Insight into Student Thinking in STEM: Lessons Learned from Lexical Analysis of Student Writing

1 Introduction

Ideally, assessment data should provide information to instructors about their students' thinking so that instructors can design appropriate instructional interventions (Bransford, 2000; Von Glasersfeld, 1994). In large enrollment courses, typical of college-level introductory STEM courses, assessments are usually limited to multiple choice instruments that can be machine scored. However, there is evidence that students may correctly answer multiple-choice questions but still harbor misconceptions, or *conceptual barriers*, which seriously compromise their learning (Nehm & Schonfeld, 2008). As a result, multiple-choice instruments may not detect critical gaps in student understanding or prevailing patterns of thinking that could otherwise be addressed through appropriate instructional intervention.

Constructed response assessments, in which students have to use their own language to demonstrate their knowledge, can provide good insight into student thinking (Birenbaum & Tatsouka, 1987). Although widely seen as more reflective of student thinking, they are significantly challenging to execute in large-enrollment courses given existing resource constraints. At UNIVERSITY, we have successfully used computerized lexical analysis in studies of students' conceptual understanding in biology, chemistry and geology. These methods can be applied across large scales to support the improvement of STEM education by providing data to instructors that can drive instructional decisions. In this paper, we summarize our work to date and describe some of the lessons we have learned in the hope that others can benefit from our work and adopt these techniques.

2 Overview of Procedure

We are taking a two-stage approach to analyzing constructed responses. First, we use *lexical analysis* to extract key terms and concepts from student writing. We then use these terms and concepts as variables for *statistical classification* techniques to predict expert ratings of student responses. This combination of analytic techniques provides insights into student thinking by categorizing written responses based on commonly used terms and phrases.

2.1 Computerized Lexical Analysis

SPSS Text Analysis for Surveys (SPSS, 2006) is lexical analysis software designed to analyze and categorize open-ended responses on surveys. The software classifies written text into categories based on the appearance of specific terms (words and phrases) it detects using *libraries* to extract all recognizable terms. Graphical tools in the software enable a variety of ways for representing the categorized data that facilitate evaluation of the responses. Although the software has several standard libraries of common terms, they do not recognize most of the technical lexicon of scientific disciplines. In our work, we have created several custom lexical libraries containing biological terms, as well as synonyms, abbreviations, and spelling variations and misspellings (AUTHORS, 2008).

One challenge for students learning cellular biology is that they often fail to account correctly for matter and energy transformations, as in cellular metabolism (Wilson et al., 2006). To examine this, we created constructed response items asking students to trace carbon backwards during cellular respiration (AUTHORS, 2008). After we created biology libraries and categories, the software was able to correctly categorize over 90% of the student responses

without human intervention. Lexical analysis allows us to categorize all compounds and processes in the responses to find patterns in student thinking.

Such insight into student understanding would be difficult to achieve if all student responses were manually evaluated. Tools such as lexical analysis software facilitate the processing of large numbers of student responses. This is essential for discerning broad patterns in student thinking and testing the robustness of those patterns by comparing student responses across assessments aligned to test similar concepts. Not only does this represent a powerful method for analyzing the outcomes of instructional change, it offers the potential for delivering a critical component of reformed science teaching – near real-time feedback that informs instructors about their students' learning.

2.2 Statistical Classification Techniques

While lexical analysis can provide formative feedback to instructors about common student conceptions, it is important to understand the relationships among the concepts identified by the lexical analysis that are indicative of expert understanding in a discipline. Humans evaluate constructed responses using a variety of *rubrics* that are designed to assess these relationships. We are combining the results of our lexical analyses with statistical classification techniques to predict how expert human raters would apply rubrics to evaluate student writing.

For example, in our research on acid/base chemistry of biological functional groups (AUTHORS, 2009), students were asked to *give an explanation of a strong acid* and then asked to *give an explanation of a weak acid*. A subset of student responses was scored using a 3-bin rubric by two experts (inter-rater reliability: intraclass correlation = 0.95). Discriminant analysis identified 14 of the lexical analysis categories that were the most useful for predicting expert ratings of the student responses. These categories were well aligned with the expert rubric, providing good evidence of content validity. The function classified 83.8% of all cases correctly ($p < .001$); we would expect to classify 33% of the cases correctly. When inter-rater reliability is calculated between the experts and the computer predictions, the intra-class correlation is 0.882 ($p < .001$).

3 Improving The Utility And Efficiency Of Computerized Lexical Analysis

Based on our analysis of constructed response items, we have: 1) learned how to structure questions so that responses are better suited for lexical analyses; 2) learned effective ways to build custom discipline-specific lexical libraries; 3) gained insights into optimal numbers and specificity of categories; and 4) learned about optimizing the granularity of the classification rubrics used to rate student responses.

3.1 Question Structure

In several areas of study, we are now in the third and fourth iteration of response gathering for particular questions. This iterative process of using student responses to refine the question has allowed us to gather data better suited to lexical analysis. One key lesson is that multiple response boxes (in which students can enter text) are very useful for questions with multiple components. For example, one original prompt investigating student understanding of general chemistry asked students to "Explain the difference between a weak and strong acid." Students were presented with only a single text box in which to respond. Our analysis of responses to this question quickly found that the majority of students were defining both strong and weak acids, instead of keeping their answer focused on the difference between them. This

complicated the lexical analysis as the software cannot determine whether the extracted chemistry terms were being applied for a description of strong or weak acids or both. For example, here are two student responses:

"When mixed into a solution, weak acids will completely dissociate, where as when strong acids are mixed into a solution, they will not be affected."

"A weak acid is one that doesn't dissociate completely and a strong acid dissociates completely in a solution."

Obviously, these two students are describing quite different behaviors for strong and weak acids in solution. However, because both responses use similar extracted terms ("dissociate", "not/doesn't", "completely") the software cannot distinguish between these responses, which can pose a significant problem for evaluating student answers.

3.1.1 Question Wording

Analysis of student responses has prompted us to re-word questions to make them more amenable for lexical analysis. A critical lesson is that any word in the stem of the question will most likely be repeated by students in their response. One way of forming appropriate questions is to craft an "ideal" answer that an instructor would like/expect to see and identify the critical terms in the response. One can then devise a question that prompts such a response without using any (or as few as possible) of the identified critical terms. Decisions about which scientific terminology to use and accept should also be made in advance. For example, in one question about free energy, the terms "spontaneous" and "non-spontaneous" appeared in the question stem. However, it was decided that the terms "exergonic" or "endergonic" reactions (the chemical term equivalents of spontaneous and non-spontaneous) were meaningful and should be categorized.

3.1.2 Response Length

We have performed lexical analysis on a variety of responses from a range of questions. Expected responses to these questions range from a single word (or molecular formula) to several sentences in length (AUTHORS, 2009). In general, responses should be of sufficient length (but not too long) to provide enough extracted terms that allow a more accurate categorization. Responses that are too long will include too many unrelated terms. This leads to categorization of too many responses in a given category, which reduces their discrimination for classification purposes. However, responses that are too short may only be categorized in one (or two) categories. This is problematic if an instructor is interested in the connections students are making between concepts (and thereby categories).

3.2 Building custom libraries

We have learned two important lessons about building custom libraries. First, it is easier if misspellings have been corrected, before analyzing the data. Second, when verbs are added into a library, it makes the extraction process more efficient if their inflections are also included, so that when the new extraction of terms is performed, all inflected forms of the words will be extracted. Verbs are not extracted by the software by default, so they have to be included in custom libraries. For example, if we are interested in extracting the verb *evolve*, we may decide that will be the term extracted and include as synonyms *evolves*, *evolving*, and *evolved.*

3.3 Number and specificity of categories

One key task of text analysis is the creation of *categories*. The software offers two options, one based on frequency of occurrence of terms or types, and the other based on linguistic analysis using semantic networks, term co-occurrence and term inclusion and exclusion. We have found that the combination of the two procedures provides the best result. We begin creating categories guided by an *expert answer*, which is used to anticipate possible responses that will be given by students. We then refine the categories based on samples of student responses. Part of this process is deciding on the level of granularity for categories. Categories should be specific enough to identify conceptions at an appropriate level to distinguish among concepts typically held by the student population of interest. Although there is no rule about how many categories should be created, we have tried to keep it fewer than 40. Fine grained categories can be easily collapsed, if required by further analysis, while broad categories are more difficult to disaggregate. For instance, in our metabolism project (AUTHORS, 2008), where we asked students to trace matter and energy during cell respiration, we created 25 categories for the component of the question asking how the carbon got into the CO_2 molecule, including relevant processes in cell respiration, such as respiration, Krebs cycle, glycolysis, photosynthesis, etc. These processes can be easily collapsed in a *metabolism* category if a reduction of variables is required by the statistical analysis. Finally, we have learned that in any *de novo* set of data analyzed by existing custom libraries and categories, typically 10% or less of student responses are *uncategorized* by the lexical analysis software. However, our experience has shown that most uncategorized responses are irrelevant to the question posed and therefore add little to our understanding of the class performance.

3.4 Granularity of expert rubrics

Much as we are learning about optimizing the granularity of categories in lexical analysis, our results also highlight the challenges of expert rating of student writing. It is often a lengthy process to develop the rubrics and train raters to ensure acceptable inter-rater reliability. Lexical analysis can help provide additional data for the iterative nature of this process. In addition to raters discussing agreements and disagreements about scoring, using the categories derived by lexical analysis for statistical classification can highlight areas where the rubrics may be vague, or unnecessarily specific. For example, when initially created a scoring rubric for student explanations of the differences between strong and week acids, we used a 4-level rubric for the raters. However, we discovered that the statistical classification was least stable between the middle two categories. Further discussion among the raters revealed that they did not believe that the distinctions between levels 2 and 3 were that clear. By reducing the number of categories, we could improve both the interrater reliability and the statistical classification accuracy. When using these techniques for formative assessment, often a binary (generally correct, generally incorrect) rubric may be sufficient for instructors to gauge the overall understanding of a concept to determine instructional intervention.

4 Conclusion

Improving STEM education requires timely and informative assessment of student understanding. While constructed responses can provide good insight into student thinking, they are generally too labor intensive to be used in large enrollment courses. Computerized lexical analysis holds the promise to support using constructed response assessments in large enrollment courses. However, application of these techniques has only recently become feasible. By

sharing lessons learned and evolving best practices, we can collectively move research and understanding of student thinking forward.

5 References

AUTHORS. (2008, March 30 - April 2). *Understanding undergraduate students' conceptions in science: Using lexical analysis software to analyze students' constructed responses in biology.* Paper presented at the NARST 2008 Annual International Conference, Baltimore, MD.

AUTHORS. (2009, April 17-21). *Using lexical analysis software to understand student knowledge transfer between chemistry and biology.* Paper presented at the National Association of Research in Science Teaching Annual Conference, Garden Grove, CA.

AUTHORS. (2009, October 18-21). *Beyond multiple choice exams: Using computerized lexical analysis to understand students' conceptual reasoning in STEM disciplines* Paper presented at the Frontiers in Education, San Antonio, TX.

Birenbaum, M., & Tatsouka, K. K. (1987). Open-ended versus multiple-choice response formats - It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*, 329-341.

Bransford, J. (Ed.). (2000). *How people learn brain, mind, experience, and school* (Expanded ed.). Washington, D.C.: National Academy Press.

Nehm, R. H., & Schonfeld, I. S. (2008). Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching, 45*(10), 1131-1160.

SPSS. (2006). *SPSS Text analysis for surveys 2.0 user's guide*. Chicago, IL: SPSS, Inc.

Von Glasersfeld, E. (1994). A constructivist approach to teaching. In L. P. Steffe & J. Gale (Eds.), *Constructivism in Education* (pp. 3-15). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wilson, C., Anderson, C. W., Heidemann, M., Long, T., Merrill, J., Merritt, B., et al. (2006). Assessing students' ability to trace matter in dynamic systems in cell biology. *Cell Biology Education, 5*, 323-331.