STUDYING C-TOOLS: AUTOMATED GRADING FOR ONLINE CONCEPT MAPS

Douglas Luckie[1], Scott Harrison[2], Joshua Wallace[1] and Diane Ebert-May[3]

[1]Lyman Briggs College of Science and Department of Physiology,[2]Department of Microbiology and Molecular Genetics,[3]Department of Plant Biology, Michigan State University

Abstract

The C-TOOLS research group has developed and released a new assessment tool, the Concept Connector, consisting of a web-based, concept mapping Java applet with automatic scoring. The Concept Connector was designed to enable students in large introductory science classes at the university level to visualize their thinking online and receive immediate formative feedback. The Concept Connector's flexible scoring system, based on tested grading schemes as well as instructor input, has enabled automatic and immediate online scoring of concept map homework. Criterion concept maps developed by instructors in the C-TOOLS project contain numerous expert-generated or "correct" propositions connecting two concept words together with a linking phrase. In this study a range of holistic algorithms as well as WordNet®[1], an electronic lexical database and thesaurus, were used to test existing and potential automated methods of scoring to supplement the answers provided by instructors. For this study 1298 student-created concept maps (with 35404 propositions) were evaluated by different "holistic" methods of automatic grading, and in particular, WordNet's computer-generated propositions were tested for validity.

Introduction

Expert-level thinking depends on a web of mental connections developed over a lifetime of education and experience (Bruner, 1960). Yet, in an attempt to turn college science students into experts, instructors often focus on passive transmission of large amounts of "content" in a short time period and then test students to see if they "got it" (NRC, 1999). In response, students tend to follow practical ways to succeed in their courses and thus adopt strategies like memorization or rote learning (Ausubel, 1963; Novak & Gowin, 1984). Visual models such as concept maps may help instructors begin to teach expert thinking as well as assess domains of student understanding. The value of knowledge scaffolding tools such as concept maps is that they reveal student understanding about the direct relationships and organization among many concepts.

Those of us that teach science should also reflect on our best practices as scientists. In our own learning as scientists, we use visual models to understand complex systems, to communicate our ideas to our peers, and to deduce testable hypotheses. Models are one of the common themes in science; they are "the main vehicle by which science actually produces its explanations and

---

[1] WordNet is an internet-based lexical database/thesaurus and registered trademark of Princeton University.

predictions" (Casti 1990, p. 31). Hence perhaps it is natural as well as appropriate that students and scientists should use visual models to describe and evaluate science.

When using concept maps in our own classroom we faced a number of logistical challenges. Initially the use of paper and pencil seems to be the most natural way to create concept maps. Students can easily create shapes, words, lines etc and add small illustrations. Yet, as students become more proficient or engaged in making a concept map, problems arise when they'd like to revise it. Erasing can become tedious and inhibit the process of revision. Using "Post-It" notes can overcome this and allow easy revision, yet a record or copy of the map is not easily generated in the active classroom. An additional challenge is scoring maps. While grading a single concept map may be less time-consuming than grading an essay or extended response, it is still more complex than grading multiple choice exams. Even if a chemistry instructor would like to use concept maps in their large introductory course of 500 students, they will point out that grading 500 maps is not practical for them.

Computer software is an avenue to address these challenges. In fact, a number of projects, like the Inspiration™ or Semantica™ commercial software, and the freely downloadable, community-oriented IHMC CmapTools software, present excellent replacements for paper-and-pencil drawing environments and may help engage the resistant student. Although computer-based tools for concept mapping are available to university faculty, few are web-based and none have automated scoring and feedback. As a result, the C-TOOLS research group developed a new assessment tool, the Concept Connector, consisting of a web-based, concept mapping Java applet with automatic scoring and feedback functionality. The Concept Connector is designed to help "make transparent" when students do not understand concepts and motivate them to address these deficiencies. Web-based concept mapping can enable students to save, revisit, reflect upon, share and explore complex problems in a seamless, fluid manner from any computer on the internet (Pea et al., 1999). Automated grading and feedback features can allow instructors to use concept mapping for reflective thought outside the boundaries of the class meeting.

The focus of this study was the automatic grading feature associated with the Concept Connector is called Robograder™. We wanted Robograder to be able to use computational power to help instructors grade concept maps. In 2003 Robograder began to use synonyms from WordNet® (Fellbaum, Ed., 1998) to add to the set of "correct" answers provided by instructors. At the time of this study, for 35404 propositions generated by students on their concept maps, 9211 could be evaluated by Robograder. The use of synonyms from WordNet enabled 971 of the 9211 propositions to be evaluated when the existing grading matrices would not otherwise make an assessment. Currently, Robograder indiscriminately accepts linking phrase synonyms independent of frequency and word sense. In theory amplifying grading rubrics by using all available synonyms should introduce errors into rubrics since multiple and conflicting meanings often exist. In our findings from this study, visual examinations of automatically graded maps indicated few false positives or false negatives despite Robograder's treatment of all synonyms of a word as interchangeably equivalent. One explanation for the observed success of indiscriminate acceptance of synonyms is that users may more likely choose words within a relevant set of

synonyms (known in WordNet as a "synset"). Thus, when developing concept maps, students appear to be less inclined to just randomly pick synonyms but rather they use semantically plausible word choices.

Background

Building the software, the Concept Connector

With both the literature providing a solid theoretical basis for using concept maps ( ) and the field of computer science providing the proper software development tools and technology, the C-TOOLS project began in 2002. A team of faculty from Michigan State University spent much of the first year of the project developing both the Java applet, called the Concept Connector (Figure 1), and the classroom problems sets with concept maps for science students. In parallel with software development was a study of how students used the tool. The Concept Connector was developed through a 'design' experiment (Suter & Frechtling, 2000) that involved testing the tool with undergraduate science-majors in biology, geology, physics and chemistry courses.
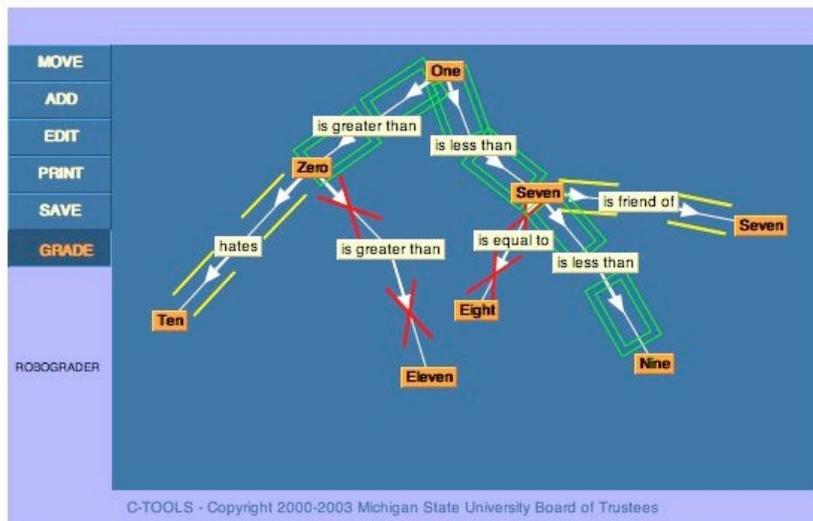
Figure 1: The Concept Connector Java applet graphic user interface (GUI). This particular screenshot shows the Java applet's GUI (blue colored areas), how the software draws a concept map, and how new colors (green and yellow rectangular halos or red X's) appear when the *Robograder* is asked to *GRADE* a concept map (http://ctools.msu.edu). A concept map contains *concepts* (orange) and *linking words* or phrases (yellow). When two concept words are connected by a phrase the unit is called a *proposition*. The drawing software is called the Concept Connector and it's autograding feature is called Robograder. WordNet is an online thesaurus Robograder can access to amplify it's grading matrix. In the image shown, mathematical terms are used to demonstrate the tool and Robograder is utilizing a math library to evaluate the propositions.

The Concept Connector is the combination of an online Java applet that serves as a map drawing tool residing in an HTML page that communicates with server-side software on POSIX-

conforming systems such as Mac OS X®, LINUX®, and FreeBSD®. The applet is small in size and is browser-compatible on every OS platform and presents a menu-driven, interactive GUI. In terms of architecture, as a technology, the project used freely available software tools and followed existing software conventions within the freeware community. By implementing and interacting with necessary software components such as cross-linking databases, resource-specific handlers, and servlets in this manner, open standards were exploited.

Using concept mapping in large introductory courses

For the C-TOOLS project, we recruited a cohort of over 1000 freshman and sophomore students enrolled in introductory science-major courses: Biology, Chemistry, and Physics, as well as non-major science courses: Introductory Biology and Geology. During class meetings, students learned how to use the web tools. The online software allows students to seamlessly create their concept map on an "easel" page, save it in a private "gallery," restore, revise and submit it to receive automatic scoring feedback. Outside of class students completed concept maps as an integral part of the course. Online concept map-based homework assignments varied widely from biology questions like "how do these organelles of the cell work?" to those in physics "where does the heat go in an oven?" and in geology of "trace the path of the water cycle." To complete an assignment students typically logged into the http://ctools.msu.edu/ website, clicked on their course assignment from a menu, and were presented with instructions and a map space seeded with approximately 10 concepts. The Concept Connector software allowed students to move concept words around, add additional concepts, organize hierarchy, and add linking words and lines. Exercises often challenged students to first construct a map individually, and submit it to the computer to receive visual feedback. They then could revise the map and resubmit. Finally, students often also worked with a partner to complete a final collaborative concept map.

Results

Study I: Testing Robograder's success rate when using WordNet, an internet thesaurus.

When an instructor creates a new concept map assignment online for the Concept Connector, if they wish to have Robograder give automatic feedback to students they must provide it with some "correct answers" or more specifically *expert propositions* in the form of a grading matrix. Instructors tend to develop an initial grading matrix by creating a spreadsheet with concept words listed both across the top and down the left side. In each spreadsheet cell that lies between two different concepts, several correct and incorrect words or phrases are provided. As a result, Robograder will know to mark something correct or incorrect with green halos or red X's, respectively, if it comes upon a student proposition that matches one in this grading matrix. While the matrix provided by the instructor is limited to the exact words or phrases inside it, Robograder can add to the matrix (amplify or supplement it) by accessing WordNet a very

sophisticated online thesaurus. It can grade synonyms of words in the matrix as if they are equivalent. In this study, we called new propositions which were derived purely by using WordNet, as "proposition derivatives" (Figure 2). The use of WordNet in this manner was previously suggested by Cañas et al (2003) and this study is exactly what was recommended, a test of how well this WordNet-based amplification approach actually works in the real world.

At the time of this WordNet study there were 35404 propositions available from Michigan State University's C-TOOLS server from student created maps. Each of these original propositions consisted of a starting concept, a linking word or phrase and a terminal concept (e.g. photosynthesis - fixes - CO_2). A random sample of 250 propositions was gathered and divided into 5 separate sets of 50 each. Manual assessment of propositions was done by hand without aid of electronic references or algorithms. Scorings for each proposition were: 1 (correct, e.g. "photosynthesis - needs - carbon dioxide"), X (incorrect, e.g. "DNA - translates - RNA"), 0 (ambiguous, e.g. "atom - is made of - neutron"), and S (structural violation, e.g. "ocans - evaporation – atmosphere"). Structural violations were for propositions with grammar problems such as spelling errors and linking phrases that do not contain a verb. Ambiguous scores were given to propositions that could only be scored as correct when viewed in a reasonably plausible context of surrounding propositions.

Examples of Synonyms, Antonyms, Troponyms and Synsets

Original Nuclear Membrane _ holds _ Nucleus (1) *Manual scorings shown ( )*
Synonym Nuclear Membrane _ contains _ Nucleus (1)
Troponym Nuclear Membrane _ superannuate _ Nucleus (X)
Antonym Plasma Membrane _ let go of _ Nucleus (X)

WordNet-derived synonyms, antonyms & troponyms were made 3 at a time:

3 synonyms of "holds" are: {{"contains", "carries"}, {"prevents"}}

           Synset #1 Synset #2

Synset #1 is valid. Synset #2 is invalid.
Nuclear Membrane contains Nucleus (1); Nuclear Membrane carries Nucleus (1)
Nuclear Membrane prevents Nucleus (X)

Figure 2: Examples of synonyms, antonyms, troponyms and synsets. In this study, we called new propositions which were derived purely by using WordNet, as "proposition derivatives." If the instructor's grading matrix indicated the word "holds" was correct, Robograder might assume the synonym "contains" was also correct and the antonym "let go of" would be incorrect. This is more often the case when evaluating student generated maps, especially so if the synonym comes from the same "synset" in WordNet as the original linking word.

Version 2.0 of the software database WordNet was used to generate "proposition derivatives" by making linking phrase substitutions with synonyms, antonyms or troponyms based on WordNet's thesaurus-like lexical capabilities (Figure 2). There were two criteria for the generation of proposition derivatives. First, derived propositions were made from linking phrases consisting of a single verb. Only 121 of the 250 original propositions met this single verb word criterion. Second, at minimum, the WordNet database had to have three available choices per lexical

relationship. An original proposition's linking verb must thus have at least nine derivatives (i.e. 3 antonyms, 3 troponyms, and 3 synonyms). Each triplet, as generated per lexical relationship (e.g. three antonyms), is called a "trio." With the single verb and triplet criteria, WordNet enabled us to construct 30 antonym derivatives, 243 troponym derivatives, and 234 synonym derivatives per proposition (Figure 2). Grading of proposition derivatives was delegated by the originating proposition sets. Graders A and B both graded derivative set 5. Grader A graded derivative sets 3 and 4. Grader B graded derivative sets 1 and 2.

The manual assessments of original and derivative propositions were scrutinized in order to both summarize and make insights into relationships that may concern automated strategies of assessment. Assessments of original and derived propositions were enumerated in order to show relative ratios of correctness, ambiguity, and grammatical errors. Trios were analyzed for fluctuations in correctness and incorrectness. In testing inter-rater reliability, Graders A and B had reproducible similarity to their scoring patterns as determined by the Kappa statistic (Cohen, 1960). The Kappa statistic ($\kappa = 0.552$) was calculated with $p_o = 0.720$ and $p_e = 0.374$ suggesting good reproducibility ($0.4 \leq \kappa \leq 0.75$). The level of significance for this degree of association is < 0.10. For the manual assessments of the 250 original propositions, 72% of the assessments between the two graders were identical (180 propositions). Opposite assessments of correctness (1 versus X) occurred 5.6% of the time. Remaining differences for the assessment of individual propositions were primarily attributable to issues unrelated to exacting qualifications of correctness. For example, 30 instances of disagreement involved only one grader assigning an S score and 26 instances of disagreement involved one grader cautiously assigning a 0 (ambiguous) score in contrast to 1 or X scorings. While our approach has statistically significant repeatability for scoring ratio properties and strong consistency for exacting qualifications of proposition correctness, further refinement would involve better synchronization between graders' approaches to assumptions of context and handling of grammatical logistics. When looking at the jointly graded WordNet derivative set ($n = 144$), the agreement between grader A and grader B was 70% ($p_o = 0.701$). The degree of association is just marginally reproducible based on $\kappa = 0.375$ and this reduction may be attributable to fewer shared contextual assumptions between graders due to loss of the original word choice. Scoring dynamics appear to be conserved; joint scorings for derivatives rise in agreement when considering just 1 and X scores, and the κ value does not suggest complete insignificance ($\alpha = 0.13$).

While synonyms and troponyms generated from the original linking word were sometimes found to be correct, antonym derivatives were always found to be incorrect. This is predictable and somewhat useful. The opposite to a "correct" answer logically should be an incorrect answer. While we predict it will not be a frequent occurrence for a student to use the exact antonym of a correct answer, including "incorrect" linking words in the grading data set has value. If WordNet can enable Robograder to increase the number of "incorrect" answers in its grading matrix, that will give it more opportunities to give precise negative feedback when invalid links are being used by students in some area of their concept map. Of 21 antonyms graded by grader A, 21 were graded as incorrect. Of the 18 antonyms graded by grader B, 18 were graded as incorrect.

Assessments for original, synonym-derived, and troponym-derived propositions are shown in Table 1 and encompass a range of assessment across all four grading categories (1, 0, X, and S). When the range of assessment is limited to 1 and X, grader A found 25.6% of synonym-derived propositions to be correct and 16.8% of troponym-derived propositions to be correct. For 1 and X scorings, grader B found 43.5% of synonym-derived propositions to be correct and 31.7% of troponym-derived propositions to be correct.

Table 1: Summary of manual assessment scores for original, synonym-derived and troponym-derived propositions.

Score	Original propositions		Synonym-derived propositions		Troponym-derived propositions	
	Grader A	Grader B	Grader A	Grader B	Grader A	Grader B
Correct	141	123	32	68	21	53
Incorrect	12	32	93	88	104	114
Ambiguous	33	13	16	0	22	0
Structural violation	64	82	0	0	0	1

The construction of trios involves random sampling from each WordNet-generated set of antonyms, synonyms, and troponyms. If conflicting meanings inside each set cause a general variation of proposition correctness, then clustering of correct or incorrect assessments within trios should not differ from a distribution of correct assessments that is random with respect to triplet structure. For the 57 synonym-derived trios assessed by grader A and the 81 synonym-derived trios assessed by grader B, the distributions showed no significant difference ($\chi^2 = 1.59$, $p = 0.66$ and $\chi^2 = 2.07$, $p = 0.56$ respectively). For the 54 troponym-derived trios assessed by grader A and the 71 troponym-derived trios assessed by grader B, the distributions also showed no significant difference ($\chi^2 = 2.64$, $p = 0.45$ and $\chi^2 = 5.57$, $p = 0.13$ respectively).

The general variability of correctness occurring within trios was investigated further by measuring how assessment score changes relate to similarities in meaning for derived proposition linking verbs. The WordNet database organizes lexical sets into subsets (termed "synsets") grouped together by similar meaning (Figure 2). Pairs of propositions occurring within trios were analyzed for having dissimilar correctness scores 1 and X, and for whether each proposition's linking verb was a member of the same synset. Shared synset membership for troponym derivatives occurred for 67% (grader A) and 49% (grader B) of all trio pairings that had an assessment score transition from 1 to X. Scoring transitions from 1 to X were next contrasted to within-trio proposition pairs where both propositions were assessed with a score of 1. Shared synset membership for troponym derivative pairs occurred for 100% (grader A) and 81% (grader B) of all such trio pairings that had a common assessment score of 1. Synonym derivatives were analyzed in similar fashion. Shared synset membership for synonym derivatives occurred for 15% (grader A) and 14% (grader B) of all trio pairings that had an assessment transition from 1 to X. Shared synset membership for synonym derivative pairs occurred for 27% (grader A) and

31% (grader B) of all such trio pairings that had a common assessment score of 1. Thus, for both troponyms and synonyms, membership of two verbs in the same synset greatly increased their chances of being correct.

Study II: Comparing human and software -based "holistic" scoring approaches

In automated grading, our ultimate goal is to follow the hierarchical scoring system developed by the Novak group (Novak & Gowin, 1984) yet since "Robograder" currently only gives visual feedback concerning the validity of the semantic relationship between linked words in a proposition, it can be used to evaluate non-hierarchical concept maps as well (like those developed by Kathleen Fisher et al). During the study period, automated scoring of student linking words successfully graded 26% of the user-made propositions existing on Michigan State University's C-TOOLS server. To increase this percentage we are studying student maps for interesting trends and testing new holistic content-independent software approaches for automated feedback.

In our first study on holistic scoring, a student concept map data set from an introductory biology course for non-majors was examined (Figure 3). Over the course of the semester all 76 students in the class completed four concept maps on different topics of biology (map 1-carbon cycle, map 2-mendelian genetics, map 3-natural selection, map 4-ecosystems). The instructor [human grader] scored each student's concept map by hand using their own holistic approach based on structure and word usage that led them to award each map a 0-5 point score. We then tested whether different algorithmic software approaches to score the same concept maps could mimic the human grader's results.

Figure 3 presents an analysis of concept map data from the biology course. It aligns the distribution of grades (0-5) given by the expert faculty to student concept maps made during a semester (top panel) with an analysis of most common "hub" concept words found in the student maps ("hub"=concept with most links; middle panel) and "holistic" grading strategies where software attempts to evaluate the same student maps via content independent approaches (bottom panels).

As a side note, the instructor predicted that the reduced student performance seen for MAP3 was based on complex interdependencies associated with the topic of "Natural Selection." The instructor also predicted that those students understanding certain critical concept words, as evidenced in MAP1 by choosing words such as "photosynthesis" to be the most highly interconnected hub, would score the highest on their concept maps. Our analysis of the data tested the hypotheses. The shift in grade distribution of maps (Figure 3 top panel, striped portions of bars) using the most popular "hub" word (Figure 3 identified in the middle panel) appears to support the instructor's prediction. If you replot that data with only maps that used "photosynthesis" as their mostly highly interconnected hub [in MAP1] as well as in the other

maps there is a subtle increase in score. Hence this is a simple proof of concept that C-TOOLS provides a well-curated data source with which to assess trends of classroom learning.
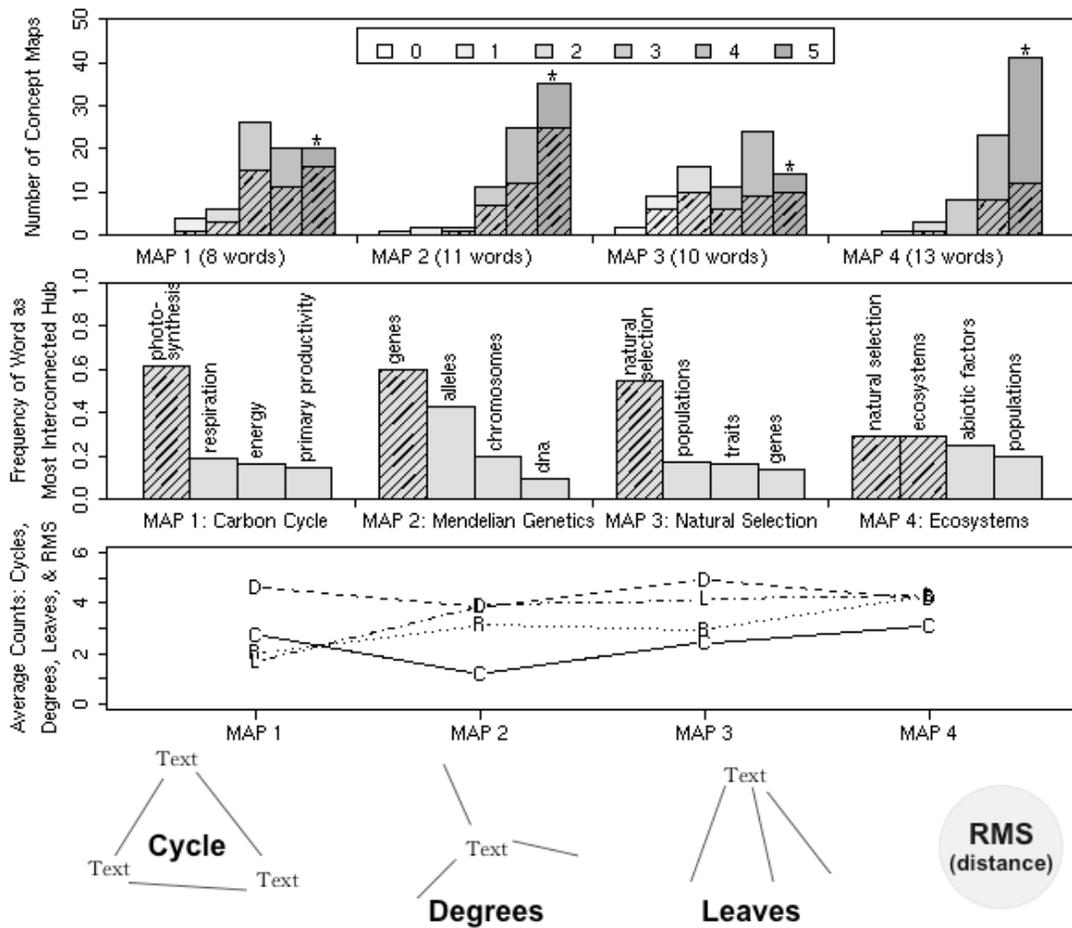
Figure 3. Human expert scoring of student maps from a non-majors biology course (top panel) and software analysis of trends in the map data (lower panels). Panel 1 (top) shows the distribution of scores (graded from 0 to 5) for each of 4 assignments given successively throughout a semester (n=76 students). The striped portions of the bars in panel 1 indicate the distribution of scores for maps that used the top "hub" concept word (for MAP1 this hub word was "photosynthesis," identified in Panel 2). Panel 3 shows trends in content-independent Gestalt scoring approaches used by software when applied to the same maps. These are the average values of 4 network topology measurements for the maps that scored a "5" (*) from each of the 4 assignments. Cycles="C", is the number of loops involving 3 or more concept words; Degrees="D", is the number of propositions connecting to a given concept word; Leaves="L", is the number of terminal ends in the concept map network; RMS, an indicator of non-branching chains within a concept map="R", is the root of the mean sum of squared distances between all concept word pairs within a concept map.

Beyond studying trends in student maps, we tested four holistic scoring approaches that score the student's concept map based solely on their networked structure (Figure 3, 4). Concept maps with a greater number of lateral connections, ie crosslinks should score higher on "cycles" or "RMS" while maps with many stratified levels of hierarchy could score higher on "degrees" or "leaves." Cycles is literally a count of the number of loops involving 3 or more concept words on a concept map while degrees is a count of the number of propositions connecting to a given

concept word. Leaves is the number of terminal ends in the concept map network. RMS is the root of the mean sum of squared distances between all concept word pairs within a concept map.
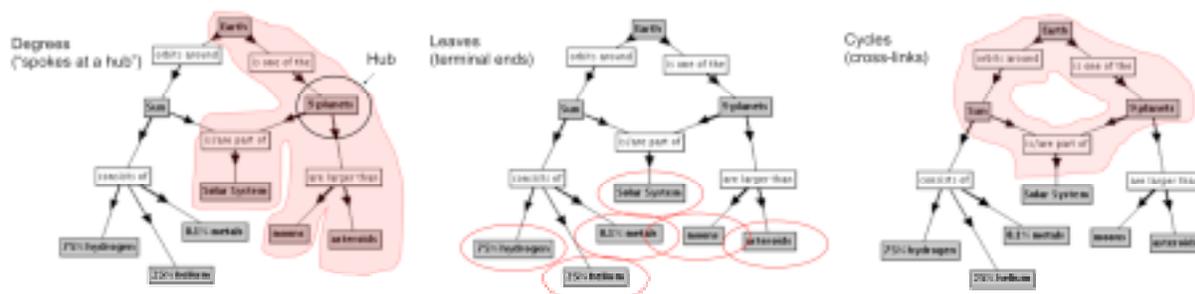
Figure 4. Four network topology measurements were studied as potential content-independent "holistic" methods to score concept maps (three examples shown). "Degrees" is the number of propositions connecting to a given concept word. "Leaves" is the number of terminal ends in the concept map network. "Cycles" is the number of loops involving 3 or more concept words. RMS (not represented) is the root of the mean sum of squared distances between all concept word pairs within a concept map and is an indicator of non-branching chains.

Automated grading approaches tested in this study were based on the network structure of the student concept maps. Methodologies using map network patterns related to hierarchy ("Leaves" and "Degrees") and cross-linking ("Cycles" and "RMS") were evaluated (Figure 4).

In the bottom panel of Figure 3, four automated scoring strategies were tested on student concept maps that received a score of 5. Interestingly, topology measurements termed "RMS" and "Leaves" correlated best with the human grader. The capacity to analyze and verify these predictions will grow in power with the accumulation of additional data and classroom-to-classroom comparisons. More studies of this sort may point to content-independent approaches that can be used automatically online to increase the number of propositions on a student concept map with which Robograder could give substantive positive and negative feedback.

Discussion

The C-TOOLS project stems from the combined activities of an interdisciplinary team of faculty from Michigan State University. This National Science Foundation-funded project developed a new assessment tool, the Concept Connector, consisting of a web-based, concept mapping Java applet with automatic scoring and feedback functionality. The Concept Connector tool is designed to enable students in large introductory science classes to visualize their thinking online and receive immediate formative feedback. Further details concerning the goals and methodology of C-TOOLS project have been previously published (Luckie, Batzli, Harrison & Ebert-May, 2003). In this study we focused on testing how successful instructor-independent software approaches were at grading student generated concept maps. These approaches can create additional "correct" linking words extrapolated from criterion maps generated by experts as well as help evaluate concept maps by content-independent "holistic" approaches.

The Concept Connector's flexible scoring system, based on tested grading schemes as well as instructor input, has enabled automatic and immediate online scoring of concept map homework. Criterion concept maps developed by instructors in the C-TOOLS project contain numerous expert-generated or "correct" propositions connecting two concept words together with a linking phrase. In this study a range of holistic algorithms as well as WordNet, an electronic lexical database and thesaurus, were used to test existing and potential automated methods of scoring to help supplement the finite set of answers provided by instructors.

WordNet Study

The results of Study I indicate that WordNet automated scoring approaches of concept maps, an approach that amplifies correctness across multiple synsets appeared to work on concept maps made by student users. Yet such an indiscriminating approach was found faulty when applied to randomly generated sets of synonyms and troponyms. Thus, the data supports that new "derived" synonyms and troponyms can be used as sets for further identifying both "correct" and "incorrect" propositions in student maps by using the original instructor provided correct and incorrect answers, respectively. Antonyms are potentially useful in their ability to derive "incorrect" answers from those provided by the instructor to be correct. Overall the use of synonyms from WordNet enabled 971 of the 9211 propositions to be evaluated when the existing instructor generated grading matrices would not otherwise make an assessment. Currently, Robograder indiscriminately accepts linking phrase synonyms independent of frequency and word sense. In theory amplifying grading rubrics by using all available synonyms should introduce errors into rubrics since multiple and conflicting meanings often exist. As mentioned earlier, in our findings from this study, visual examinations of automatically graded maps indicated few false positives or false negatives despite Robograder's treatment of all synonyms of a word as interchangeably equivalent. One explanation for the observed success of indiscriminate acceptance of synonyms is that users may more likely choose words within a relevant set of synonyms (known in WordNet as a "synset"). Thus, when developing concept maps, students appear to be less inclined to just randomly pick synonyms but rather they use semantically plausible word choices.

Although it may appear from the data provided that there is only a 10% gain by using WordNet's synonyms for automatic grading, this is only from the standpoint of automating the assessment at the proposition level. At the larger concept map level, there are highly interconnected concept words that follow a pattern of classroom consensus and also correspond to student performance (Luckie, Harrison, & Ebert-May, 2004). Better understanding of the linking words around major hubs would aid us to analyze the formative dynamics of how users in a classroom interconnect concepts and, potentially, knowledge domains. Analysis and further improvements to Robograder^{TM} cannot just be limited to synset hierarchies of each individual linking word since there are content-dependent dynamics of semantic overlap that influence how words can sensibly connect to other words (Banerjee & Pedersen, 2003).

Holistic Scoring Study

Currently, instructor-provided and WordNet-supplemented grading matrices together enable Robograder's automated scoring of student linking words to successfully grade 26% of the user-made propositions existing on Michigan State University's C-TOOLS server. To increase this percentage we are pursued a case-study of student maps from a non-majors biology course for interesting trends and to test four new holistic content-independent software approaches for automated feedback.

As mentioned above, automated grading approaches tested in this study were based on the network structure of the student concept maps. Methodologies using map network patterns related to hierarchy (leaves and degrees) and cross-linking (cycles and RMS) were evaluated. Software-based and content-independent holistic scoring approaches that focused on topology measurements termed "RMS" and "Leaves" correlated best with the human grader's own holistic approaches. The capacity to analyze and verify these predictions will grow in power with the accumulation of additional data and classroom-to-classroom comparisons. Yet if only one topology measurement was found to have good validity and reliability in evaluating student concept maps, all maps would be able to get reliable semantic and topological automated and immediate feedback to students online.

More studies of this sort may point to content-independent approaches that can be used automatically online to also increase the number of propositions on a student concept map with which Robograder could give substantive positive and negative feedback. Visual models are common in science, as well as the use of more than one assay to collect data. We hope the Concept Connector and its feature Robograder will aid our colleagues in large introductory science classes to use an additional tool in teaching and assessment. Beyond the multiple choice exam and rote learning, challenging students to wrestle with the new ideas of science within the boundary of a concept map with responsive feedback may begin to help direct them toward expert knowledge and meaningful learning.

References

Ausubel, D. (1963). *The Psychology of Meaningful Verbal Learning*. Grune/Stratton. New York, NY.

Banerjee, S., & Pedersen, T. (2003) *Extended Gloss Overlaps as a Measure of Semantic Relatedness*. Paper presented at IJCAI 2003 – 18th International Joint Conference on Artificial Intelligence.

Bruner, J. (1960). *The Process of Education*. Harvard University Press. Cambridge, MA.

Cañas, A. J., Valerio, A., Lalinde-Pulido, J., Carvalho, M., & Arguedas, M. (2003). *Using WordNet for Word Sense Disambiguation to Support Concept Map Construction*. Paper

presented at SPIRE 2003 – 10th International Symposium on String Processing and Information Retrieval.

Casti, J. L. (1990). *Searching for certainty: what scientists can know*. New York, W. Morrow, 496 p.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Collins, A., Joseph, D. & Bielaczyc, K. (2004) Design research: Theoretical and methodological issues. *Journal of the Learning Sciences, 13*(1), 15–42.

Fellbaum, C. Ed. (1998). *WordNet – An Electronic Lexical Database,* MA: MIT Press.

Fisher, K. M. (2000). *SemNet software as an assessment tool*. In J.J. Mintzes, et al (eds.), Assessing science understanding: A human constructivist view. Academic Press. San Diego, CA.

Harrison, S. H., Wallace, J. L., Ebert-May, D., & Luckie, D. B. (2004). C-TOOLS automated grading for online concept maps works well with a little help from "WordNet" Paper presented at CMC 2004 – 1st International Conference on Concept Mapping.

Ihaka, R., & Gentleman R. (1996). R: A Language for Data Analysis and Graphics, *Journal of Computational and Graphical Statistics, 5*, 299-314.

Luckie, D. B., Batzli, J. M., Harrison, S., & Ebert-May, D. (2003). C-TOOLS: Concept-Connector Tools for Online Learning in Science. *International Journal of Learning 10*: 332-338.

Luckie, D., Harrison, S., & Ebert-May, D. (2004). *Introduction to C-TOOLS: Concept Mapping Tools for Online Learning*. Paper in review for CMC 2004 – 1st International Conference on Concept Mapping.

National Research Council. (1999). *Transforming Undergraduate Education in Science, Mathematics, Engineering, and Technology*. National Academy Press. Washington, DC

Novak, J. (1990). Concept Mapping: A Useful Tool for Science Education. *Journal of Research in Science Teaching, 27*(10), 937-949.

Novak, J. D., & Gowin., D. D. (1984). *Learning How to Learn*. Cambridge Press. New York, NY.

Pea, R., Tinker, R., Linn, M., Means, B., Bransford, J., Roschelle, J., His, S., Brophy, S., & Songer, N. (1999). Toward a learning technologies knowledge network. *ETR&D. 47*(2): 19-38.

Suter, L., & Frechtling, J. (2000). Guiding principles for mathematics and science education research methods. *NSF Report* 00-113.

Wittrock, M. C. (1992). Generative Learning Processes of the Brain. *Educational Psychologist, 27*(4), 531-541.