USING LEXICAL ANALYSIS SOFTWARE TO UNDERSTAND STUDENT KNOWLEDGE TRANSFER BETWEEN CHEMISTRY AND BIOLOGY

Time and resources often prevent instructors from using constructed response assessments in large undergraduate science. We investigate the utility of using lexical analysis software to categorize student responses and uncover undergraduate student misconceptions in chemistry and biology. Students were assigned a question set consisting of four questions relating to free energy and acid/base chemistry. Student responses were analyzed using SPSS Text Analysis for Surveys, using a custom library of science-related terms. The resulting analyses of student responses suggest concept connections made by students as well as student difficulties in understanding of these topics. Only 37 out of 158 students linked reaction spontaneity with thermodynamics. Student responses for two acid/base questions were rated using a scoring rubric by two independent scorers. Analysis of this question set showed student deficiencies in predicting pH behavior of functional groups in biology. After this scoring was complete, discriminant analysis was used to create classification functions that could predict human expert scores for the two acid/base questions with 83.8% and 77.0% accuracy ($p < .000$). This study suggests that computerized lexical analysis may be useful for automatically categorizing large numbers of student open-ended responses.

Kevin C. Haudek, Michigan State University
Rosa A. Moscarella, Michigan State University
Mark Urban-Lurain, Michigan State University
John E. Merrill, Michigan State University
Ryan D. Sweeder, Michigan State University
Gail Richmond, Michigan State University

Acknowledgement: This material is based upon work supported by the National Science Foundation under award 0736952. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

Introduction

Recently, the NRC outlined a vision for the future of biological science education which includes recommendations for explicitly linking chemistry and biology education. (NRC, 2002). This call for integration of mathematics and physical sciences with biology was also echoed by college instructors (Bialek & Botstein, 2004). Although there have been efforts to couple chemistry and biology during laboratory exercises and attempts at combining courses and/or curricula (Reingold, 2004; Wenzel, 2006) there have been few studies on studying students' conceptual transfer between these disciplines and any existing barriers to this coupling chemistry and biology. However, there are some reports that indicate potential problems that students have, in concepts such as free energy, oxidation-reduction and equilibrium (Schwartz & Serie, 2001). Other interesting studies have followed the student learning progression in college chemistry (Claesgens, Scalise, Wilson, & Stacy, 2009). Such studies may help identify principles that are transferred by students to other STEM disciplines, along with the students who have this ability. If curricula integration is to occur, it seems critical to understand the concepts and principles where students currently have difficulty transferring knowledge between disciplines.

Based on the assumption that understanding biological systems requires some knowledge of chemical principles, many institutions set general chemistry as a pre-requisite for general biology. However, in our introductory biology course we have encountered students that still fail to apply some basic chemistry concepts, such as conservation of matter (Parker et al., 2007). We are seeking tools that not only assess students' conceptions in these areas, but also that provide evidence for determining the effectiveness of subsequent instructional interventions.

Assessment

Assessment is a key component of engaging student learning with understanding (Bransford, Brown & Cocking, 1999). Assessment, especially formative assessment, becomes a central point in student learning because it enables the teacher to follow student progress and to make adjustments to instructional strategies as needed. It is then imperative that effective assessments should help detect student problems in understanding, rather than just evaluating whether a given response is right or wrong.

However, detecting the nature of student problems with multiple choice questions is often difficult, because they do not reveal students' reasoning. One way we choose to investigate student understanding and reasoning is by using *constructed response* (also referred to as open-response or essay) questions focused on biologically relevant chemistry topics. *Constructed response* assessments require students to respond to questions using their own language, allowing the potential to reveal misunderstanding and conceptual barriers (Birenbaum & Tatsouka, 1987). However, the large enrollment numbers of many undergraduate introductory science courses often discourages the use and evaluation of large numbers of open-ended response questions. Instructors often must resort to computer-graded multiple-choice assignments and exams for simplicity. It is sometimes suggested that sampling of responses to an assignment in large classes can provide the necessary feedback to instructors; however such sampling is unlikely to provide an accurate depiction of the whole class, as a few exceptionally good (or poor) responses may alter the instructors' perception of the class's understanding. Rapid

reading and grading of these responses by the instructor may also lead to errors in scoring or missing important conceptual connections made by students. To counter these problems, we are using lexical analysis software to analyze and evaluate student responses to provide rapid feedback for the instructor. We believe this lexical analysis software provides several key advantages to the evaluation of formative constructed responses assessments.

Previous Work

In previous work, we used lexical analysis software to analyze students' constructed responses in an introductory biology class (Moscarella et al., 2008). The instructors of that class focused on model-based reasoning to trace matter and energy in biological systems, which could be assessed by asking students to trace a phenomenon backwards. We have also used lexical analysis software to identify *conceptual barriers* through the analysis of students' open-ended responses (Richmond, Urban-Lurain, Parker, Merrill & Merritt, 2008). These previous results suggest that lexical analysis software can help instructors identify students' misconceptions or other difficulties in constructing new knowledge through the analysis of the vocabulary they use in their responses.

In this study we extend this work to evaluate students' understanding of the basic chemistry that may be related to conceptual problems students have in cellular and molecular biology. Further, we attempt to use lexical analysis software to evaluate students' understanding of important biological chemistry topics and use the results of the lexical analysis to predict how experts would score the student responses.

Methods

Question Sets

During the spring and fall terms of 2008, students from an introductory Cell and Molecules biology class were asked to answer an online question set (consisting of a total of four (spring) or six (fall) questions) for homework credit (approximately 0.01% of course grade per question. Students were given full credit for any genuine effort at responding, whether correct or not). The question sets were designed to address topics common to the introductory chemistry and biology courses. For this analysis, we focused on four questions, further divided into two broad topics containing two questions each. The questions analyzed here focused on two major underlying themes of the courses; energy (free energy) and matter (acid/base chemistry). Both sets contained a standard definition-type question and one question relating a chemical concept to a biological system. All students were asked to answer the acid/base chemistry questions, while students were randomly assigned to different free energy questions. Responses were collected in an online course management system for analysis. Questions with multiple parts were presented as a single page with multiple response boxes to students. The question sets were the following:

(A) Free energy:

1. Is it possible for a chemical reaction that is non-spontaneous to become spontaneous at a different temperature? Why or why not?

2. Polymerization reactions in which amino acids are synthesized into a peptide chain are endergonic, yet this reaction occurs continuously in your cells. Suggest a mechanism by which cells can drive this non-spontaneous reaction.

(B) Acid/base:

3a. Give an explanation of a strong acid.
3b. Give an explanation of a weak acid.

4a. Consider two small organic molecules in the cytoplasm of a cell, one with a hydroxyl group (-OH) and the other with an amino group (-NH2). Which of these small molecules (either, both or neither) is most likely to have an impact on the cytoplasmic pH?

A. Compound with amino group (correct response)
B. Compound with hydroxyl group
C. Both
D. Neither

4b. Explain your answer for the above question.

Scoring

A total of 158 responses for question one and 153 responses to question two were collected. A total of 382 student responses were collected for questions three and four, of which a subset was used for further analysis (see below). The question sets were given in the second half of the term, so that students had seen both topics in both chemistry and biology courses.

One goal is to determine if computerized lexical analysis can be used to develop reliable scoring functions for student responses. For responses to question set B, two members of our research group, with expertise in Chemistry and Biology, independently evaluated a subset of the answers and rated them according to the following rubrics:

For questions 3a & b:

Level 1: Correct definitions of strong and weak acids (i.e., Strong acids ionize completely in solution, weak acids only partially ionize in solution)
Level 2: Correct definitions with minor errors in additional facts or reasoning (i.e., Strong acids ionize completely in solution and have very low pHs. Weak acids don't dissociate completely in water)
Level 3: Correct definition for one acid, incorrect for the other (i.e., Strong acids ionize completely, weak acids do not ionize)
Level 4: Totally incorrect/irrelevant response for both acids (i.e., Strong acids have a lower pH, weak acids have a higher pH)

For question 4:

Level 1: Correct description of basic nature of amino (i.e., Amino groups act as weak bases in a cell)

Level 2: Partially correct explanation/irrelevant to question (i.e., Amino groups are molecules with a higher pH level than the hydroxyl.)

Level 3: Totally incorrect explanation (i.e., Amino group has two H atoms it may give up, but hydroxyl has only one OH molecule it may give up.)

For the rating of the responses for questions 3a & b, both expert evaluators gave a single score to both explanations. For question 4b, only students who selected the correct multiple choice response in question 4a (***compound with amino group***) had their explanations evaluated further using the above rubric (a total of 129 responses).

For questions 3a and b, 150 of the collected responses were independently evaluated and used to help train the lexical analysis software. Further, the software was then used to predict the remaining student responses, of which another 101 responses were scored by one of the evaluators.

Computerized lexical analysis

For the lexical analysis, we used SPSS Text Analysis for Surveys v. 2.1 (SPSS-TAFS, SPSS Inc., 2007), which allows the analysis of open-ended questions through the classification of responses into categories. The data are imported into the software, which extracts key terms from the responses that will be used to categorize them. SPSS-TAFS has two ways to create categories: by linguistic analysis using semantic networks, term co-occurrence and term inclusion and exclusion; and by term or type frequency. To create the categories for this project, we combined both strategies, based on an "expert answer" for each question. We took advantage of a custom library with biological terms we previously built in the software (Moscarella et al., 2008). Once the data were categorized, the individual responses and the associated categorized data were exported for subsequent data analysis. Responses could include more than one category if it contained multiple relevant terms.

Analyses and findings

A category (denoted by *italics* in this text) in lexical analysis is the name of a group of similar terms and synonyms that is defined by either the software and/or the user. Categories can also contain functions (and, or, not) that allow the user to specify very particular phrases or combinations of terms to include in that category. For example, in the results of question 1 (see Table 1) the category *Heat* contains the terms: "heat", "extra heat", "input of heat", "delta q" and the function: heat energy, which is designed to select students who use the terms "heat" ***and*** "energy" in their response.

Reaction spontaneity question

Our lexical analysis resulted in a total of 19 categories, including the categories *Yes*, *No* and *Don't Know*. This was in order to classify student responses to the first part of the question (Is it possible?) separate from categorization of their explanations (Why

or why not?). Category results showed that student overwhelming answered affirmatively to the question; 122 responses were *Yes* and a total of 22 were *No* or *Don't Know*. All responses that were not categorized by the software as *Yes, No* or *Don't Know* were manually evaluated and responses were of the affirmative nature. The remaining categories of student explanations to the question (minus *Yes, No* or *Don't Know* categories) are shown in Table 1. All student explanations were analyzed together using the software, regardless of their original Yes/No answers to the question. The category with the most student responses was *Increase temperature* (95 responses), which contained a function to select only responses with the terms "temperature" ***and*** "increase". It was important not to count the term "temperature" alone, since this word appears in the question stem. There is still some uncertainty whether the word "temperature" in the stem remains too big of a cue to prompt student response. We also note that some responses were classified as *Heat* (36 responses). Surprising to us, although many students got this question correct (~85%), very few responses invoked explanations involving free energy or components of free energy, *enthalpy* (12 responses) and *entropy* (26 responses). *Free energy* (20 responses) or the related *exergonic* (5 responses) and *endergonic* (4 responses) terms were used only sparingly. In fact, by far more students gave responses in the category *activation energy* (32 responses) in attempting an explanation. This may be due, in part, to the relative placement of the question set in a biology course shortly after chapters relating to enzymes and cellular metabolism.

Table 1: Categorized student responses to question 1. *Note*. Only categories pertaining to the explanation included in the student response are shown. Total percentages are greater than 100% because any response may have multiple categories.

Category	Number of responses	% of total responses
Increase temperature	95	60.1
Compounds	50	31.6
Heat	36	22.8
Activation energy	32	20.3
Energy	28	17.7
Entropy	26	16.5
Free energy	20	12.7
Quicker / Kinetics	13	8.2
Enthalpy	12	7.6
Types of energy	11	7.0
Catalyst and enzyme	10	6.3
Change	8	5.1
Exergonic reaction	5	3.2
Endergonic reaction	4	2.5
Bond	3	1.9
Break	2	1.3

We have additional data for this question from another semester. Although the absolute numbers differ, the category trends remain remarkably similar. The large majority of students (~ 70%) agree that a reaction can change spontaneity at different

temperatures. Their response patterns are remarkably similar as well; the most popular category remains *increase temperature* (88/190). As before, *free energy* (24/190) and *entropy* (13/190) terms are not commonly used in student explanations.

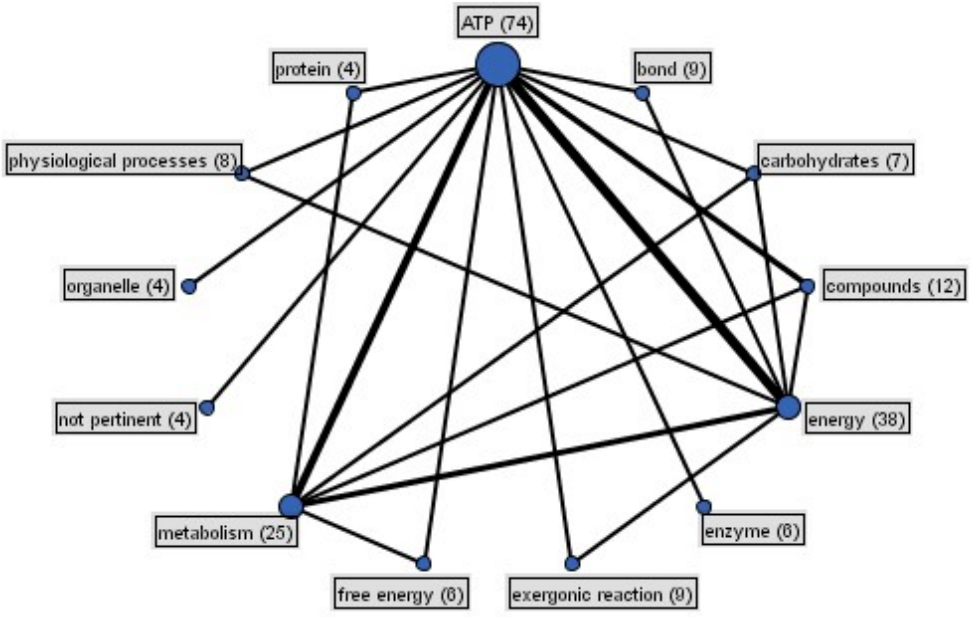
Polymerization question

Our lexical analysis of the polymerization question produced 23 categories. Seventy-four out of 153 responses correctly indicated ATP as a potential solution to the energy needs of a cell. Sixty-two responses contained the category *Energy*, showing that these students understood the need for energy input.

Figure 1 maps the relationships among categories for the 74 student responses that included ATP. The relative sizes of the nodes show the frequencies of the categories and the thickness of the lines represent the number of shared responses between two nodes. For simplicity, only categories sharing four or more responses are shown.

Of the 74 responses that included ATP, 38 of them were used terms that were categorized as *Energy* and 25 used terms categorized as *Metabolism*. Note that the largest number of responses share these three categories (ATP, Energy and Metabolism). This suggests that at least some portion of students attempt to explain the role (Energy) or production (Metabolism) of ATP inside the cell, even when not explicitly prompted to do so. At this level of analysis, it is impossible to tell if the students could provide a chemical explanation of how ATP acts as an energy management molecule. However, the observation that students attempt to connect these concepts (even at a qualitative level) should be encouraging for an instructor.

Figure 1. Category web showing the relationships among categories included in students' responses that include ATP.



Strong and weak acid question

For question three (strong/weak acid), a total of 150 student responses were independently scored by two members of the research team. Analysis of the raters' scoring of the student responses showed very good inter-rater agreement. On the

strong/weak acid question, Cronbach's Alpha was .973 and the single measure interclass correlation was .948 ($p < .000$). The two raters were in absolute agreement on 136 of the 150 responses (see Table 2.) Only student responses which were scored at the same level by both raters were used for further analysis.

Table 2: Number of student responses to strong/weak acid question scored at each rubric level for which raters were in complete agreement.

Level				
1	2	3	4	Total
91	9	18	18	136

The lexical analysis generated 27 categories for the strong and weak acid explanations. These categories included relevant terms that can potentially reveal students' understanding of acid-base principles applied to biological systems, such as: pH, ionization, conjugation, etc. We used the lexical classifications of each students' answers as the independent variables in a discriminant analysis (Spicer, 2005), with the expert classification of the student answers as the dependent variable. Stepwise analysis resulted in 12 of the categories being selected for prediction (Table 3). The three discriminant functions had good classification accuracy (Wilks Lambda = .072, Chi-square 335.01, df = 36, $p < .000$). Table 3 shows the resulting structure matrix containing correlations for each of the categories on the three functions. These are similar to loadings in factor analysis and can be conceived of as the correlation between the categories and the derived functions.

Table 3: Structure Matrix resulting from analysis of strong/weak acid question.

Categories		Function		
		1[a]	2[b]	3[c]
Strong acid explanation:	hydrolysis	-.111	.288	-.248
	donate	-.071	.143	.257
	solution	.323	.237	-.178
	hydrogen	-.062	.101	.187
	not pertinent	-.086	-.132	-.016
	conjugate	-.060	.156	-.134
	ionization	.446	.056	-.417
Weak acid explanation:	incomplete dissociation	.624	-.094	.222
	not dissociate	-.157	.354	.279
	solution	.294	.056	-.269
	not pertinent	-.122	.106	-.168
	electrolytes	.011	.031	.424

[a] Function 1 accounts for 69.9% of the variance
[b] Function 2 accounts for 25.2% of the variance
[c] Function 3 accounts for 5.0% of the variance

Analysis of these categories can be very informative and reassures that the statistical prediction model is using reasonable terms aligned with the expert rubric. For instance, in this analysis the purpose of function 1 is to discriminate bin 1 responses from all other bins. The categories with highest correlation to this function are *strong acid ionization* and *weak acid incomplete dissociation* (see Table 3). These are two critical term categories for students to use in their understanding of strong and weak acids. Also note the largest negative correlations for function 1 are categories (*not dissociate, not pertinent*) exemplary student response would ***not*** contain. In this case, function 1 appears to predict students' ability to correctly talk about ionization states of strong and weak acids. Further analysis of the classification functions shows that function 2 separates responses from bins 3 and 4. Again, note the strong positive correlation of some improper categories (i.e. *weak acid not dissociate, strong acid hydrolysis*) with this function (see Table 3). Students using these types of terms do not have a complete understanding of acid/base chemistry and would end up with a lower score.

Perhaps equally as informative are the categories that were not used in the prediction model. Categories not used may be due to the few number of responses or correlation between the category not used with one category that was used. As an example, both categories *weak acid donate* and *weak acid hydrogen* were not selected as useful variables in the discriminant analysis. However, this is most certainly due to high correlation between these categories and the *donate* and *hydrogen* categories in the *strong acid*. That is, if a student used donate or hydrogen in response to a strong acid, they were very likely to use the same term again in describing the weak acid, so that use of these categories again in the weak acid description provided no better predictive power of student responses. Interestingly, the category *ions* was strongly correlated to itself in both the strong and weak explanations. However, it was not used in the analysis in either case, suggesting that the term "ions" is not a good predictor of student understanding.

To test the utility of the classification function, we used a cross-validation classification in which each case is classified by the functions derived from all cases other than that case. We used prior probabilities of group membership based on group size, though the results with equal prior probabilities are very similar. The results are shown in Table 4. Correctly classified cases are shown in bold on the diagonal. The function classified 83.8% of the cases correctly; by chance, we would expect to classify 25% of the cases correctly.

Table 4: Classification percentages of cross-validated student responses for acid question classified at each level

	Computer Predicted Rating			
Expert Rating	1	2	3	4
1	**93.4**	3.3	3.3	0.0
2	33.3	**33.3**	22.2	11.1
3	5.6	5.6	**66.7**	33.3
4	11.1	0.0	11.1	**77.8**

Note that, most cases that were incorrectly classified were still within one category of the expert raters. The worst cases were the level 2 responses, where only 33.3% were correctly classified. This is likely due to that fact that only 9 of the original student responses were rated as level 2 by the experts. The low number of responses in a given rubric level makes it more difficult for the software to uncover patterns in the response categories. Therefore, the fewer response examples at a certain level, the more difficult it is for the software to correctly predict new responses at that level.

To further test the classification function, we had one of the experts rate an additional 101 student responses and compared those ratings with the predicted ratings as shown in Table 5. These data were not used to derive the discriminant functions in the initial analysis. The discriminant functions correctly classified 80.2% of these data. Again, the poorest result was for the responses rated as level 2 by the expert as this was the category with the smallest number of original responses. This suggests that it may be reasonable to collapse levels 2 and 3 to form a three level rating scheme.

Table 5: Percent of second sample of student responses for acid question classified at each level

	Computer Predicted Rating			
Expert Rating	1	2	3	4
1	**85.0**	8.3	0.0	6.7
2	50.0	**16.7**	0.0	33.3
3	16.7	8.3	**75.0**	0.0
4	0.0	4.3	8.7	**87.0**

Functional group pH question

For question set four, a total of 129 student responses were independently scored with the three level rubric (described in the Methods section) by two members of the research team. Analysis of the raters' scoring of the student responses showed very good inter-rater agreement. The Cronbach's Alpha was .961 and the single measure interclass correlation was .925 ($p < .000$). The two raters were in absolute agreement on 113 of the 129 responses (see Table 6.) Only student responses which were scored at the same level by both raters were used for further analysis.

Table 6: Number of student responses to amino question scored at each rubric level for which raters were in complete agreement.

Level			
1	2	3	Total
41	14	58	113

The lexical analysis generated 29 categories for the amino explanations. These categories included relevant terms that can potentially reveal students' understanding of acid-base principles applied to biological systems, such as: pH, ionization, conjugation, etc. We used the lexical classifications of each students' answers as the independent variables in a discriminant analysis (Spicer, 2005), with the expert classification of the student answers

as the dependent variable. Stepwise analysis resulted in six of the categories being selected for prediction (Table 7). The two discriminant functions had good classification accuracy (Wilks Lambda = .284, Chi-square 135.185, df = 12, $p < .000$). Table 7 shows the resulting structure matrix for each of the categories on the two functions.

For question 4b, we note that the highest coefficients for function 1 are *base/basic, accept hydrogen* (a property of a base) and *acid/acidic* (a negative coefficient; see Table 7). The large positive effect on function 1 of *base* and *accept hydrogen* are in good agreement with our scoring rubric, as it would be difficult for a student to achieve a high score without addressing the basic nature of the amino functional group. The large negative effect of *acid/acidic* may be due to the fact that a number of students expressed a misstatement that amino groups only existed as amino acids (i.e. covalently linked to a carboxylic acid). Such a misstatement would obviously result in a lower score on the scoring rubric and may account for part of the strong negative association of *acid/acidic* with function 1.

Table 7: Structure Matrix resulting from analysis of amino-hydroxyl pH question.

	Function	
Category	1[a]	2[b]
base/basic	.517	.349
acid/acidic	-.378	.629
amino group	.280	.128
hydrogen	-.039	.605
charge	.121	.137
accept hydrogen	.345	.388

[a] Function 1 accounts for 91.9% of the variance
[b] Function 2 accounts for 8.1% of the variance

This seems to suggest that function 1 in the predictive scoring model relates strongly with acid/base chemistry of these biological functional groups. Students who explain the basic nature of the amino group would be assigned a higher score and predicted to be classified in bin 1 of the scoring rubric, which is aligned with a human evaluation. It is important to define the difference between the categories *hydrogen* and *accept hydrogen*, seeing as they give dramatically different correlation results. The category *hydrogen* is a more general category designed to include all response that use the term "hydrogen" in any manner. The category *accept hydrogen* is a subset of responses in the hydrogen category. The *accept hydrogen* category contains a function which only selects responses that include the term "hydrogen" and the term "accept" or "pick-up". The disparate correlation results for these two categories indicate the level of sophistication that responses may be categorized and to which the classification function can distinguish between student groups.

Again, there are numerous categories that do not add value to the classification functions. A few categories are surprising that they do not add predictive value, such as *ionization*, a key component in describing weak and strong acids, and *raise pH*, which is the correct result of adding an amino group. The non-use of *ionization* is probably due to its high correlation with the category *hydrogen*, which is used as a scoring variable.

Although, *raise pH* shares some correlation with the *base/basic* category, it may be that responses that indicate an increase in pH supply wrong reasoning for this effect. Therefore, *raise pH* is not a good indicator of the quality of student response.

To test the utility of the classification function, we used a cross-validation classification in which each case is classified by the functions derived from all cases other than that case. We used prior probabilities of group membership based on group size, though the results with equal prior probabilities are very similar. The results are shown in Table 8. Correctly classified cases are shown in bold on the diagonal. The function classified 77.0% of the cases correctly; by chance, we would expect to classify 33% of the cases correctly. The worst cases were the level 2 responses, where only 42.9% were correctly classified. This is likely due to that fact that only 14 of the original student responses were rated as level 2 by the experts.

Table 8: Percent of cross-validated student responses for amino question classified at each level

	Computer Predicted Rating		
Expert Rating	1	2	3
1	**82.9**	12.2	4.9
2	21.4	**42.9**	35.7
3	6.9	12.1	**81.0**

Conclusion

By using lexical analysis software, we were able to successfully categorize student open-ended responses to two sets of chemistry questions about topics that are foundational for cellular and molecular biology. For free energy, we determined that students agree that reaction spontaneity can change with temperature, but relatively few attempt to use Gibbs free energy equation (or parts thereof) to explain this phenomenon. However, after biology instruction in cell metabolism, students not only provide ATP as a possible solution to a cell energy question, but also many of them successfully link ATP to its role or creation inside a cell. This becomes a more insightful observation for an instructor as opposed to whether a student chooses an "ATP" response from an identical multiple choice question.

We have also demonstrated that multiple choice questions do not necessarily reflect proper student knowledge of a topic. In analysis of the amino-hydroxyl pH question, although 129 students selected the correct answer on the multiple choice part of the question (question 4a), fewer than half of these students could supply a correct chemical explanation of why that selection was the right answer (question 4b). Therefore, one option for improving multiple choice questions then, is to couple a multiple choice response with a constructed response that can be subjected to lexical analysis.

In our analysis of questions relating to acid/base chemistry, we were able to use discriminant analysis to create classification functions that predict how expert raters would score these responses with 83.8% and 77.0% accuracy ($p < .000$). The results suggest that computerized lexical analysis can successfully be used to categorize student open-ended responses. We do not suggest that these classification functions be used in place of human scorers. However, it does seem feasible to design a series of questions in

an online homework set containing open response questions. Using computer predicted scoring functions, student responses could be evaluated by the software and students could be directed to a next question based on whether their predicted score is "correct" or "incorrect". This would allow homework sets be tailored to individual students and to give students more relevant practice in topics where they demonstrate a lack of understanding.

The lexical analysis software provides varying levels of detail and analysis for the user. The first, most broad level of response evaluation provided here is a simple overview of all the student responses (see analysis of question 1; reaction spontaneity). This "snapshot" allows the instructor to quickly scan the entire class and see all categories for term usage and frequency to get a general sense of how students answered the question. This can be more informative and inclusive than reading a select number of student responses to gauge the level of a class. The second level involves looking at connections between categories (see analysis of question 2; polymerization reaction). This is significant for complex questions because it allows the instructor to see past the number of terms and to see connections between categories. This level better highlights the connection trends made by the class as a whole and indicates if students are attempting to connect the ideas/concepts at all. Finally, the deepest level of sophistication is to use the software to predict the quality of student responses (see analysis of question 3; strong/weak acids) to provide formative feedback for the instructor. At this level, statistical software uses the combination of all categories and responses to provide scoring functions applied to all student responses. In the examples we provided here, it seems feasible that this computer predictive scoring could be applied in certain circumstances. In addition, the utility of the predictive model in characterizing which terms, functions and categories are useful (or not useful) predictors of student responses is instructive. Instructors can attempt to pinpoint either key principles or student misconceptions for a given topic/question. Looking at student connections and terminology on key concepts could be beneficial for instruction leading to changes in student learning.

We also see utility of this software for application in Just in Time Teaching (Novak, Gavrini, Christian & Patterson, 1999) by having students write short responses to conceptual questions before class and characterizing their responses for the instructor. Again, a major advantage of the software is that the instructor can get an idea of the level of understanding of the whole class in a few moments as opposed to reading select responses or quickly scanning a number of them. For example, by asking question 4 and seeing that over 60% of the students had little understanding of the concept, the instructor could adjust the next class material to address common student mistakes about functional groups and pH. These techniques can allow instructors in large enrollment course to use constructed response assessments much more frequently than would be possible relying only on human raters to evaluate the assessments. Such formative assessments can have a positive impact on student learning.

We believe further refinement of our scoring rubric and lexical analysis could result in improved classification functions. For example, in discriminant analysis of responses to amino-hydroxyl pH question, we note that classification function 1 accounts for 92% of the variance seen between the groups. It may be future analysis of these responses are scored only in a 2-bin system. This new scoring rubric may be based on

acid/base knowledge as predicted by function 1. Similarly, for questions 3a and b, one choice for refinement may be to collapse bins 2 and 3 into a single bin, providing a more populated bin to analyze. Perhaps this will improve classification functions for the middle portion of our scoring rubric. As well, it appears that we have many non-informative (and possibly redundant) categories that may be eliminated or combined to provide more discriminant functions or a simpler response distribution for instructor analysis.

For several of these questions, we are now on our third iteration of collecting student responses. Through the process, we have made some useful observations about the structure and nature of questions to facilitate computerized lexical analysis. Firstly, any words that appear in the stem of the question will not yield much (if any) utility in analysis of response. This statement may seem obvious, but we have found even small modifiers in the stem, such as "slowly", lead students to talk about topics, rates and kinetics for instance, that they may not bring up themselves and therefore may skew the analysis. The software allows for functions, that is, writing rules for only certain combination of words, such as "increase temperature" for question 1 above. These functions are very valuable to select meaningful responses, but even so, question stem words continue to be a problem in analysis. Secondly, the more specific parts a question may be reduced to, the better. Our original attempts at question 3 (above) consisted only of a single question and response area to the current question 3a & b. In our original analysis, we discovered that many students offered examples of weak and strong acids (both correct and incorrect) and that it was difficult to get the software to recognize when a student was defining a strong acid or when a student was defining a weak acid. Our solution to this was splitting the definition part of the question into two response boxes and adding separate response boxes entirely for examples. We feel we have attained significant improvement in analysis of this question due to these changes. Finally, when asking students to explain a response to a multiple choice question, it is a common practice for them to explain why they *didn't* select another answer. A common student explanation looks like "I chose X because Y and Z didn't…." This means that in analyzing all responses for students that chose X, the terms Y and Z will be present in the analysis. The instructor then must make sure the other terms in the analysis are linked to a description of X and not to Y or Z (or vice versa).

We believe that lexical analysis can be a useful tool for instructors in evaluating constructed response assessments. The ability to see the term usage and category connections for an entire class at once glance is valuable for the instructor to assess class understanding as well as modify instruction when necessary. Further analysis can help identify vital concepts in student explanations and/or misconceptions. Coupling this lexical analysis with computer-determined classification functions of student responses opens the possibility for the critical evaluation of large number of constructed response items.

References

Bialek, W., & Botstein, D. (2004). Introductory science and mathematics education for 21st-century biologists. *Science, 303*(5659), 788-790.

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. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience and school*. Washington, DC: National Academy Press.

Claesgens, J., Scalise, K., Wilson, M., & Stacy, A. (2009). Mapping Student Understanding in Chemistry: The Perspectives of Chemists. *Science Education, 93*(1), 56-85.

Moscarella, R. A., Urban-Lurain, M., Merrill, J. E., Richmond, G., Patterson, R., Parker, J. M., Merritt, B. W., Wilson, C. D., Long, T. M. (2008, March 30 - April 2). *Understanding undergraduate students' conceptions in science: Using lexical analysis software to analyze students' constructed responses in biology.* Paper at NARST 2008 Annual International Conference, Baltimore, MD.

National Research Council, (2002). *BIO2010: Undergraduate Education to Prepare Biomedical Research Scientists;* National Academies Press: Washington, D.C,; *http://www.nap.edu/books/0309085357/html/*

Novak, G. M., Gavrini, A., Christian, W., & Patterson, E. (1999). *Just-in-time teaching: Blending active learning with web technology*. Upper Saddle River NJ: Prentice Hall.

Parker, J., Anderson, C., Merrill, J., Heidemann, M., Long, T., Merritt, B., Richmond, G., Sibley, D., Urban-Lurain, M., and Wilson, C. (2007). Where Has All The Carbon Gone?" A Thought Paper On Frameworks For Assessing Biology Understanding *Proceedings of the Conceptual Assessment in Biology Conference,* Boulder, CO.

Reingold, I. D. (2004). Inverting organic and biochemistry: A curriculum tweak that benefits all. *Journal of Chemical Education, 81*(4), 470-+.

Richmond, G., Urban-Lurain, M., Parker, J., Merrill, J., & Merritt, B. (2008, March 30 - April 2). *Assessment-informed instructional design to support model-based reasoning in college-level biology.* Paper presented at the NARST 2008 Annual International Conference, Baltimore, MD.

Schwartz, A. T., & Serie, J. (2001). General chemistry and cell biology: An experiment in curricular symbiosis. *Journal of Chemical Education, 78*(11), 1490-1494.

Spicer, J. (2005). *Making sense of multivariate data analysis*. Thousand Oaks, Calif.: Sage Publications.

SPSS. (2007). *SPSS Text analysis for surveys 2.1 user's guide*. Chicago, IL: SPSS, Inc.

Wenzel, T. J. (2006). General chemistry: expanding the learning outcomes and promoting interdisciplinary connections through the use of a semester-long project. *CBE Life Sci Educ, 5*(1), 76-84.