ASEE's 123rd Annual • Conference & Exposition • New Orleans, LA • June 26-29, 2016

Paper ID #16419

Applying Natural Language Processing Techniques to an Assessment of Student Conceptual Understanding

Christian Anderson Arbogast, Oregon State University

Christian Arbogast is a graduate student in the School of Mechanical, Industrial, and Manufacturing Engineering at Oregon State University. His academic and research interests include adapting computer science techniques to supplement traditional qualitative analysis and the mechanical design process.

Dr. Devlin Montfort, Oregon State University

Dr. Montfort is an Assistant Professor in the School of Chemical, Biological and Environmental Engineering at Oregon State University

Using Natural Language Processing Tools to Enhance Qualitative Evaluations of Conceptual Understanding

Abstract

This work-in-progress briefly surveys a selection of open-source Natural Language Processing (NLP) tools and investigates their utility to the qualitative researcher. These NLP tools are widely used in the field of lexical analysis, which is concerned with automating the generation of useful information from human language using a variety of machine processes. Recent research shows that the statistical analysis of software recognized linguistic features can benchmark certain mental processes, such as cognitive load. This investigation generates those linguistic indicators using transcripts from a multi-year, interview based study and compares them to a qualitative analysis of a subject's conceptual understanding of various engineering topics. Our intermediary findings indicate a correlation between changes in the linguistic indicators introduced in this paper and a qualitatively coded analysis of conceptual understanding over time. Future work will involve increasing the breadth of the dataset to further establish the fidelity of this approach and expand on the premise of using automatically generated linguistic indicators to aid the qualitative researcher.

Introduction

Improving learning outcomes for students in our engineering programs may be an idea with common support but the overall fuzziness of that goal hardly contributes to a common set of actionable processes. However, as a diverse field of researchers, we can hope that a diversity of small steps will eventually coalesce around that ideal. One particular subgoal, as set out in [1], is increasing an instructor's capability for accurate formative assessment, or the process of making student learning readily visible using a variety of in-situ tools. Formative assessment differs from traditional assessment (ie, traditionally scored tests or homework) in that it aims to illuminate some of the underlying knowledge structures held by the student, not just their ability to meet a normative proficiency of skills. In that way, formative assessment is a manner of understanding a student's conceptual understanding of a subject area, or an individual's underlying and internalized framework of how the world works [14].

One complication of recognizing student conceptual understanding is that it tends to be largely personal to the student [14], which renders it a time consuming and expensive undertaking. Concept inventories [8] have been popularized as a method of recognizing student conceptual understanding but need to be designed by an expert to test specific concepts in isolation. Some critics have commented that concept inventory tests can be valuable pedagogical tools but lacking without a complementary statistical analysis of the results [7]. It seems that the gold standard for assessing student conceptual understanding is the individual attention of an experienced qualitative researcher who is well versed in the methodology, or what can be described as a craft of inductive reasoning [17].

Increasing the number of researchers interpreting a dataset can lead to greater reliability of the conclusions drawn but it also leads to a significant increase in the cost of the study.

Developments in the computational power available to researchers, as well as significant advancement in software and machine learning methodologies have greatly increased the size and complexity of datasets that a single researcher can effectively analyze. In particular, many Natural Language Processing (NLP) tools have recently become available that give a researcher the ability to extract meaningful information from recorded human language. The goal of this study is to investigate the use of NLP techniques in extracting linguistic indicators of an individual's knowledge and recognize evidence of changes in a student's conceptual understanding over time.

Literature Review and Theoretical Basis

In a broad sense, text analytics is the process of retrieving data from unstructured text in a meaningful way. Lexical analysis is a series of procedures that assign natural language meaning by examining the syntax of that text. These analysis techniques are widely used in the field of computational linguistics, which focuses on creating statistically formed meaning using features of natural human language. In our study, these definitions refer to various steps in extracting the meaning behind text.

Many commercial and academic fields have been using automated tools to analyze text since the early 1980s [10]. Early researchers focused on developing the computational approaches for automating analysis of linguistic structure in order to facilitate machine translation, speech recognition and synthesis, and keyword recognition [9]. The previous approaches can be seen as developing sophisticated ways of finding information already embedded in a dataset. Now, researchers are moving on to more meta-level analyses. This ranges from the simple frequency tracking of emotion words for the purpose of estimating Facebook user sentiment [19] to automatically assessing the correctness of student responses in a classroom setting using complex machine learning algorithms and artificial intelligence [15].

A recent application [6] attempted to use automated text analysis to judge student understanding by extracting high level concepts from open ended written responses to survey questions. Their work focused on using thematic analysis. The dataset was comprised of open ended, written responses. That study demonstrates some positive benefits of the approach in terms of the speed of analysis and potential of the approach but drawbacks included making a large proportion of incorrect categorizations. Other researchers have started to explore using NLP to interpret open ended student response, such as the development of a tool for assigning reviewers and facilitating team formation [22]. A commonalty between these two approaches is the attempt to extract complex meanings from language usage, largely to mixed results.

Our approach does not focus on creating conceptual abstractions directly from the literal content of textual data but rather uses features of the syntax, or relationships between communicated words, to generate indicators linked to conceptual understanding. Differences between expert and novice problem solving strategies contrast an expert's highly developed and interconnected knowledge framework and a novice's superficial and fragmented one [4]. We would expect an expert to fluently move between concepts related on a deep conceptual level, whereas a novice may struggle greatly to connect unified concepts that have surface level conflicts. The relative

ease with which an individual traverses linked concepts within a problem domain is not only descriptive of conceptual understanding in that area but also appears linked to external physiological effects, including the lexical syntax of communication.

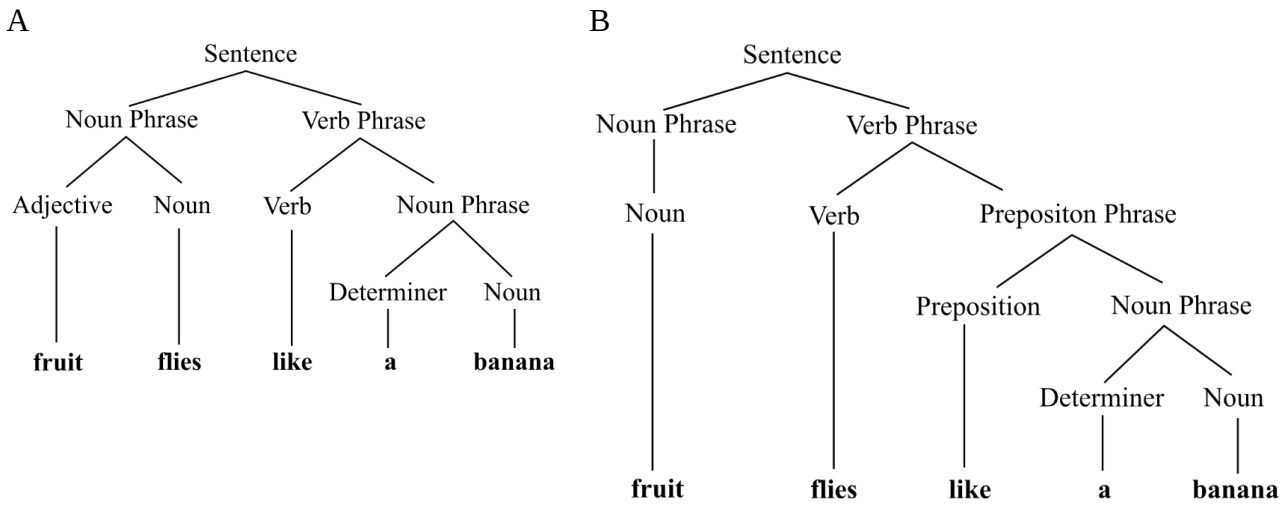
Researchers in the area of Cognitive Load Theory [2] pose a connection between the cognitive processes involved and certain linked physiological effects. In fact, many researchers look for a demonstrable physiological effect, such as electroencephalography monitoring (EEG), as an indicator of how a subject is processing information. A 2012 study along the lines of that approach aimed to examine the concurrence of various physiological indicators for a subject under high versus low cognitive load [3] and found that there was a very high correlation between measures such as handwriting degeneration, object position recall, EEG, eye-tracking measures, certain lexical features, and the subject's cognitive load. Demonstrating the co-occurrence of those physiological indicators of high cognitive load is promising as it could mean that a software based lexical analysis may be as useful as eye tracking systems or an EEG machine in showing features of real-time cognition. The low cost of dissemination for the software-based approach would certainly be a boon for researchers looking to connect cognitive load and more nebulous features like conceptual understanding.

A recent study that built off of the connection between lexical indicators and cognitive load found that software generated measures of lexical syntax, complexity, and diversity were very strong predictors of English as Second Language learners' conceptual grasp of conversational English [12] and related task performance. Our study is largely informed by the work of [12] and seeks to investigate the application of a similar procedure. Our approach is to automatically generate lexical indicators using various software, make an inference about the cognitive load of the participant, and determine if there is any correlation to a traditional qualitative assessment of conceptual understanding.

Natural Language Processing Tools

The methods described in this section involve complex linguistics procedures and computational grammars, which are entire fields unto themselves. Fortunately, many research tools from those fields have been developed with open-source ideals in mind and can be easily combined into existing research. An expanded explanation of the underpinnings of these techniques will be detailed in work subsequent to this work-in-progress. For the sake of brevity, we will be naming the procedures adapted to this study and briefly discussing what they do, rather than heavily detailing how they work.

The first step in our procedure is to determine the Part of Speech (POS) for each word. The meaning behind the words we use are largely ambiguous and dependent on the context of their use. The POS of each word will be important to our analysis. Consider the two different POS parsing trees shown in Figures 1a and 1b that chart the decomposition of one part of a common humorous phrase.



Figures 1a and 1b:
Illustrating the ambiguity of possible syntactic POS parsing trees for a common humorous phrase

To a human reading the sentence: "Fruit flies like a banana", humor arises as a result of the two simultaneous existing and conflicting meanings of the sentence. Is the subject a tiny, but enormously aggravating insect or the general flight characteristics of fruit? After slight reflection, it is easy for a person to settle the conflict and judge the correct meaning based on the context of the conversation but it is much harder for a software program to do so. In order for a machine to proceed, it must make probabilistic judgements based on the lexical syntax, or context, much like a person would. It must employ sophisticated algorithms to create probabilities of various interpretations being correct. Even then, it cannot be sure.

Our analysis makes use of the Stanford POS tagger [21], which is software that accepts text input and determines the part of speech for each word in the text sample according to the Penn English Treebank tag set, as described by [18]. The software allows the user to 'train' a POS model using a customized textual dataset if wanted, which may take into account application specific jargon and regional speech differences. However, we have opted to use the included English language Bidirectional Model which has a reported accuracy or over 90%, even when judging unknown words [18]. This program is written in Java and released under GNU General Public License, which allows for copying, modification, and redistribution. It is used in our implementation without any modification.

The next stage of lexical analysis is to simplify the words used into what are called lemmas. This is the process of grouping different forms of similar words into single entities, or meanings. The ability to do so is largely determined by the part of speech of the word and access to a database of word relations for a specific culture. The meaning behind context of each word, rather than the literal characters is the subject of our future lexical analysis. Figures 2a and 2b demonstrate this effect.

Figure 2a: Example of lemmatizing multiple tenses of a word into single lemma
Figure 2b: Demonstration of part-of-speech dependence on lemmatizing

Figure 2a shows how different tenses of the word "have" can be collapsed into one meaning, or lemma. Figure 2b shows the importance of part of speech in lemmatizing. "Cats" may refer to the concept of a small feline but the adjective "Cattiest" is not typically a descriptor of that same animal. Our ability to perform lemmatization is built around use of the Python Natural Language Tool Kit (NLTK), a software package of computational linguistics tools for the Python programming language. Python NLTK is an open source library. Using Python NLTK allows us to interact with Princeton WordNet [13]. WordNet is an online lexical database, or machine readable dictionary which groups words of similar meaning based on their part of speech. The purpose of this step is to isolate the meaning of the words into lemmas before further analysis and comparison.

The third tool employed in this study is a lexical statistics analyzer, which has been adapted from the source code release by [12]. It was originally developed to generate lexical indicators of English as Second Language learners' written text but can be applied for any purpose. It can generate information about the lexical complexity, density, and variation of text based on syntax, as well as word choice sophistication based on words used and their part of speech.

We will be focusing on the Uber Index (Figure 3) developed by [5] that highlights certain features of communication of an individual as a reflection of lexical diversity. Lexical diversity can be seen as measure for how varied the structure of speech is and this category of assessment has significant research history within the field of linguistics. This indicator reflects lexical diversity by relating the total number of lexical words used (T) in a text sample with the number of unique lexical words used (N).

$$Uber\ Index: \quad U = \frac{log^2 T}{log T - log N}$$

Figure 3: Uber Index [5].

This index was chosen for analysis because the Uber Index is seen as as a better representation of lexical diversity for texts of varying length than other similarly used indices of lexical diversity, as found by [11]. Our interview based data collection methodology did not specifically manage the duration of participant response and this is an effort to reduce the influence of differing length of text samples on an assessment of lexical diversity.

Indicators of lexical diversity are one of the simpler lexical characterizations but this serves a specific purpose in this study. Uber is independent of the specific meaning of the words used.

We will be examining problem solving within core engineering concepts and expect to encounter large amounts engineering jargon. More complex lexical indicators usually make comparisons to the language usage of a typical population within a certain culture, using the meaning of words as a feature of analysis. By focusing on base level syntactic features, we can avoid highlighting engineering terminology that is rarely used in a more general population. Furthermore, focusing on Uber as a metric may prevent some of the uncertainties involved with more complex lexical indices from overshadowing the development of the research methodology.

The open-source tools in this discussion are interacted with using a custom GUI developed in Python, which gives the user a simple point-and-click approach to using these tools simultaneously on an arbitrary text dataset. These individual software tools are typically integrated into more complex, ad-hoc applications or accessed individually using a command line interface. Providing an easy to use interface is intended to promote these software resources to researchers not already familiar with them.

Methodology

Our dataset is based on a preexisting selection of interview trasnscripts conducted with entry level engineers over the first three years of their professional carreers. The subjects were self-selected as seniors from the student body of a large public university in the Pacific Northwest. Over the course of those three years, intensive interviews were periodically conducted where they were asked to solve ill-structured questions centered around engineering fundamentals. Those questions took the form of classroom-like written problems but without some of the information that they might have expected to receive in a classroom setting. The questions were repeated over the course of the three year study in order to make comparisons between their question responses as they gained experience in their chosen engineering field.

The transcripts were excerpted so that responses to similar problem statements over the three year study could be directly compared. If the subject's only response to a question was the phrase "I don't know" in year one and their response involved a lengthy discussion in year three, we would not directly compare the lexical diversity of the two texts due to low content levels. Interviewer speech was removed so only the participant's words were present. The first stage involved a traditional qualitative analysis of the interviews using the coding framework shown below in Figure 4. The qualatiave analysis was focused on teasing out evidence of the interviewee's conceptual understanding of the subject. To that effect, we modeled the coding framework after [16], which made a good comparison between levels of conceptual understanding and features of problem solving that tracked with our interview questioning.

The interview exerpts were then analyzed to generate lexical indicators using the Natural Lanuage Processing Tools described previously. These were interacted with through a Python based GUI that acted as a wrapper for the various extermal software tools, as shown in Figure 5. Once these two methods of analysis were completed, the results of each analysis were compared.

Level of Understanding	Code	Description
Conceptual	**C1**	*Implicit* ***relationship recognition*** [C]
	C2	***Reformulation*** *of problem (****cross-domain*** *analogy)* [C]
	C3	***Reflection*** *to see if answer makes* ***sense*** [C]
	C4	***Reflection*** *on appropriateness of* ***approach*** [C]
	C5	***Reformulation*** *of problem (****same-domain*** *analogy)* [C]
	C6	Step-by-step Approach (**Action Sequence**) [P]
	C7	Statement of **not enough information** [P]
	C8	***Recalling*** *Equation or definition* [P]
Procedural	**C9**	**Naming a Procedure** [P]

Figure 4: Coding Framework

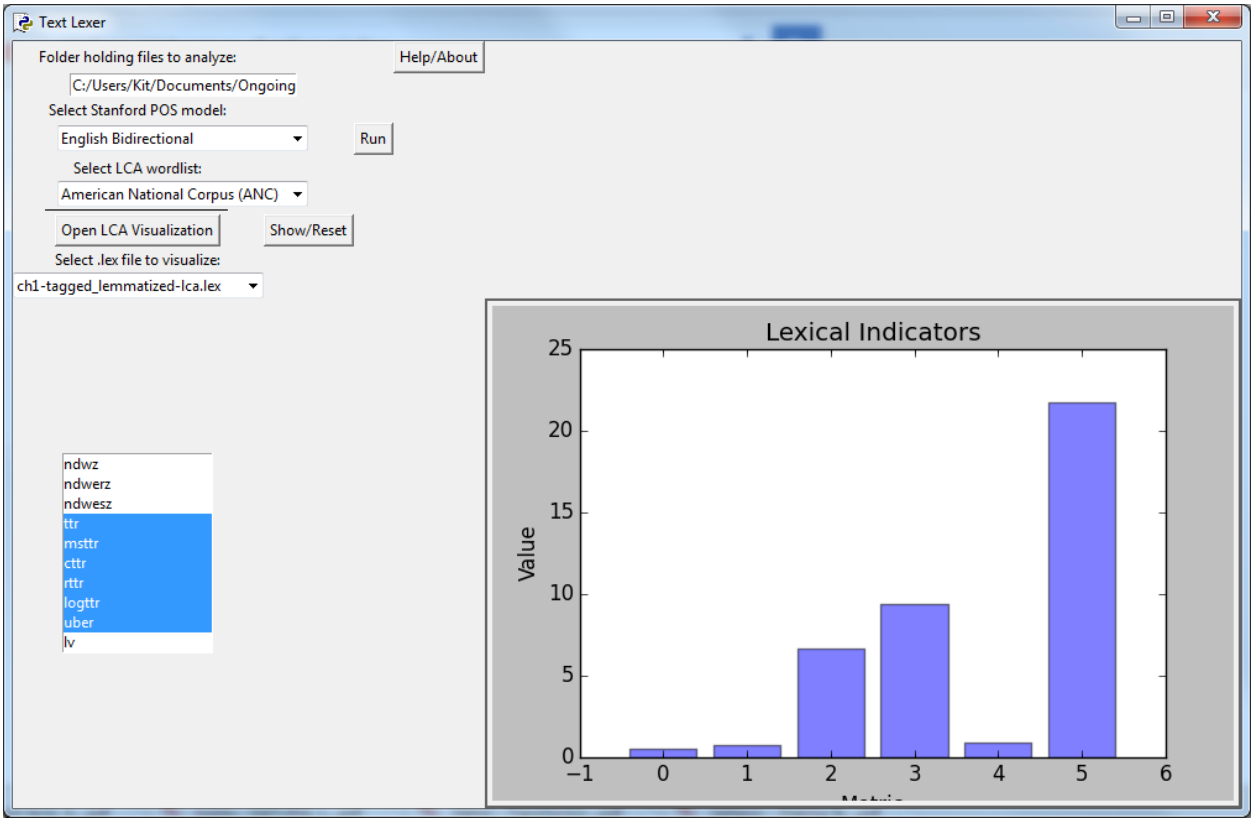


Figure 5: Python based GUI for POS tagging, Lemmatizing, Lexical Analysis and visualization of results.

Intermediate Results and Discussion

Please note: These finding are intermediary and not complete enough to warrant characterizing the overall statistical relevance of the approach. For this reason, tables of raw values are not included to avoid misrepresentation. Instead, a single representational sample is included to provide a point of discussion on the general qualities of this analysis approach.

The following interview excerpt was selected on the basis of how similarly the question was posed in both interview periods. This was done to isolate response from the engineer to a specific concept. The first example involves the interviewer directly asking where the highest velocity in the residential pipe network would occur. This is not the entirety of the interview session but gives us a chance to discuss the application of the coding framework. The subject matter of the excerpt is part of the Pipe Network portion of the interview. The lexical indicators generated from the entire question set are later discussed.

2011
Interviewer: Okay. What about highest energy?

Engineer A: Probably at the reservoir because it has the least amount of head loss.

[interviewee does not offer further rationale]

2013
Interviewer: Okay. Um, what about the highest energy?
Engineer A: Energy? Well potential energy would be at the reservoir.
Interviewer: Where at the reservoir?
Engineer A: Waiting to go down the pipe.
Interviewer: So like, at here at the bottom?
Engineer A: Yeah. I'm going to put my PE [potential energy] right here. And then that switches to kinetic energy, and then it has to travel really far. So I would say, here [points], is where my kinetic, once it gets switched from potential to kinetic. And then it goes and, the friction is going to make it slow down. But energy is not created or destroyed so it goes into the friction and becomes heat or something else.

This excerpt show how a subject's response can change from year-to-year. The most basic change in the text is the significant increase in quantity of information offered by the subject. However, that is not enough to make a judgment of the student's conceptual understanding. The 2011 response would warrant a Code of C8 or C9 (see Figure 4 for list) based on the simple naming of an engineering term. This is very low on the Conceptual-Procedural scale of Conceptual understanding. Together with lack of reflection on underlying features of the problem, we could not say a high level of conceptual understanding was demonstrated in this response, alone. However, in 2013, the subject not only names a procedure (a low value

conceptual understanding code) but reflects on the underlying analysis technique of conservation of energy (a high value code of C2). This is a pretty dramatic example of a demonstrated high level of conceptual understanding in an interview response. A comparison of the accompanying lexical indicators follow.

Figure 6 shows a sampling of the available a lexical complexity and diversity indicators for the above person over the course of the first three years of their professional engineering career, as generated using the software tool. The X-axis describes the name of the indicator and the Y-axis is a normalized percentage, which reflects relative sizes of the text data.

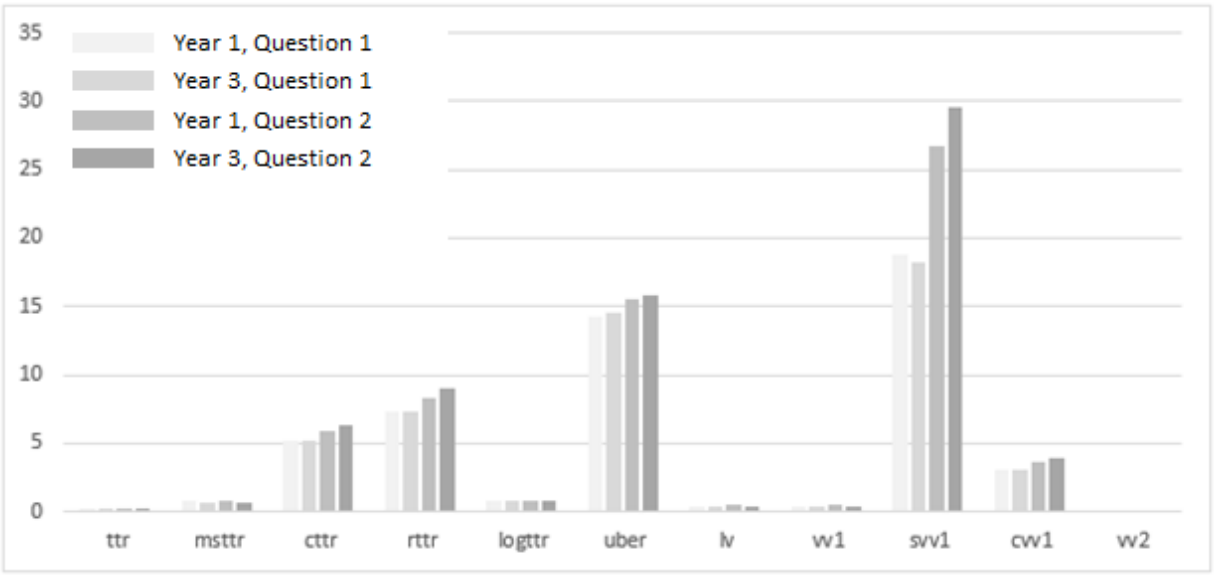


Figure 6: Changing Lexical indicators for a single person over time

Note the common trend in the results. This is a side effect of how these lexical indicators are constructed. They are essentially made of different combinations same linguistic feature 'building blocks', as seen in [12], which means that the indicators are mathematical transformations of one another. Our selection of the Uber Index as the preferred index of comparison may not dramatically highlight a specific lexical feature in the case of this individual but can also be seen not to introduce outliers into the results pool. Seemingly emergent trends will allow us to take two snapshots and interpolate how that person's response has changed over time.

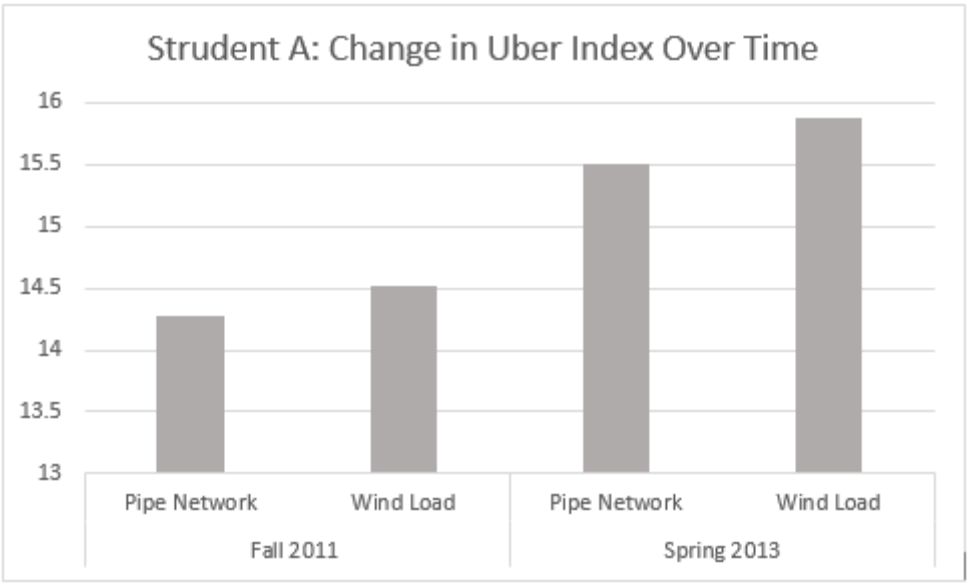


Figure 7: Change in Uber Index of Person A over Time.

Note the increase in Uber across both problem sets. The data was selected in such a way as to minimize difference in response across time. The question prompt given to the student was introduced in largely the same manner for the purpose of drawing comparison between the student's responses. A slight trend in Uber index across both problem sets has emerged (although the selected scale of the Y-axis erroneously magnifies this change). This shows an increase in lexical diversity for the same problem, year-to-year. This is a possible indication that the cognitive load of the individual is lower during the second iteration of question responses. This may reflect on a changing proportion of germane load to intrinsic load (in Cognitive Load Theory), as a result of increased conceptual understand of the individual. This would demonstrate higher proportion of long term memory access compared to short term and indicates access of more well developed schema which hints at growth of expertise.

State of Ongoing Work

We are generally seeing a correlation between the change in values of the lexical indicators generated and the qualitative assessment of conceptual understanding over time. However, our set of analyzed data is not yet sufficiently large enough to draw statistical conclusions. The small example shown in the results and discussion section may not be representative of the entirety of that student's interview. While we are focusing on the Uber index of lexical diversity as a likely metric for validation of the methodology, we do not yet know if this will prove to be the most insightful lexical feature speech.

It should be noted that this approach cannot assess the "correctness" of the response. This approach would not be able to differentiate between a well exercised, fluently accessed conceptual understanding consisting of many misunderstandings and an expert conceptual understanding of that same knowledge domain. The posited relationship between conceptual

understanding and displayed lexical features cannot incorporate this level of analysis without using interpretations of the natural language meaning of the words used. Further refinement of the methodology is needed before that research direction can be explored.

Our intent is not to try to automate high level analysis, rather we aim to investigate the decomposition of those high level features into base elements that could yield a secondary level of insight onto a traditional qualitative analysis. A smaller intent is to create a software package that is easy to implement, share, and modify if it is found to be of use to researchers. The visual display can summarize any of the generated lexical indices, as specific indices may be more relevant depending on the nature of the analyzed text. An optional feature is the ability for the researcher to represent newly supported lexical indices from the literature, or even create their own lexical indices out of basic linguistic feature 'building blocks' as a further path of research. A major goal of this ongoing project is to bring a Natural Language Processing based approach to a field of research where it shows promise and is relatively absent.

1. Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn brain, mind, experience, and school*. Washington, DC: National Academy Press.

2. Chandler, P., & Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction*. http://doi.org/10.1207/s1532690xci0804_2

3. Chen, F., Ruiz, N., Choi, E., Epps, J., Khawaja, M. a., Taib, R., … Wang, Y. (2012). Multimodal Behaviour and Interaction as Indicators of Cognitive Load. *ACM Transactions on Intelligent Interactive Systems*, *2*(4), 22:1–22:36.

4. Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1979). Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science*, *5*, 121–152. http://doi.org/10.1207/s15516709cog0502_2

5. Dugast, D. (1979). *Vocabulaire et stylistique. I Théâtre et dialogue. Travaux de linguistique quantitative.* Geneva: Slatkine-Champion.

6. Goncher, A., Boles, W. W., & Jayalath, D. (2014). Using automated text analysis to evaluate students ' conceptual understanding.

7. Heller, P., & Huffman, D. (1995). Interpreting the force concept inventory: A reply to Hestenes and Halloun. *The Physics Teacher*. http://doi.org/10.1119/1.2344279

8. Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141. http://doi.org/10.1119/1.2343497

9. Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266.

10. Hobbs, J. R., Walker, D. E., & Amsler, R. A. (1982). Natural language access to structured text. *Proceedings of the 9th Conference on Computational Linguistics*, *1*, 127–132. http://doi.org/10.3115/991813.991833

11. Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84. http://doi.org/10.1191/0265532202lt220oa

12. Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *Modern Language Journal*, *96*(2), 190–208. http://doi.org/10.1111/j.1540-4781.2011.01232_1.x

13. Miller, G. a. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41. http://doi.org/10.1145/219717.219748

14. Montfort, D., Brown, S., & Pollock, D. (2009). An Investigation of Students' Conceptual Understanding in Related Sophomore to Graduate-Level Engineering and Mechanics Courses. *Journal of Engineering Education*, (April), 111–129. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2168-9830.2009.tb01011.x/abstract

15. Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming Biology Assessment with Machine Learning: Automated Scoring of Written Evolutionary Explanations. *Journal of Science Education and Technology*, *21*(1), 183–196. http://doi.org/10.1007/s10956-011-9300-9

16. Rittle-johnson, B., & Schneider, M. (2013). Developing Conceptual and Procedural Knowledge of Mathematics. *Oxford Handbook of Numerical Cognition*. http://doi.org/10.1093/oxfordhb/9780199642342.013.014

17. Saldana, J. (2009). An Introduction to Codes and Coding. *The Coding Manual for Qualitative Researchers.*, (2006), 1–31. http://doi.org/10.1519/JSC.0b013e3181ddfd0a

18. Santorini, B. (1991). Part-of-speech Tagging Guidelines for the Penn Treebank Project.

19. Settanni, M., & Marengo, D. (2015). Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. *Frontiers in Psychology*, *6*(July), http://doi.org/10.3389/fpsyg.2015.01045

20. Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. http://doi.org/10.1207/s15516709cog1202_4

21. Toutanova, K., Klein, D., & Manning, C. D. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1 (NAACL '03),* 252–259. http://doi.org/10.3115/1073445.1073478

22. Verleger, M. A., & Beach, D. (2014). Using Natural Language Processing Tools to Classify Student Responses to Open-Ended Engineering Problems in Large Classes Using Natural Language Processing Tools to Classify Student.