peers are thinking and writing on subjects of interest. A highly connected database will provide greater opportunity for this.

Studies of information retrieval systems suggest that there are generally serious problems with keywords as a means of accessing information from databases. Furnas, et al. (1983) point out that imprecision in the way humans name and refer to objects might lead to reduced performance in retrieval tasks. Random pairs of people were shown to use the same word for an object only 10 to 20 percent of the time. The usual implication of this is that users of a database cannot be very confident that they are retrieving a large proportion of relevant documents, and a

small proportion of irrelevant ones; it is often difficult to anticipate the exact keywords used by the system designers or database indexers.

The traditional solution to this problem has been to restrict keyword indexes to a small set of valid descriptors. This has been shown to improve precision in meaning by increasing consistency of keyword use between indexers and users (Tinker, 1966). However, this is only suitable in applications where domain vocabulary is already standardized among database users. In the case of students building a CSILE database, knowledge from any domain might be contributed, so it would be impossible to anticipate what keywords might best be included in a restricted set of descriptors.

Another way to overcome the problem or imprecise keywords in information retrieval systems is to allow an object to be identified by many not necessarily unique words (Furnas et al., 1983). In related studies, Gomez et al. (Gomez & Lochbaum, 1985; Gomez, Lochbaum, & Landauer, 1990) demonstrated that retrieval success can be improved if the number of different names for a data object is increased. It might be expected that large index vocabularies could become awkward and ambiguous, with many objects sharing the same keywords, but the rich indexes generated by allowing "unlimited aliasing" in these studies were seen to facilitate retrieval without imposing any obvious human performance cost.

The design of CSILE is directed toward students *constructing*, rather than just exploring or retrieving information from a database of their collective knowledge. So, although it is clear that enriching keyword indexes in CSILE databases might improve retrieval success, the problem of how to get student users of the system to generate and assign these useful keywords to their notes remains. Skills of assigning and searching by keywords are neither presupposed for students using CSILE, nor do the students receive any specific instruction to develop these skills. The overall design challenge in CSILE is to provide *procedural facilitation* (Scardamalia & Bereiter, 1984; Scardamalia et al., 1989) for students to structure and elaborate their own individual and group knowledge. The CSILE user interface should support the efforts of students and foster appropriate and effective use of keywords, without presuming the skills of an expert database user, and without providing any substantive help which would direct the content of students' notes.

The version of CSILE used during the course of this study required students to assign at least one keyword before any note could be stored. Students could either select from a scrolling list of all keywords used in the database, or type one which may or may not be in the list. No other assistance was provided for the selection of keywords. To retrieve information stored in each other's notes, students formulated Boolean search statements based on keywords, authors, topics, or other criteria, through a simple "point-and-click" dialogue. The notes returned by searchers were displayed, one at a time, in a "read window".

This paper reports on young students' use of keywords as a mechanism for building and discovering connections between pieces of the collective knowledge

base. Based on observed, sub-optimal patterns of keyword usage, two methods of procedural facilitation are suggested. A simulation verified that keyword standardization and enrichment processes could be valuable in boosting connectivity in the database in a way that could promote the integration of student knowledge.

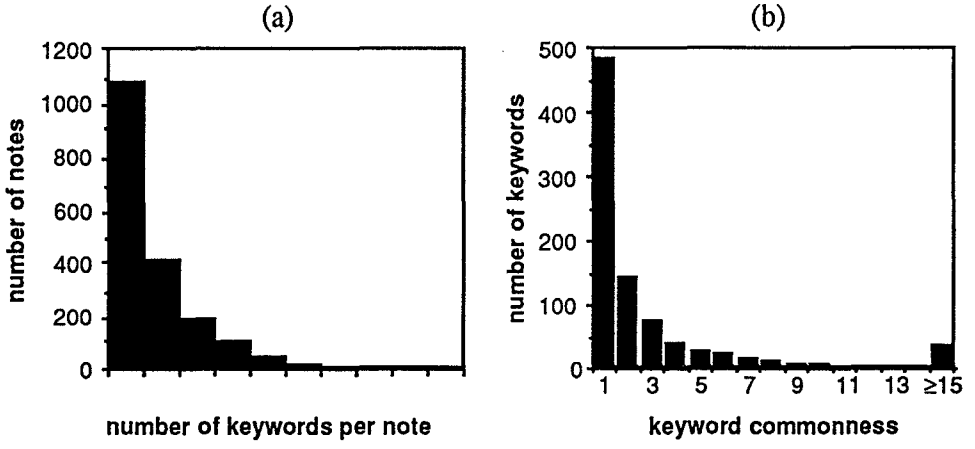
Analysis of Keyword Usage

A CSILE database, generated over nine months by two grade 5-6 classes (age 10-11), was examined for this study. CSILE was used as a regular part of learning activities in the classrooms, and the topics represented include many of their units of study. A total of 1893 student-generated text notes were captured and analysed, along with system-generated transaction logs of student activities with the system.

Results and Discussion

Although it is possible to assign several keywords to each note, only one is required before a note can be stored. In 57.5% of the notes examined, only one keyword had been assigned (see Figure 2a). The maximum number of keywords per note was 10, although fewer than 10% of the notes had more than 3 keywords. Some students probably would not bother with keywords at all if they were not required, but others have indicated in interviews that they liked to put more keywords on their notes so that more students would read them. Some students seemed to appreciate the value of good keywords in note retrieval and building connections in the database.

Figure 2. (a) Frequency distribution of notes by number of keywords. (b) Frequency distribution of keywords by commonness. (The commonness of a keyword is the number of notes to which a keyword has been assigned.)



Of the 892 different keywords used, 54.3% were assigned to only one note; thus, most keywords were very uncommon in the database (see Figure 2b).

Although the most common keyword was assigned to 256 notes, only 3.8% of the keywords were assigned to 15 or more notes. There are several possible reasons for students not assigning the same keyword to multiple notes. The interface may provide disincentives to do so in that it is difficult and time-consuming to scroll through over 800 terms to find an applicable keyword. Students often create "private" or unique keywords to act like "file names" for their own use in retrieval, not thinking in terms of others retrieving their notes. The reasons for uniquely identifying one's own notes with keywords might be more apparent to the students than those for connecting one's note to other notes by assigning common keywords.

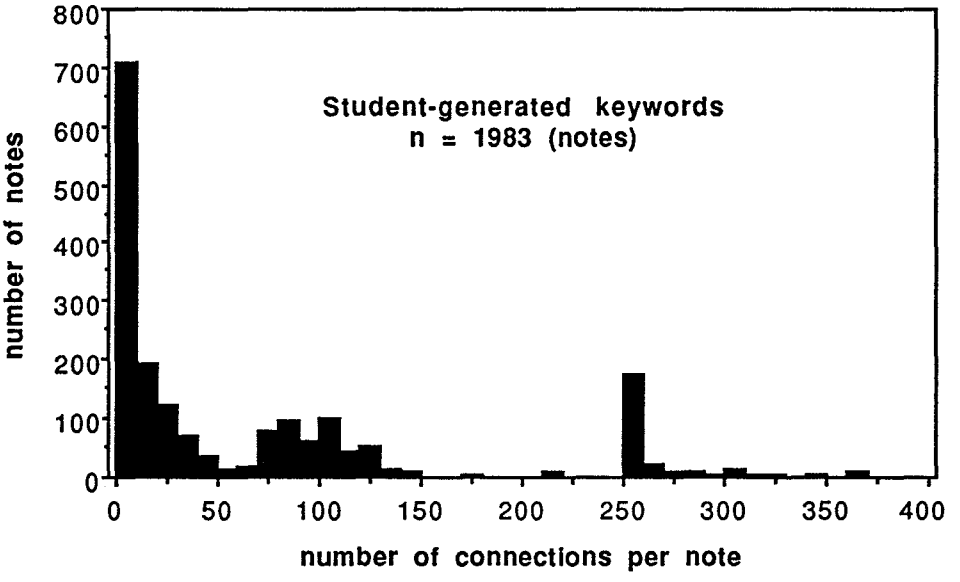
Most of the students' keywords could be identified as vocabulary relevant to the knowledge domains under study. The keyword list is not an exhaustive lexicon of any domain, but the terms students chose are certainly relevant to their fields of inquiry. There are, however, two main problems with the keywords students had invented. First, many phenomena reduced the potential for the formation of keyword connections due to a lack of standardization of terms: slight variations in spelling, case, punctuation, and abbreviation, unnecessary suffixes (e.g. "-ed", "-ing", "-s"), and the addition of articles (e.g. "the...", "a...") or pronouns (e.g. "my..."). There were many cases where a single term could have stood in the place of several independently entered "versions" of the same keyword. This is certain to reduce the potential for notes to share keywords. The second problem was with terms having little semantic value in describing the contents of a note. Unfortunately, some of the most common keywords fell into this category: "comment", "vocabulary", "book review", "junk", "plan", and "question". Although these keywords suggest something about the type of knowledge represented in the note, they do little to indicate specific information. There were also numerous examples of uncommon keywords which did not seem to indicate their notes' contents. In all, 282 keywords out of 892 were identified as problematic in either of these two main ways.

A simple measure of connectivity was generated to indicate the richness of the keyword connections the students had built into their database. Each keyword in a given note may connect that note to few or many other notes, depending on the keyword's commonness. The total number of other notes with which a given note shares one or more keywords is a measure of how connected that note is. Connectivity across the whole database is described in Figure 3 as a frequency distribution of numbers of keyword connections per note.

A large portion of the database (37.5% of all notes) had fewer than 10 keyword connections per note due to the frequent use of uncommon keywords. As many as 9% of the notes were completely unconnected. On the other end of the distribution, 13.5% had 250 or more connections due to the use of very common keywords, often in combination. Because the most common keywords were mainly not indicative of note contents (the second problem noted above), and the better

keywords were less common, the amount of connectivity which might be of any use to students in retrieving related notes is really quite small.

Figure 3. Frequency distribution of number of keyword connections per note.



Out of 5514 searches conducted by students over the period of construction of the database, only 20% use a keyword as a search term. Keyword searches which either return very few or very many notes don't facilitate the discovery of interesting relations between notes. Without keywords providing a major role in information retrieval, it is likely that they have little apparent function to many students. In fact, students tended to pay very little attention to keywords. To view the keywords and other note information, the user is required to open a special "info" window; this was done for less than 8% of the notes which were viewed in the read window.

Simulation of a keyword enrichment facility

Based on the previous analyses, it was possible to suggest two simple forms of procedural facilitation which might eliminate some of the identified problems with students' keywords, as well as increase the connectivity in the knowledge base. First, the computer could suggest modifications to new keyword entries to help students *standardize* their keyword vocabulary without any substantive interference in the keyword selection process. Second, a keyword *enrichment* facility could suggest additional keywords for each note, based on the overlap of textual contents of notes with the user-generated keyword list, without creating any new keywords.

The use of these facilities has been simulated, *a posteriori*, on the students' own database.

Keyword Standardization

In order to eliminate problems of the lack of standardization of keyword form as described above, the following "soft" rules were applied:

- encourage correct spelling
- encourage singular words
- encourage root words (no suffixes)
- discourage extraneous punctuation
- discourage pronouns and articles
- discourage numerals
- ignore case
- discourage "invalid" words (as determined by teachers or designers)

As a form of procedural facilitation for real users, it would be important that these rules be instantiated as suggestions by the computer of an alternative keyword when a "problem" keyword is entered, so that the user could retain ultimate control over the choice of keywords. For the simulation, however, it was assumed that users would accept all suggested alternatives, and changes were applied to all instances of the 282 problem keywords.

Keyword Enrichment

It was hypothesized that many of the keywords which represent domain vocabulary would actually appear in the texts of more notes than those to which they had been assigned as keywords. If these terms were assigned as keywords, there should be an increase in the number of keywords per note, commonness of keywords, and overall connectivity through the database. This would be instantiated by having the computer suggest additional keywords at the time of creation or modification of a note, or later as an updating process. Either way, the user should be able to accept or reject the computer's suggestions. For the simulation, though, it was assumed that users would accept all suggested additional keywords.

The corrections and eliminations of problematic keywords were specified manually by creating a list of changes which were to be made. A computer program then effected the changes in the database. Because some "invalid" keywords were deleted from notes, some notes were left without any keywords. These notes were discarded from the following analyses, because it could not be determined what alternative keyword the student author might have assigned. This reduced the total number of notes analysed to 1392.

Following this process, a computer program scanned the text of every note, comparing each word with the main list of keywords (those which had been standardized). Where matches were found, the keyword was assigned to the note, generating a new database with standardized and enriched keywords. The following

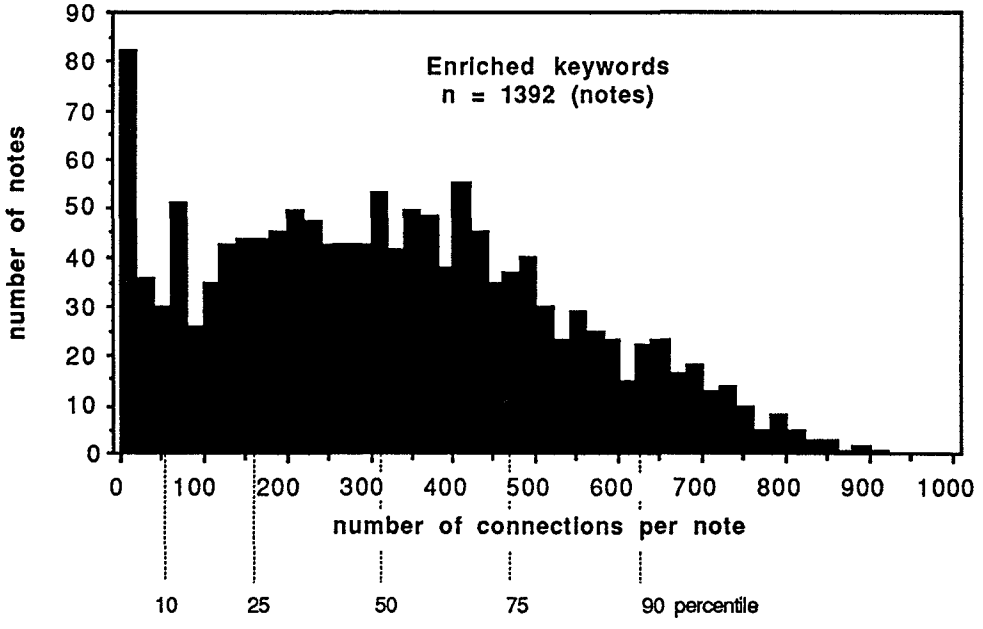
analyses compare this keyword-enriched database to the same notes in their unaltered form.

Results and Discussion

The unaltered database of 1392 keywords contained a total of 865 different keywords. After the standardization process, there were 24% fewer (656). The median commonness increased from 1 to 5 notes per keyword. The proportion of keywords used only once fell from 55.0% to 23.9%, and the proportion used 15 or more times increased from 3.4% to 27.1%. The enriched keywords had an improved likelihood of being shared by multiple notes.

The number of keywords per note also increased dramatically. Whereas the unaltered notes had no more than 10 keywords and a median of 2 per note, the enriched notes had as many as 35 keywords, with a median of 6. The proportion of notes with only one keyword fell from 44.3% to 6.8%. The data confirm the hypothesis that many of the keywords representing domain vocabulary are actually used within the text of students' notes more often than assigned as keywords. It is potentially very empowering to make this phenomenon visible to students, so that they know that other students are contributing information related to their own notes, and that the keywords are a means of accessing those related notes.

Figure 4. The distribution of connectivity after keyword enrichment.



The median level of connectivity increased from 8 to 315 connections per note (see Figure 4). Only 0.8% of the notes remained completely unconnected, 3.7%

had fewer than 10 connections, and 5.9% had more than 20. A statistical comparison of the connectivity distributions before and after enrichment is based on the null hypothesis that if both connectivity distributions are the same, it can be assumed that equal numbers of notes from each set of connectivity scores will fall within any range of percentile scores from the combined distribution. Table I shows percentile scores of the combined distribution, and numbers of notes from each independent distribution which fall within each range of the combined percentile scores. A Chi-square test for homogeneity of the distributions shown in table I indicates that the two distributions do not contribute equally to their combined distribution ($\chi^2 = 1848$, df = 5; $p < 0.001$). Connectivity was radically increased by the enrichment process.

The *a posteriori* simulation of keyword enrichment might have exaggerated the potential to improve connectivity over the evolution of a database. The first notes produced on a topic will not gain as much from the enrichment facility as later notes, because users have not yet provided the domain-relevant keywords with which to cross-reference their notes. The fact that students create keywords as they create their notes remains a strength of the system, however. If a process for updating keyword connections to older notes is available as new keywords are created, these estimates of connectivity are tenable.

Table I. Frequency of notes in each of the unaltered and enriched connectivity distributions falling within the ranges of certain percentile scores from the combined distribution.

combined percentile *score*		10th *2*		25th *7*		50th *48*		75th *315*		90th *509*	
Unaltered	301		716		249		126		0		0
Enriched	20		28		78		571		418		277

Conclusions

For CSILE to meet its objectives as a communal knowledge-building environment, students need to consciously utilize some mechanism for creating and discovering connections between their collective ideas. Keyword connections might form the basis of such a mechanism, but an examination of unaided keyword usage has revealed several critical problems. Students tend to assign one or very few keywords to each note. They tend not to re-use the existing keywords from their own and others' notes, creating new ones instead, even when similar keywords already exist. Many keywords have little value in terms of describing notes'

contents. These facts combine to create a database with very low connectivity. This may explain the relative infrequency of keyword searches.

It should be beneficial in CSILE to facilitate an increase in the number of keywords assigned to each note by suggesting keywords which other students (or the same student at other times) felt were important descriptors of other notes. This would not only increase the probability of retrieving a given note through a keyword search, but boost the connectivity amongst the notes, as demonstrated by our simulation.

The advantage of providing this sort of procedural facilitation to students using CSILE goes beyond simply improving information retrieval performance. Students using the current version of CSILE do tend to use important domain vocabulary as keywords for their notes, but they do not benefit from the potential for building meaningful connections between their notes. If students can be informed that several other notes have a keyword which is in their note, the decision to assign the keyword becomes a decision to build connections between notes in the database. This should contribute to the sense among the student users of constructing a collaborative knowledge base, rather than merely entering personal notes into a public database.

Automatic full-text indexing could increase connectivity in the database as much or more, but enriched keyword indexing is expected to have greater benefits for student users. Analysis of expert on-line searchers reveals that expertise in information retrieval stems largely from experience, and a familiarity with the vocabulary and contents of a database (Oldroyd, 1984). CSILE users are in the advantageous position of searching within a database of their own creation. Awareness of the content of the communal database can be enhanced by keyword enrichment suggestions at the time of creation and storage of notes. Connections may more profitably be explored through searching if users are more familiar with each other's keywords, and this will in turn expose the students to more of the database. The intentional construction of an integrated knowledge base can only be enhanced as users become more familiar with the contents and vocabulary of their database, and aware of connections among their contributions.

Enrichment of keyword complements of students' notes can radically *increase* connectivity, but can it really *improve* connectivity? An optimal level of connectivity would exist if there were keyword connections between all notes which really have related contents, and none between those which do not. An increase in connectivity would be an improvement only to the extent that keywords assigned to multiple notes point to related concepts in those notes. It is possible to imagine many cases where either the creation of poor keywords or the poor application of facilitated enrichment would yield levels or sorts of connectivity which might not profit the builders of a large database. Further research will reveal whether implementation of the suggested procedural facilitation in the CSILE user interface will improve the construction of integrated student knowledge bases.

References

Furnas, G. W.,Landauer, T. K.,Gomez, L. M., & Dumais, S. T. (1983). Statistical semantics: Analysis of the potential performance of key-word information systems. *The Bell System Technical Journal, 62*(6), 1753-1803.

Gomez, L. M., & Lochbaum, C. C. (1985). People can retrieve more objects with enriched key-word vocabularies. But is there a human performance cost? In *Proceedings of Interact '84* (pp. 257-261). Amsterdam: North Holland.

Gomez, L. M.,Lochbaum, C. C., & Landauer, T. K. (1990). All the right words: Finding what you want as a function of richness of indexing vocabulary. *Journal of the American Society for Information Science, 41*(8), 547-559.

Oldroyd, B. K. (1984). Study of strategies used in online searching 5: differences between the experienced and the inexperienced searcher. *Online Review, 8*(3), 233-245.

Scardamalia, M., & Bereiter, C. (1984). Development of strategies in text processing. In H. Mandl,N. Stein, & T. Trabasso (Eds.), *Learning and Comprehension of Text* (pp. 379-406). Hillsdale, NJ: Lawrence Erlbaum Associates.

Scardamalia, M.,Bereiter, C.,McLean, R. S.,Swallow, J., & Woodruff, E. (1989). Computer-supported intentional learning environments. *Journal of Educational Computing Research, 5*(1), 51-68.

Tinker, J. F. (1966). Imprecision in meaning measured by inconsistency of indexing. *American Documentation, 17*, 96-102.

Acknowledgements

This article reports on only a small part of the CSILE project. The development of software and provision of hardware to design and implement CSILE is made possible by support from Apple Computer, Inc., External Research Division.

The author wishes to acknowledge the support of Marlene Scardamalia and the entire CSILE research and development team for sharing their resources to make this research possible. In particular, thanks go to Jim Hewitt for his tracking and analysis programs. Thanks also to Charles Laver and Jim Webb who teach at Huron Public School, and the Grade 5 and 6 students who contributed their time and knowledge to this study. The author is also indebted to Bob McLean and others who proof-read and commented on versions of this article.