AN ABSTRACT OF THE THESIS OF

Christian Arbogast for the degree of Master of Science in Mechanical Engineering presented on March 16, 2016.

Title: Assessing Student Conceptual Understanding: Supplementing Deductive Coding with Natural Language Processing Techniques.

Abstract approved:

Devlin B. Montfort Bryony L. DuPont

Assessing student conceptual understanding is a valuable method for gauging specific student learning outcomes but can be difficult and time consuming to measure. This research investigates the potential of automating some intermediary steps of a qualitative analysis of student conceptual understanding with tools from the field of Natural Language Processing (NLP), Computational Linguistics, and Cognitive Psychology. This investigation centers on interviews conducted with newly graduated engineers over the first three years of their professional lives. Lexical features of those interview transcripts were measured using a variety of open-source NLP software. A theoretical framework was developed to link those lexical indicators to features of cognitive load and conceptual understanding. Results of the application of NLP to create indicators of student conceptual understanding were compared to the results of a traditional qualitative assessment. It was found that certain lexical indicators, primarily the Uber Index, could be useful descriptors of conceptual understanding when certain conditions are met but that the limitations of the approach must further addressed in future research. Factors that limit this application include variances in text length and the magnitude of change in conceptual understanding in an individual.

Assessing Student Conceptual Understanding:
Supplementing Deductive Coding with Natural Language Processing Techniques

by
Christian Arbogast

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Presented March 16, 2016
Commencement June 2016

Master of Science thesis of Christian Arbogast presented on March 16, 2016

APPROVED:

Co-Major Professor, representing Mechanical Engineering

Co-Major Professor, representing Mechanical Engineering

Head of the School of Mechanical, Industrial, and Manufacturing Engineering

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Christian Arbogast, Author

ACKNOWLEDGEMENTS

I wish to deeply thank Dr. Devlin Montfort for providing me the research opportunity and inspiration to move forward in an unexpected field. Your patience as I figured out how to become a graduate researcher has been greatly appreciated. Thank you for your friendship and guidance.

To Dr. Bryony DuPont: Thank you for taking on an interdisciplinary student with the promise of so little return in return. Exposure to the people and ideas from the Design Engineering Lab at OSU has helped shape the way I view engineering as a practice.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION AND MOTIVATION

An objective that stems from the National Research Council (NRC) commissioned exploration of student learning, How People Learn, is the need for ongoing research on the topic of formative assessment, or the practice of "making students' thinking visible by providing frequent opportunities for assessment, feedback, and revision …" (Bransford, Brown, & Cocking, 2000). This is an important aspect of bridging the gap between the academic study of knowledge and learning, and real-world application to improve student learning outcomes in the classroom. Rather than informing one specific research direction, this recommendation seeks to present the value of targeting formative assessment to a diverse research community, with the goal of establishing a field of knowledge in the area. Many perspectives are needed to generate the foundational science of theory, rigorously test through application, and refine landscape of the practice of formative assessment.

In a broad sense, targeting how people develop patterns of thinking, organize knowledge, and evolve expertise has long been a goal for human learning and development researchers; from the early pioneers, Vygotsky and Piaget, to modern researchers such as Michelene T.H. Chi and Ruth A. Streveler (Michelene T.H. Chi, Feltovich, & Glaser, 1979; John-Steiner & Mahn, 1996; Piaget, 1997; Streveler, Litzinger, Miller, & Steif, 2008). The influence of those eminent researchers has greatly shaped our current understanding of how humans, collectively, learn and serves as a strong undercurrent to the process of improving engineering education. The practical impacts are varied, but a very important one includes the creation of schemes to model complex psychological systems as researcher-recognizable cognitive structures and describe how those structures change throughout the process of knowledge acquisition (Blake and Pope 2008). Many engineering education researchers are exploring the criticality of student conceptual understanding to the development of expertise, within the interconnected framework developed by the aforementioned researchers (Bransford et al., 2000; Litzinger, Lattuca, Hadgraft, & Newstetter, 2011; Svinicki, 2010).

Despite the increasing focus on the cognitive factors underlying student learning, the challenge of translating elements of a broadly applicable understanding of human learning into curriculum level practice persists. In a 2010 study, Borrego, Froyd, & Hall investigated the diffusion of engineering education innovations among U.S. engineering departments. Each innovation had been developed with a modern awareness of the cognitive factors involved in student learning. A finding of this work was that departmental adoption of the practices dramatically under paced overall awareness of the selected innovations. The low rate of adoption may be explained in many different contexts but the fact that more than 80% of respondents were aware of the innovations appears to demonstrate the motivation and willingness of engineering faculty to seek out research based tools to improve educational outcomes. The desire to improve student learning stands in contrast to research findings into the actual efficacy of our engineering programs. A 2009 study compared the conceptual understanding of fundamental mechanics of materials principles among a particular group of U.S. sophomore and senior engineering students to find that the seniors did not show any significant difference in conceptual understanding (Montfort, Brown, & Pollock, 2009). Newcomer and Steif (2008) showed that many students were unable simultaneously apply the fundamentally related Statics concepts of force and moment equilibrium despite the class being structured in such a way to reinforce the interrelation of those concepts.

Despite targeted research into student conceptual understanding of core engineering concepts, the instructional and research devices being developed to encourage and make student learning visible have yet to be shown to be practically effective. Perhaps the tools needed to assess the state of a student's conceptual understanding are unappealing to instructors due to some of the inherent barriers of translating research into practice, such as common constraints driven by cost, time, or overall situational appropriateness (Henderson & Dancy, 2011). An alternative explanation of low implementation is that it is quite hard to recognize when the instructional approaches are actually effective in increasing student conceptual understanding. Much of the established research into student leaning and cognition has focused on the development and substantiation of theory, while the overall efficacies of practical implementation are largely untested. This is partly the motivation behind NRC's push for improved methods of student formative assessment. Traditional methods of assessing student conceptual understanding are incredibly time intensive and newly developing assessment tools have yet to reach broad consensus of efficacy. Can the findings of student learning researchers be immediately applied to the classroom? Does the maturation of the field directly translate into improved student outcomes?

To answer these questions, better real-time modeling of the development of student conceptual understanding is needed. The goal of this research is to test the applicability of research findings from the fields of Natural Language Processing, Computational Linguistics, and Cognitive Psychology. Those individual fields take different approaches to assessing features of cognition and learning, which have not yet been widely applied within engineering education. Applying those unique research approaches may help to overcome some of the endemic challenges in assessing student conceptual understanding. The effect of fusing different research areas must be rigorously tested before conclusions may be drawn. This study compares a traditional assessment of student conceptual understanding using qualitative research methods and newly developing methodologies from other fields.

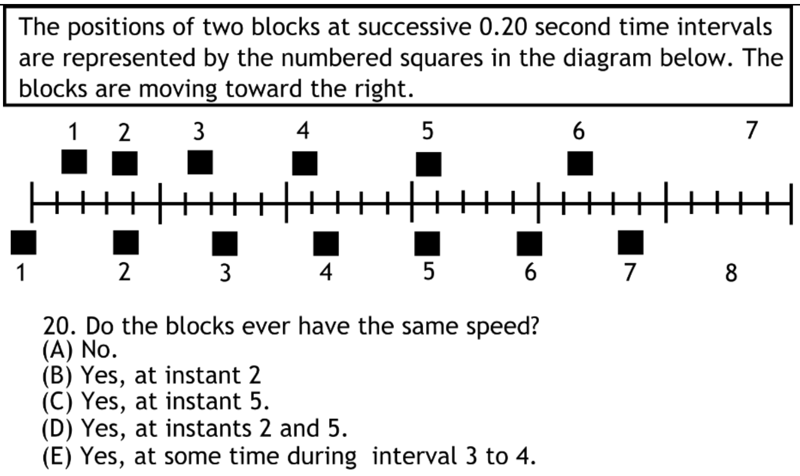
Of particular interest is ability to apply software-based Natural Language Processing tools that allow for automatic recognition of certain features of human speech. The immediate goal of this research is to investigate the impact of automating some intermediary steps of a qualitative analysis of student conceptual understanding. If some level of automation proves successful, benefits could include greater repeatability of analysis and increased interrater reliability. This would reduce the amount of time needed for a beginning qualitative researcher to gain proficiency. To test this goal, a standard qualitative assessment was performed on subjects participating in an interview-based study and compared with the results of an automated analysis of those same subjects. The results were interpreted with regard to their application in a traditional qualitative assessment. In this way, we may shed light on the context in which certain practices of assessing student learning can be supplemented with these new techniques, and perhaps even describe a new direction of research in our field.

1.1 A Common Approach

Current approaches to describing student conceptual understanding, skill acquisition, and the efficacy of our engineering programs tend to track with the guidelines laid out by researchers of effective student assessment, notably Bass & Glaser (2004), Baxter & Glaser (1998), Norton

(2009). The methodologies recommended by these researchers largely center on relating student performance to well-formed, measurable learning goals by comparing ideal and actual demonstrations of student learning outcomes. Such methods tend to look at the artifacts generated, both tangible and intangible, such as answers to test questions, implemented procedures, skills, reflections, habits, etc. They use what a student generates as a basis for assessing conceptual understanding. Unfortunately, using a student's demonstrated solution to a problem as a direct indicator of conceptual understanding can be challenging, as highlighted by Montfort, Brown, and Pollock (2009). Engineering students, in particular, can be proficient in applying rote methodologies while simultaneously holding very little real understanding or even understandings contradictory to their memorized problem-solving algorithms.

Therefore, solely relying on a student's demonstrated problem solving can easily misrepresent their underlying conceptual understanding of a subject. Integrating a wide portfolio of student work and one-on-one interactions can illuminate the underlying knowledge structure the student holds, but this poses challenges to the practitioner in the time required to perform the evaluation as well as the subjective nature of the endeavor. A popular instrument for gaining an understanding of a student's underlying conceptual framework is the Concept Inventory. The first widely used implementation in this genre was the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992), which tests an individual's competence in applying an isolated fundamental concept of physics. In this diagnostic tool, a simple scenario is presented to a student along with a variety of possible solutions that are designed to highlight common student conceptual difficulties and competencies (see example in Figure 1).



The positions of two blocks at successive 0.20 second time intervals are represented by the numbered squares in the diagram below. The blocks are moving toward the right.

20. Do the blocks ever have the same speed?
(A) No.
(B) Yes, at instant 2
(C) Yes, at instant 5.
(D) Yes, at instants 2 and 5.
(E) Yes, at some time during interval 3 to 4.

Figure 3:Example of Force Concept Inventory problem. Adapted from Edward, Richard, Redish, & Steinberg (1999)

The positive reception of the physics-based concept inventory has led to the implementation of this technique in the broad areas of Chemistry, Biology, and Engineering. Within engineering, the Foundation Coalition (foundationcoalition.org) has curated a broad collection of concept inventories that deal with subject matters that cover many engineering core classes. Each Concept inventory must be designed by an expert to test an individual concept in isolation, such that a student who answers the problem correctly can be seen to have demonstrated deep understanding of that concept. The designer of the concept inventory must have broad insight into common student misconceptions within the field (Evans, Gray, & Krause, 2003). A substantial benefit of the Concept Inventory is that it is packaged similarly to a traditional multiple choice assessment. As such, it is easy to distribute, quick to assess the outcome, and possible build a quantitative model of the knowledge capabilities of a group of students. Various concept inventories have since been created for application in many fields of study. However, some critics of the technique (Heller & Huffman, 1995; Smith & Tanner, 2010) have suggested that while concept inventories can be valuable as a pedagogical tool and as an evaluation metric for a course, they are still lacking when it comes to making decisions about individual students.

The problem arises in that student’s misconceptions are largely personal, being organically formed as a subliminal self-explanation or understanding of a physical phenomenon (Chi & Roscoe, 2002). If student responses to a concept inventory correlate strongly with experts’, this could indicate that the student holds a well-developed framework for understanding the underlying phenomena and relationships, similar to experts (Michelene T.H. Chi et al., 1979). However, student misconceptions triggered by distracters in a concept inventory, as termed by Hestenes et al. (1992), likely do not represent a common misunderstanding shared by all novices, due to the personal nature of those misconceptions. In addition to this drawback, Heller and Huffman (1995) posed the necessity of factor analysis or other statistical interpretation of student response patterns to validate the meaningfulness of the information. While the individual inventory items may accurately target intended misconceptions, at issue was the possibility that the interplay between items in the overall collection could engender further, unanticipated misconceptions resulting from the structure of the assessment instrument. A meta-analysis of response trends may account for this.

Concept inventories can give a good representation of how closely a student’s performance tracks with a prescribed understanding, emphasizing normative student thinking, but they do not necessarily describe an individual’s reasoning. In particular, concept inventories are not designed to track changes in incorrect reasoning and these changes may be vitally important to guide instructional feedback. Additional analysis is still needed to link non-expert student performance on a concept inventory to an individual student's conceptual understanding of a subject.

1.2 Moving Beyond Concept Inventories

Progressing an assessment of student conceptual understanding beyond a concept inventory scoring poses a challenge for researchers. Evidence of student learning is not neatly available in worked out homework problems or recorded test answers, in fact previously mentioned research shows these artifacts to be commonly misleading. Researchers interested in analyzing student learning must first generatively gather that data. It can be challenging and researchers are often faced with a large, ambiguous set of evidence. The materials typically used to assess deep conceptual understanding include written responses to open-ended questions and audio recordings of one-on-one interviews. The complexity of the data necessitates human interpretation of the subject's behavior in order to grasp the reasoning behind the actions, a process described using the umbrella term of qualitative analysis.

Qualitative analysis is a methodology rather than a specific approach, one which can be likened to practicing a craft of inductive reasoning (Saldana, 2009) while implementing an informal collection of best research practices. The reliability and importance of findings is determined by the competence of the researcher and there are few safeguards against misinterpretations. This is not a prescriptive process and is largely heuristic driven, based on expertise and intimate familiarity with the dataset (Saldana, 2009). It often takes the efforts of multiple researchers to gain greater confidence in the generalizability of the research findings, as often demonstrated through metrics such as inter-coder reliability (Hruschka et al., 2004). This process is often hampered by limited resources and time available in a classroom setting.

Many analytical approaches and software tools have been developed to aid the qualitative researcher operating under time and resource constraints. Software suites falling under the category of Computer Aided Qualitative Data Analysis Software (CAQDAS) are often employed to manage large quantities of data and automatically perform a variety of complex analysis procedures (Carcayry, 2011; Jones, 2007). The core features of a CAQDAS program are the ability to search content, apply codes, create links between data, provide query tools, allow the user to make annotations, and provide the ability to create inter-data networks (Lewins & Silver, 2009). Three examples of widely known commercial CAQDAS programs used are Atlas.ti, NVivo, and MAXQDA (Saldana, 2009). Each of these programs offer the researcher a slightly different approach to performing a detailed qualitative analysis, the relative merits of which are not covered in this discussion. There is a large body of work that exists (Carcayry, 2011; Jones, 2007; Lewins & Silver, 2009) to aid the researcher in becoming a more proficient practitioner of qualitative methodologies using these complex software suites. Those materials typically give a broad framework of best practices for approaching rigorous qualitative analysis with CAQDAS. It is up to the user to iteratively develop the expertise needed for effective research. Ultimately, it is the still the interpretive insight of the researcher that is the basis for the quality of the research.

2 BACKGROUND

2.1 Introduction to Automated Text Analysis Tools

The development of automated research tools to analyze the meaning contained in text, with respect to the natural language used in daily communication, came into focus in the early 1960s with the refinement of grammar-rule based approaches to automatically parsing language. The approach quickly matured and developed into practical applications (Hobbs, Walker, & Amsler, 1982; Sager, 1981). Early researchers focused on developing the computational approaches for automating analyses of linguistic structure in order to facilitate processes such as machine translation, speech recognition and synthesis, and keyword recognition (Hirschberg & Manning, 2015). Despite long standing academic interest in programmatic parsing of human language, the limited computational power and analysis techniques of the time limited the practicality of such approaches. However, constant advances in computation approach, rapid increases in casually available computational power, and the proliferation of open source software tools have dramatically increased the feasibility of small scale text analysis implementations.

The widening availability of text analysis software has led to a variety of implementation strategies and goals. These may range from a simplistic count of words with a predefined meaning, such as tracking the frequency of emotion words to estimate Facebook user sentiment (Settanni & Marengo, 2015), to a more dynamic analysis, like automatically assessing the correctness of student responses in a classroom setting using complex machine learning algorithms and artificial intelligence (Nehm, Ha, & Mayfield, 2012). To understand the similarities and differences in these different approaches, it will be useful to define some of the broader categories of technologies. Lexical analysis describes a general methodology by which the content of spoken language or written text is assigned a natural language meaning through the use of syntax examining processes. In this process, literal character sequences are given language-based meaning and can be further examined. This type of analysis is widely used in Computational Linguistics, a field that fuses statistics-based modeling of the meaning contained in text and a study of the use of human language. Traditional computer science approaches are combined with the study of computational linguistics to form the discipline of Natural Language Processing (NLP), which is intended to "…learn, understand, and produce human language content" (Hirschberg & Manning, 2015) . It is a highly interdisciplinary field that can be applied in a variety of contexts for many different purposes.

2.2 Trial Implementations of Automated Text Analysis Tools

Developing a natural language processing approach is a growing interest in the academic field of assessing student understanding but has yet to be formalized into a common methodology. NLP is still in relative infancy and there are many equally valid approaches to addressing specific features of the process, without methods emerging as universally superior (Joshi, 2003). A recent application (Goncher, Boles, & Jayalath, 2014) attempts to use automated text analysis to judge student understanding by extracting high level concepts from open ended written responses to survey questions. Their method was to employ Leximancer, commercially available automated text analysis, to automatically code a text sample, with the intention of highlighting a student's thinking and reasoning. A brief examination of the Leximancer product website shows that the software is meant to be ready for use with minimal input from the researcher. Goncher, et al. used open ended, written responses to survey questions as a dataset. This study demonstrates some positive benefits of the approach in terms of the speed of analysis but at the cost of reporting a mix of correct and incorrect categorizations of student understanding. The automated portion of this study generated coded findings that were consistently different from those of human researchers. Furthermore, software derived thematic conceptual extractions were based on meaningless vocabulary 87% of the time (Goncher et al., 2014).

In a similar vein, work conducted by Verleger & Beach (2014) also apply NLP tools to interpret open ended student response. However, their study was directly focused on the development of a custom software tool to influence team formation and assigning reviewers based on the content of academic papers. The process involved extracting complex content analyses based on student writing. The software they created looked at natural language text in order to generate coded measures such as audience readability and the occurrence of supportive argument rationale, which was then compared to a manual coding. Their tools had an accuracy as compared to human reviewers of 60 to 85 percent. In contrast to the commercially available Leximancer software, Verleger and Beach's implementation involved considerable 'training' of the software to be able to detect idiosyncratic features of their specific dataset. This involved much more input on the the part of the researcher but led to more accurate results.

These two examples are representative of the ambitions researchers have for implementations of NLP based software but display lackluster results. However, these cases also exemplify the potential trajectory of the approach seen in the field, as a whole. Developing targeted implementations of NLP based tools to be used on a specific datasets appear to yield much more useful results than generally exploratory applications.

2.3 Implications and Contemporary Approaches

Automatically analyzing the abstract content of a text sample could be an extremely powerful tool once strong evidence of broad generalizability and accuracy is established. A corollary approach to enhancing a qualitative analysis would be to avoid direct, high level abstraction but rather decompose those features into simpler elements to be used in a more traditional research investigation. A promising path is looking at the underlying structure of written text for contextual grammatical features, or lexical structures. A goal of this approach is to glean some information about the content by examining the way in which it was communicated, possibly revealing phenomena with meaningful implications. A major inspiration for applying this approach to judging student conceptual understanding comes from the work done by Lu (2012), where natural language processing techniques were used to reveal a strong correlation between the written structure of ESL student's work, task performance, and their proficiency with English language communication.

Lu's 2012 study employed NLP tools to evaluate transcript-based data from over 400 students and was able to extract information describing meaningful lexical structures independently from the literal content of the communication. Researching the application of natural language processing techniques to extract lexical information (the structure and context of the communication) of written text may avoid the pitfalls of trying to automate high level concept analysis. It is not anticipated that such a technique will eliminate the need for an in-depth qualitative analysis of student conceptual understanding but rather to supplement it. This work is largely exploratory in nature, combining elements from historically independent fields with the intent of identifying useful commonalities. The analysis will focus on pre-existing records of communication between individuals and researchers, specifically targeting text interview transcripts.

3 THEORETICAL BASIS

3.1 Interactions between Conceptual Understanding, Cognitive Load, and Lexical Features

Before searching for evidence of how a student's conceptual understanding changes over time, it is necessary to describe what we mean by the term and what it means to 'see' evidence of conceptual understanding. Conceptual understanding is the fundamental knowledge a person has about a specific phenomenon, without performing a rote calculation or methodology. It is an underlying personal truth that describes how and why world works (Montfort et al., 2009). This internalized framework of understanding is more akin to experience-reinforced intuition than academic knowledge. Conceptual understanding can be used as an indicator or labeling tool to help describe and make sense of the phenomena of student learning, but does not specifically explore the underlying mental processes that lead to development of expertise in a subject. There are a variety of approaches for bridging the gap between theorized descriptions of knowledge formation and underlying phenomena. Such tools may involve innate features within a person or that subject's varied social experiences. This study focuses on one particular avenue of tying the formation of conceptual understanding to a wider set of mental features by means of employing Cognitive Load Theory (CLT).

CLT is an area of research that posits a connection between how the physical brain functions and the thought processes it performs. A simplistic understanding of CLT is that a person has a finite capacity for varying types of mental processes, based in the physiology of the brain itself (Kalyuga, 2009). This capacity can be modeled as limited resource and the way in which that resource is used during a particular mental task, or allocated when performing simultaneous tasks, can yield valuable information about the individual. The amount and type of mental resources devoted to a task is termed cognitive load. One particular assertion in CLT is that there is a measurable relation between working and long term memory, as indicated by different types of cognitive load. Germane cognitive load has been connected to deep transferrable knowledge structures (Kalyuga, 2009) that are often representative of conceptual understanding. As a person's conceptual understanding of a subject increases, the level of cognitive load required to make deep connections within that conceptual domain is lowered. An expert in a field is able to make deep conceptual connections between features without this dominating available cognitive resources. Contrastingly, a novice's cognitive resources may be monopolized when employing a simplistic, surface level analysis. We can then use measures of cognitive load to indicate how well structured a conceptual area is for an individual, which implies a certain level of conceptual understanding in that area. An individual showing low cognitive load while displaying features of a highly interconnected conceptual domain points to high conceptual understanding in that area.

CLT partly focuses on exploring the cognitive factors that affect a learner's acquisition and development of schemas, or structures that contains a deep understanding of a particular problem domain. Abstract thought is seen to have a related physiological component. The idea that a person has only a limited capacity for cognitive processing at any given time gives rise to the concept that tradeoffs between various cognitive processes (with associated physiological bases) exist in an individual (Chandler & Sweller, 1991; Sweller, 1988). Sweller (1988) notes the conflicting ability to engage in schema attainment and revision versus directed goal attainment; problem solving processes with implications of differing levels of conceptual understanding. In other words, if a person's cognitive resources are dominated through elementary problem solving approaches within a poorly connected concept area, it is very hard to simultaneously develop or revise concept connections inside that same area.

Measuring cognitive load, therefore, becomes a bridge between measurable physiological performance and certain thought processes. Testing methods for measuring relative levels of cognitive load in an individual was central to Chen et al.'s (2012) study which found that cognitive load could be measured in real time. In this single study, many modes of measurement, such as eye tracking, hand writing analysis, EEG monitoring (targeted measuring of brain waves), and lexical analysis were found to be strongly inter-correlated and that they reliably pointed to levels of cognitive load. This leads to an assumption that one individual measure may be enough of a basis to characterize an individual's level of cognitive load. Chen's summary includes lexical analyses that are shown to be reliable measures of cognitive load. Specifically, Chen used lexical density, which takes into account the variety and sophistication of the words used. Using lexical analysis may preclude the need for expensive physio-cognitive measuring equipment in order to extract information about a person's thought process and can be applied retroactively. Lexical analysis techniques are available through open-source software that is quick to implement and simple to disseminate. By using lexical analysis techniques, we can infer levels of conceptual load and make judgements about an individual’s conceptual understanding.

3.2 Natural Language Processing Tools

NLP tools involve complex linguistics procedures that rely on 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. For the sake of brevity, we will be briefly discussing what these procedures do, rather than heavily detailing how they work. The terminology, definitions, and use of software in this section are detailed by Bird, Klein, and Loper (2009).

An important first step in applying NLP is determining the Part of Speech (POS) for each word in a text. The meaning behind the words we use are largely ambiguous and dependent on the context of their use. The meaning of each word will be important to for analysis in many ways. Consider the two different POS parsing trees shown in Figures 2a and 2b that chart the decomposition of one part of a common humorous phrase.

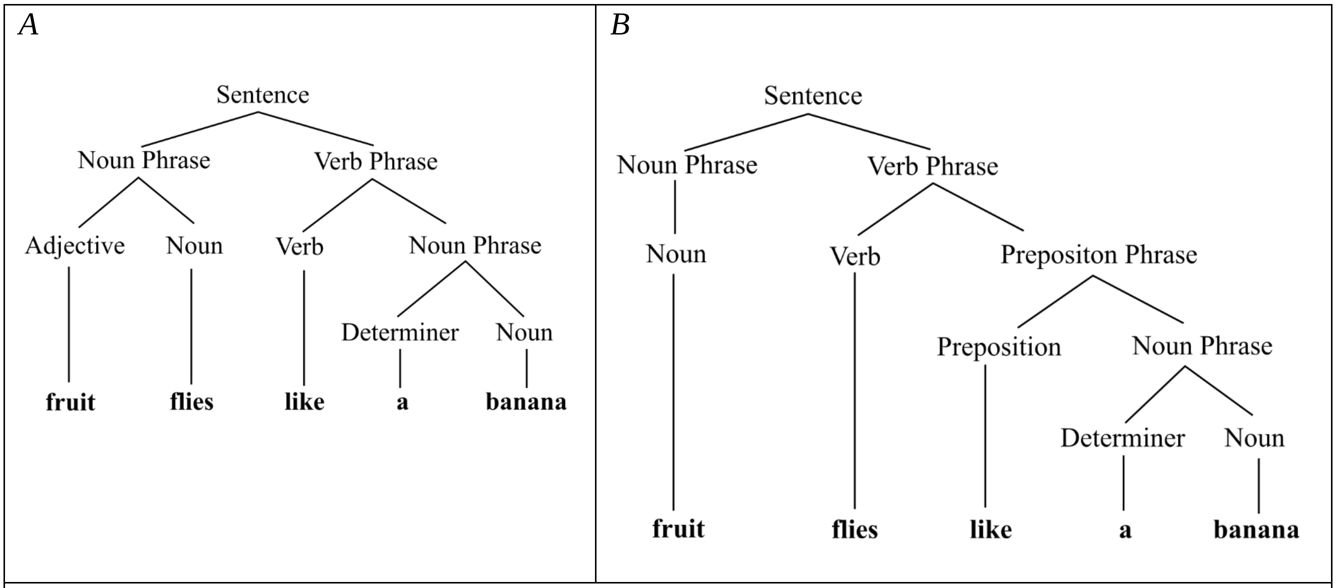


Figure 4A and B: Illustration of possible syntactic POS parsing tree ambiguity

To a human reading the sentence: “Fruit flies like a banana”, humor arises as a result of the two simultaneously existing and conflicting meanings of the sentences. Is the subject a tiny, but enormously aggravating insect or an incorporeal concept that happens to shares flight characteristics with a certain yellow 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 decide from the context if it is more likely to be a claim about insects’ diets or fruits’ modes of transportation. It must employ sophisticated algorithms to create probabilities of various interpretations being correct. Even then, it cannot be sure.

The analysis makes use of the Stanford POS tagger (Toutanova, Klein, & Manning, 2003), which is a piece of 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 (Santorini, 1991). 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 of over 90%, even when judging unknown words (Santorini, 1991). 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. A lemma is a core meaning, or concept, that can define a group of related words. Lemmatizing 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 relationships for a specific culture. The meaning behind each word, rather than the literal character sequences are the subject of our future lexical analysis. Figures 3a and 3b demonstrate this effect.

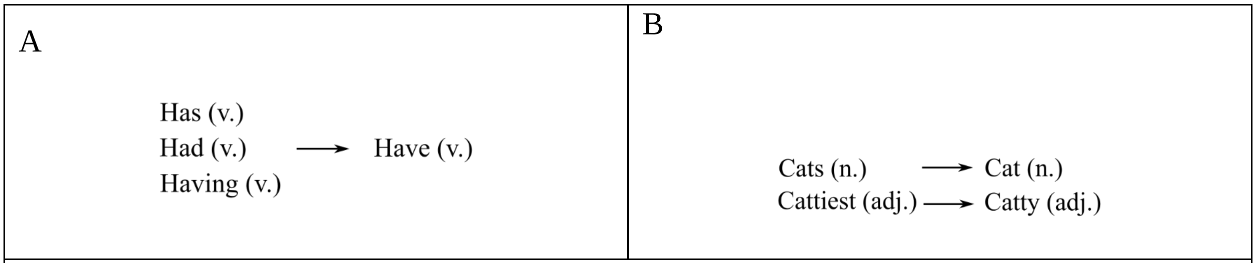


Figure 5A and B: Lemmatizing by word tense and by part of speech

Figure 3A shows how different tenses of the word "have" can be collapsed into one meaning, or lemma. Figure 3B demonstrates the importance of part of speech in lemmatizing. "Cats" may refer to the concept of a small feline but the adjective "Cattiest" is likely not 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. This study relies on a custom python program developed to interact with the Stanford POS Tagger and create carry out the lemmatization process. Using Python NLTK allows us to interact with Princeton WordNet (Miller, 1995). 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 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 (Lu, 2012). It was originally developed to generate lexical indicators of English as Second Language learners' written text but can be adapted 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. Possible lexical statistics generated using an adaption of Lu's code are of a similar type that Chen et al. (2012) correlated to measures of cognitive load. This study generates lexical indicators (ie, statistical descriptors of contextual language use) for a text to generate implications of a person's cognitive load, as informed by Chen's methodology.

Computational linguists have developed many statistical indicators to describe features of communicated language (Lu, 2012). Some indicators are more useful for certain types of text than others; for example: short vs long, complex vs simple, word choice independent vs dependent, et cetera. The lexical indicator most appropriate to the interview transcripts analyzed in this study is the Uber Index (Equation 1), as developed by Dugast (1979) to examine the relationship between the lexical types of words (ie, function descriptions) used and overall text length. The Uber Index is a descriptor of the lexical diversity of a text, or the relationship between the total numbers of lexical words (T) used and total number of lexical word types (N) used in a text. A high Uber index value would indicate that complex sentence structures are being used in an individual's communication.

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

Equation 1: Uber index from Dugast (1979)

This index was chosen prior to analysis because Uber is seen as as a better representation of lexical diversity for for texts of varying length than other similarly used lexical indices, as found by (Jarvis, 2002). This relaxation of a text length requirement is useful to the study, as detailed further in the Methodology Section. Another reason Uber is appropriate for this study is that it is very closely related to the specific indicator used in Chen's study without placing value on the specific words used (Šišková, 2012). This study focuses on conceptual understandings of core engineering ideas. It was anticipated that individuals trained to solve these problems in a similar setting would use similar words in their responses even if their conceptual understandings were quite different. That complication can be avoided by focusing on lexical features that are independent of word choice. Uber is appropriate for our short text samples and linked to a component tested for in Chen's 2012 study, which allows us to generate representations of cognitive load from on our analysis.

If lexical structures can be linked to task performance and cognitive load, independently of the content of the communication, perhaps some metrics methods can be developed that that allow a researcher to link those same lexical structures to task performance that a researcher may associate with infer an individual’s conceptual understanding in a targeted subject area using only text-based records. Recent improvements of freely available NLP software (Bird et al., 2009; Miller, 1995; Toutanova et al., 2003) have lowered the barriers to implementing such an approach, especially in automated applications. The tools in this discussion are interacted with using a by means of custom GUI developed in Python, which gives the user a simple point and click approach to using the described NLP tools simultaneously. This research is a preliminary investigation and uses simple implementations of the NLP tools on an easily accessible dataset to test the validity and utility of the approach.

4 Methodology

4.1 Procedural Overview

The research findings of this study are the results of a direct comparison between a qualitative-research based assessment and an exploratory, Natural Language Processing based assessment of an individual’s conceptual understanding of core engineering concepts. The research subjects involved in this study had been educated at the same university and engineering program. More details of of the participants, selection process, and multi-year study timeline are described in the next section.

A representation of the qualitative research process is shown in Figure 4, below. A coding framework was developed and applied to interview transcripts using the qualitative research suite, AtlasTI. This process aimed to assess the relative change in a subject's conceptual understanding of a Mechanics of Materials and Fluid Dynamics based topic, on a three year time scale. Figure 5 shows the software based NLP process that tracks the change of lexical traits from the same time span After the individual analyses had been completed, the results were compared for the purpose of finding correlations between the approaches and identifying potential effects of using the results of the automated NLP process to inform a future qualitative analysis of conceptual understanding.

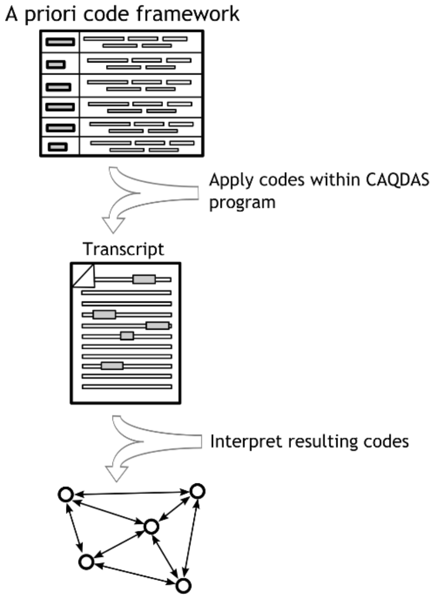


Figure 4: Overview of Qualitative coding method

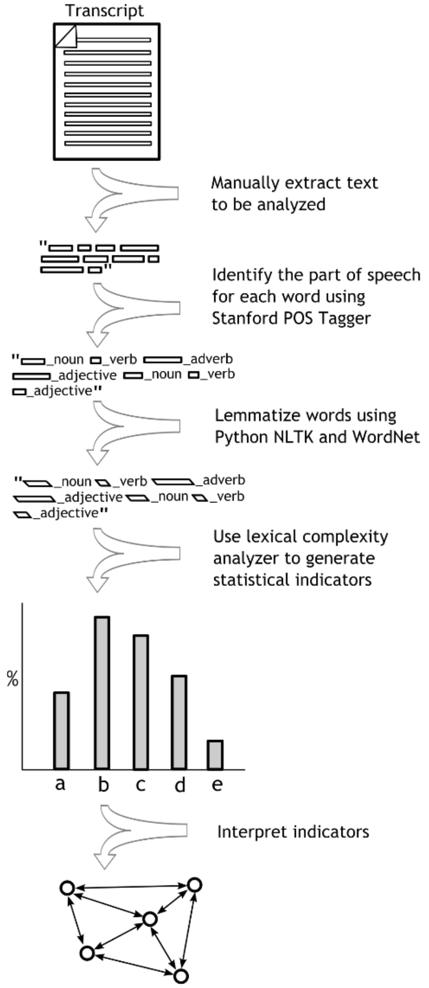

Transcript
Manually extract text to be analyzed
Identify the part of speech for each word using Stanford POS Tagger
"_noun _verb _adverb _adjective _noun _verb _adjective"
Lemmatize words using Python NLTK and WordNet
"_noun _verb _adverb _adjective _noun _verb _adjective"
Use lexical complexity analyzer to generate statistical indicators
%
a
b
c
d
e
Interpret indicators

Figure 5: Overview of Automated Natural Language Processing method

4.2 Dataset and Interview Methodology

This paper draws on three years of longitudinal data on entry level engineers' conceptual understanding of solid and fluid mechanics. This study began with 12 participating subjects that were selected from the senior-level civil engineering student body at a land-grant public university were tracked as they transitioned into engineering professionals. Interviews were initially conducted in person to allow for a natural rapport to develop between the interviewer and subject, then continued on a regular basis remotely. Over the course of the three year period, some subjects dropped out and some did not participate in the entire interview sequence. Only six subjects were found to suitable for direct self-comparison over the course of the interview time frame. Despite the small sample size, a large quantity of data was collected for each of the six subjects over a multiyear time span. Each participant in the study was asked to engage in five sets interviews between 2011 and 2013, which covered their senior years of undergraduate education though their second year of professional engineering practice. Figure 6 shows the timeline of the interviews and the relationship of repeated interview questions over that three year period.

The average length of these interviews was 90-minutes, so about 7.5 hours of raw interview audio recordings were generated for each of the subjects. Transcripts were created from the resulting audio. In each interview, the subject was asked questions about a pair of solid and fluid mechanics problems, designed to highlight the conceptual understanding of the participant. Upon reaching the third year of the study, the interviewees were reintroduced to the same interview questions they first saw in 2011 to allow for direct comparison. The overall quantity of interview data generated for each of the six subjects is quite extensive, however this study only makes use of a subset of that data, which pertains to specific concept areas described later. The subset is comprised of about 3 hours of recorded interviews for each of the six participants.

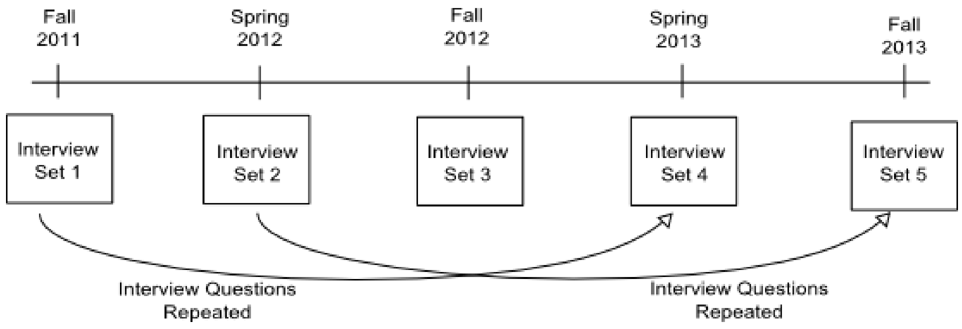


Figure 6: Timeline of Interviews

The students initially had a similar educational background, having progressed through a standardized curriculum. This study was aimed at capturing emergent phenomena of changing conceptual understanding as the students transitioned into varying professional engineering vocations.

The interview sessions were semi-structured and relied primarily on verbal interaction as method for gathering data (Case & Light, 2011). Galletta and Cross (2013) explain the process of designing a semi-structured interview protocol, collecting the data, and analyzing the results. A standard list of questions was prepared for individual interview sets and the experienced interviewer was able to reflexively interact with the students to tease out the underlying meaning in the student's response. (Barriball & While, 1994) have highlighted the advantages of conducting semi-structured interviews. Positive features of this interviewing structure include the ability to react to differences in the personal and educational histories of the interviewees, as well as the ability to clarify issues raised by the sample group on a person-by-person basis. (Barriball

& While, 1994) also emphasize the benefit of asking reactive, probing questions to decrease the influence of social pressures felt by the respondents in the interviews. Audio of the interviews was recorded and text transcriptions of the sessions were created for ease of analysis. The two forms of media were compared for accuracy and cross referenced to provide insight into the analytical information presented by the interviewee and to allow interpretation of nonverbal discourse.

The problems posed in the interviews were designed to represent material similar to what the subjects encountered in their coursework and targeted conceptual areas that the interviewees would possibly be expected to engage with in their professional life. In this way, the study intended to contrast how conceptual understanding developed in an academic context changed as that understanding was refined through practice. The interview questions were also open-ended enough to allow for the interviewer to pursue many possible lines of questioning. Visual depictions of two representative problems are shown in Figures 7 and 8.

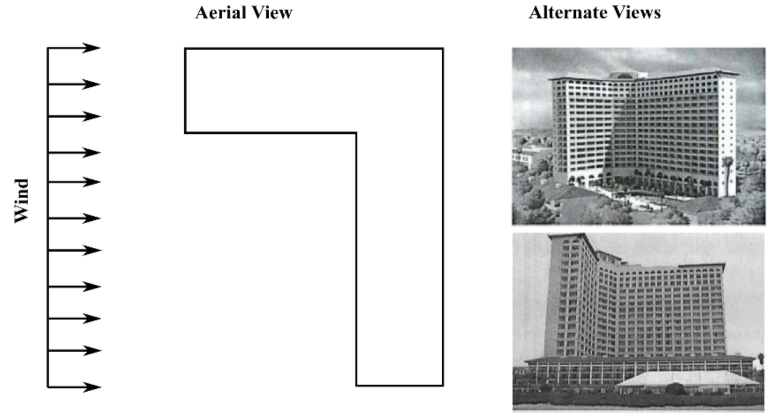


Figure 7: Static Wind Loading Visual Description

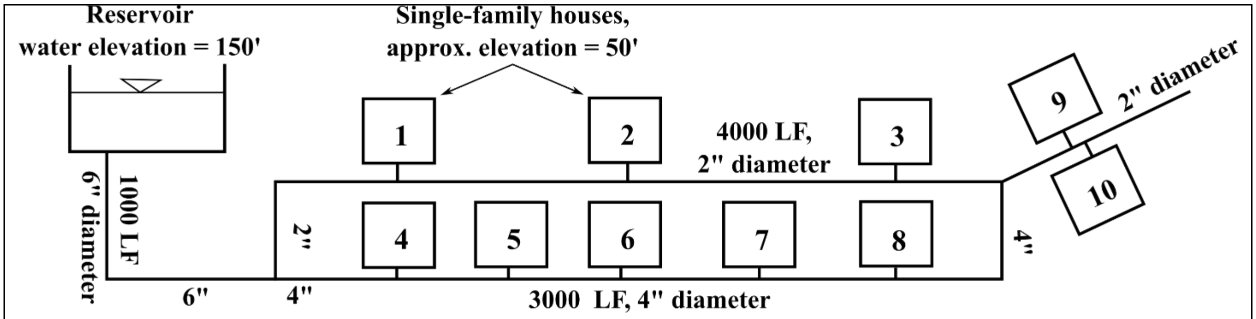


Figure 8: Residential Pipe Network Visual Description

These two questions are only a part of the interview process and make up roughly half of the analytically based interview prompts. The questions posed in the interviews were designed to avoid encouraging the subject to apply rote methodologies or equations. In the structure loading problem (Figure 7), the subject was asked where the building would deflect the most under a given uniform load rather than being asked to quantify the value of maximum strain. The interview protocol used with the pipe network problem of Figure 8 questions about the effects of increased demand on the system rather than calculating frictional losses within a pipe. In general, the types of questions followed a similar format: What information would be necessary in order to describe or the system? How will system changes affect the relationship between state attributes? If a specific output was required, how should the parameters of the system be altered? In this way, the interviewer was trying to tease out what the subject intrinsically understands about the problem domain.

4.3 Coding Protocol and Framework

A basis for the validity of this study is the comparison of the newly applied technique of lexical analysis against the standard practice of a qualitative coding. Each transcript was qualitatively analyzed to describe the conceptual understanding displayed therein. An a-priori coding framework was developed and is represented below in Table 1.

Level of Understanding	Code	Description
Conceptual	**C6**	*Implicit **relationship recognition***
	C5	***Reformulation** of problem (**cross-domain** analogy)*
	C4	***Reflection** on appropriateness of **approach***
	C3	***Reflection** to see if answer makes **sense***
	C2	***Reformulation** of problem (**same-domain analogy**)*
	C1	**Relationship** recognition (**same domain**)
	P3	Step-by-step Approach (**Action Sequence)**
	P2	***Recalling Specific** named equation or definition*
Procedural	**P1**	**Naming general Procedure**

Table 1: Coding Framework

The theoretical basis of this coding framework is largely informed by the works of Rittle-johnson and Schneider (2015), Rittle-Johnson, Siegler, and Alibali (2001), and Van Merrienboer and Sweller (2005). A major feature is the division between codes indicative of conceptually heavy problem (beginning with the ‘C’ prefix) and procedurally based (beginning with ‘P’ prefix) problem solving strategies. These codings contain various explicit and explicit measures of the conceptual/procedural solving continuum but it should be noted that the processes can profoundly influence each other and may be hard to isolate (Rittle-Johnson et al., 2001). The implicit and explicit measures have been adapted from examples given by Rittle-johnson and Schneider (2015) to reflect specific features we might expect the interview subjects to display in their interview question sessions. A second major feature of this coding framework is a relative ranking of codings displaying procedural and conceptual processes. This ranking reflects assessments the cognitive complexity required for each of the coded features. It is suggested that the complexity of the problem solving feature employed is directly related to a learner’s conceptual understanding of the concept domain and cognitive load experienced by the individual (Van Merrienboer & Sweller, 2005). That is to say, a novice in a subject devotes a more substantial amount of cognitive load to a simple procedural approach than an expert might, precluding the possibility of devoting the needed cognitive resources for a more conceptual reflection.

This coding framework is applied to inform a qualitative assessment of conceptual understanding. No direct quantitative values will be drawn from its use. While a hierarchical ranking of codes and a transition from procedural to conceptual processes is supported by research, this specific coding framework has not yet been directly verified. The relative meaning of the magnitudes associated with each code is untested. We can’t conclude that a C6 shows twice as much conceptual understanding for a subject when compared to C3. Similarly, codes may have differing implications from person to person. To counter these ambiguous unknowns, this framework will only be used to make comparisons for specific individuals over time, not for making interpersonal assessments. A large number of Procedural codes and few Conceptual ones generated in a particular interview session will be evidence of a mostly procedural understanding. Similarly, a grouping of high ranked Conceptual codes will evidence strong conceptual understanding. A neutral response will be generated from a broad mix of middling ranked Conceptual and Procedural codes.

4.4 Application of NLP and Lexical Analysis

Before the previously laid out natural language processing steps can be applied, some manipulation of the dataset is needed. The raw texts were complete transcripts of the entire interaction between the subject and the interviewer. This tended to include ‘off topic’ conversation, such as introductions and idle small talk before interview questions were posed. Interview questions were prepared in advance and typically introduced using very similar if not identical wording. All response to standardized interview questions was deemed to be ‘on topic’ and used in the analysis, while ‘off topic’ conversation was ignored. Similarly, interviewer speech was removed to isolate the participant’s speech. Next, we decided to separate the individual subject matters of Fluid Mechanics and Mechanics of Materials. This was to allow for a more direct comparison if the interviewee showed differing levels of conceptual understanding between subjects. The ‘chunked’ text was then processed using previously described NLP software. The following series of excerpts tracks the transformation of a short question response as it is manipulated by researchers and the NLP tools into a measurable lexical quantity. A very short response was chosen to limit the amount of data and illustrate the process clearly. The following is the raw interview text as transcribed by the researchers:

Interviewer: Okay. What about highest energy?

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

The text is then manually trimmed to remove interviewer speech and other non-participant response:

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

The first NLP process of POS tagging can then be performed using the Stanford POS tagger (Toutanova et al., 2003), with associated output displayed below, and the resulting lexical parsing tree is shown in Figure 9:

Probably_RB at_IN the_DT reservoir_NN because_IN it_PRP has_VBZ the_DT least_JJS amount_NN of_IN _________NN ._.

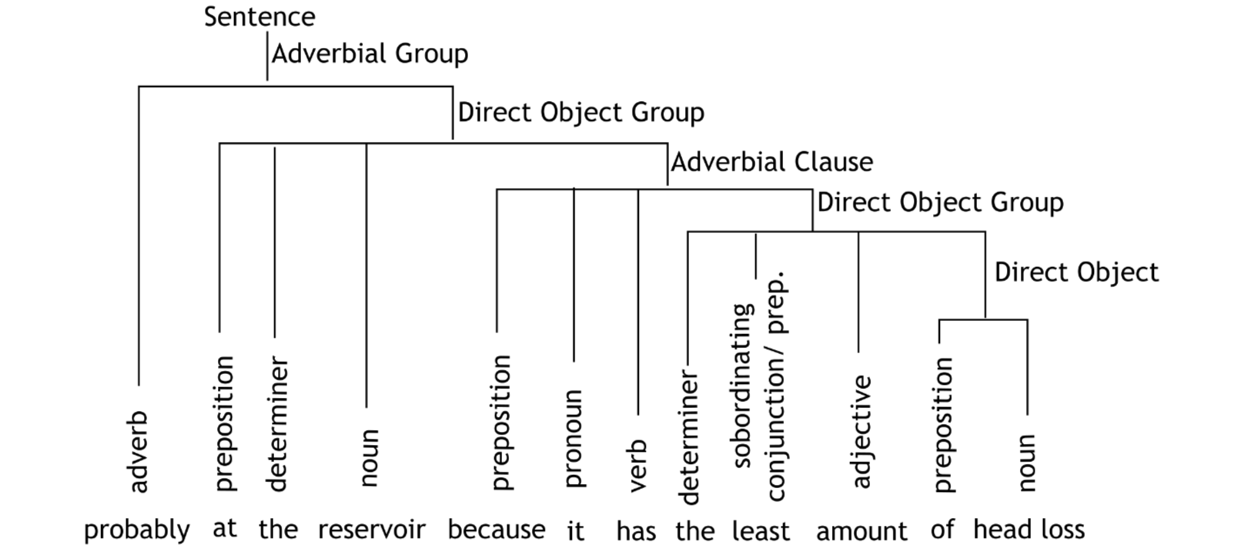


Figure 9: POS parsing tree for NLP application example.

Two transformations of the text are shown. The first is the output of the program using shorthand notation for POS tags. The second displays a POS parsing tree generated using the output, with the shorthand POS tags replaced by longhand descriptions. Two interesting features can be demonstrated here. First, the POS tagging of the sentence is highly accurate. The parsing tree describes a POS structure that recreated the intended communication of the interviewee. The second notable feature is the recognition of the engineering phrase ‘head loss’. Note that this term has been replaced with the stand in ‘____’ because it has unknown meaning to the POS tagger. The POS tagger was able to recognize that ‘head loss’ denoted a particular concept separate from the individual words ‘head’ and ‘loss’. It did this by examining the context of the sentence and judging that ‘head loss’ was the subject of a noun-phrase, indicating a single entity. It should be noted that the term was replaced but the lexical information of intra-sentence relationship and POS type was retained for use in later lexical analysis.

The next transformation is lemmatization using a custom implementation of Python NLTK (Bird et al., 2009) and WordNet (Miller, 1995):

> Probably_RB at_IN the_DT reservoir_NN because_IN it_PRP have_VBZ the_DT least_JJS amount_NN of_IN ________NN

This example is not very especially dramatic due to the simple word choice the interviewee used. One example of lemmatization actively occurs in the excerpt. The word ‘has’ was transformed into its lemma, ‘have’. A similar transformation would have occurred if the text included the words ‘had’ or ‘having’. With this process complete, the lexical analyzer (Lu, 2012) can be applied to compute the Uber statistic. The Uber statistic forms a representation of the lexical diversity of the excerpt and a manual calculation example is shown below:

> Probably^{1} at^{2} the^{3} reservoir^{4} because^{5} it^{6} have^{7} the^{3} least^{8} amount^{9} of^{10} ___^{11}

$$Uber = \frac{log^{2}T}{logT - LogN} = \frac{\log 12^{2}}{\log\left(\frac{12}{11}\right)} = 30.82$$

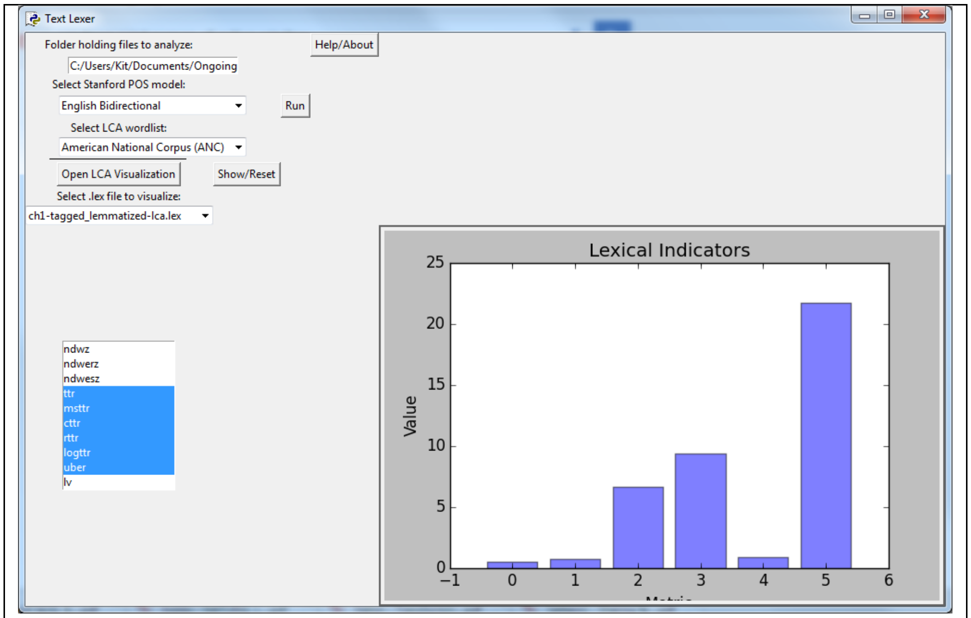
Equation 2: Demonstrated Uber Calculation

In this simple demonstration of calculating an Uber index value, each unique lexical word is given an increasing, superscripted numerical label. The only repeated lexical word in the short sentence was ‘the’. The total number of lexical of words (T=12) and number of lexical word types (N=11) are used to calculate the Uber index (Equation 1). It should be noted that while Uber is appropriate for varying length texts (Jarvis, 2002), the sample sentence used is many orders of magnitude smaller than the overall transcript dataset. Consequently, the numerical results cannot be directly compared. Indeed, lexical statistics are intended to describe characteristics of entire manners of communication and quickly lose demonstrable meaning when applied to small fragments of text. Nearly every word in this example is unique, showing that and Uber value of 30 is an extremely high lexical diversity. In this case, only one word was repeated and this is unlikely to occur in actual speech. Having virtually no repeated lexical word types in a text dataset constitutes an extreme case.

As an illustration, consider a text made of two exact duplicates of the sentence "probably at the reservoir because it have the least amount of ___". In this case (T=24, N=11) and the Uber value would be about 5.6. If the sentence was duplicated 10 times (T=120, N=11, the resulting Uber value would equal about 4.2 (note that this is a decreasing logarithmic function as N goes to zero). Together, these examples provide a reference of extreme high value (30) and an extreme low value (5). It may be hard to predict the actual upper and lower bounds of lexical diversity expected from a participant in this study, but the extremes discovered in this simple example inform the range of values we might expect to see. In addition to this single sentence example, values of Uber index from other studies (Jarvis, 2002; Lu, 2012) tend to fall in the range of 11-26. These values were established for different datasets but indicate the general range we can expect to see in typical speech.

Although these methods rely primarily on the freely available available software programs already mentioned previously, some programming was needed to combine the disparate elements and create an all-in-one GUI. These open source tools were written in differing languages and program inputs/ output are not standardized. These tools typically see use in ad-hoc applications. Figure 9 shows a screen capture of the all-in-one software graphical user interface that allows the simple importation of text, selection of metrics, and display of analysis results. The software created specifically for this research makes it easy to select text files for analysis and quickly display visual results in one window. It gives the researcher easy control over variables that influence the POS tagging, lemmatizing, and lexical analysis procedures. In addition, it is not tied to any specific application. It is flexible enough to allow the researcher to simultaneously run lexical analyses on an arbitrary number of texts, manipulate many options, and chose which lexical indicators should be displayed.

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 secondary goal of this project was to introduce a computational linguistics approach to a field of research where is shows promise. As such, the interface for the tools needs to be simple to apply and robust in operation.



Text Lexer

Folder holding files to analyze: C:/Users/Kit/Documents/Ongoing

Select Stanford POS model: English Bidirectional

Select LCA wordlist: American National Corpus (ANC)

Select .lex file to visualize: ch1-tagged_lemmatized-lca.lex

Figure 9: Python based GUI for POS tagging, Lemmatizing, Lexical Analysis and visualization of results. Results are exported in a '.csv' spreadsheet format.

5 Results and Discussion

After the qualitative coding and lexical analysis processes had been completed, the results were compared on an individual and group-wide basis.

5.1 Overview

The following table compiles and visually compares the results from both the qualitatively coded process and the automated, NLP-based lexical analysis (Uber index). The results for the six pseudonymous study participants have been grouped by interview problem set to show change over time in a particular subject area. Change in conceptual understanding is represented in three possible states: increase, decrease, or no change. The evaluation was the result of a qualitative assessment of the quantity and values of codes generated according to the coding framework, shown previously in Table 1. Change in Uber index is presented as two possible values: increase or decrease. This result was a simple quantitative comparison of the value generated by the Uber formula, show previously in Equation 1. The color of each cell indicates if the two approaches agree in their assessment of change in student conceptual understanding over time. A green shading represents that the results of the qualitative analysis and lexical analysis agree. A red shading shows disagreement. No shading shows that no direct comparison can be drawn as an effect of unchanging conceptual understanding.

Subject Area Name	L-Building 2011⟹2013	Pipe Network 2011⟹ 2013
Phillip	🟩 C.U. ↑ Uber ↑	🟩 C.U. ↑ Uber ↑
Roberta	🟩 C.U. ↑ Uber ↑	⬜ C.U. ⇕ Uber ↓
Brian	⬜ C.U. ⇕ Uber ↑	⬜ C.U. ⇕ Uber ↑
Terrance	🟧 C.U. ↑ Uber ↓	🟩 C.U. ↑ Uber ↑
Zander	🟧 C.U. ↓ Uber ↑	🟧 C.U. ↓ Uber ↑
Sally	⬜ C.U. ⇕ Uber ↑	⬜ C.U. ⇕ Uber ↓

Symbols

C.U.↑/ C.U.⇕/ C.U.↓ : Conceptual Understanding Increase/ No Change/ Decrease

Uber↑/ Uber↓ : Uber Index Value Increase, Decrease

🟩 : Agreement between Uber and Conceptual Understanding Change

🟧 : Disagreement between Uber and Conceptual Understanding Change

⬜ : No direct comparison between Uber and Conceptual Understanding Change

Table 2: Relative change in conceptual understanding and Uber index by subject area, over time

5.2 Change in Conceptual Understanding

5.2.1 Results

The qualitative coding of conceptual understanding for the majority of participants shows either positive or no change at all, with equal frequency. A Decrease in conceptual understanding is only seen in one individual from the study, simultaneously occurring in both problem domains. Only one other person showed consistent change, in the case of conceptual understanding increasing across both domains. Additionally, most individuals showed Increase of conceptual understanding in the L-Building problem rather than the Pipe Network, but only by a small margin. Occurrences of No Change in conceptual understanding was seen moderately often in both problem domains. Decrease in conceptual understanding was seen the least, in only two out of twelve occurrences.

Most codes created in this study were grouped around moderate value Procedural problem solving features or low Conceptual problem solving features. No occurrence of the highest ranked conceptual code (C6) was seen and only sixteen C5's were recorded. High ranked Conceptual codes were rare in the overall study. Additionally, high ranked Conceptual codes did not typically indicate a low number of Procedural codes would be present. No significant tradeoff, or mutual exclusion between Conceptual and Procedural was seen.

5.2.2 Discussion

The relative lack of change demonstrates the durable nature of conceptual understanding and knowledge frameworks. Once a conceptual understanding is developed, it is very hard to change and doing so often requires a metacognitive approach. This is still surprising given that the participants in the study are actively using, and presumably building upon, their engineering knowledge in their daily professional lives.

Interviewees often voluntarily brought up connections between the concepts discussed and their professional work, particularly when giving supporting evidence. A general trend, which existed outside the scope of the coding framework, was that many more personal examples or explanations were drawn when discussing the L-building problem. Job responsibilities for some of the participants included designing or assessing physical structures, analyzing loading scenarios, and generally dealing with mechanics of materials subject matter. This might be expected to raise the occurrences of increasing conceptual understanding but this was only reflected in a small number of participants.

Participants also tended to respond with a coding that fit an expectation created by the interviewer's question. A Procedural code often resulted when the interviewer asked a simply directed question, such as: "Where is the highest pressure in the system?". The responses to this type of question were often short and cited a single piece of evidence or simple reference to an equation. There was usually no further exploration once the participant had given, as seen by themselves, a 'correct' answer. Conversely, the highest value Conceptual codes were only portrayed when participants were asked to describe general features of a system or make judgments of the relative importance of problem features, or relationships. This is not entirely unexpected. When there is no perceived single answer, solvers are more likely to follow connected concepts to seemingly divergent ends within a conceptual understanding. A response of this type would indicate that the individual holds a well-developed conceptual understanding. It is perhaps an unsurprising correlation to note the infrequency of such responses with the low occurrences of positive changes in conceptual understanding.

5.3 Change in Uber Index

5.3.1 Results

The change in Uber index was fairly constant across the entire dataset, with nine out of twelve cases showing increase. To further examine Uber response, the individual Uber indices for all participant interview sessions are compiled in Figure 10.

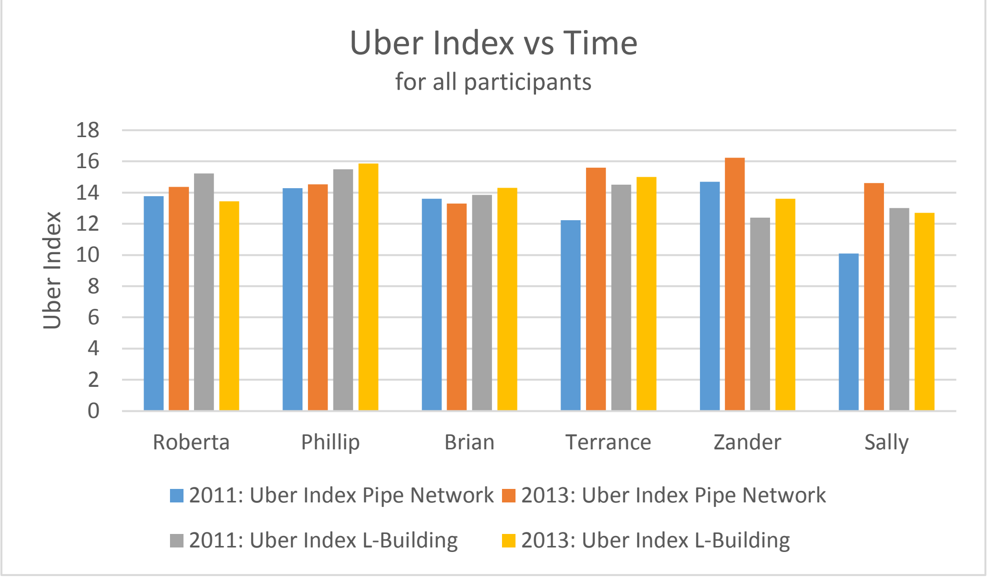


Figure 10: Combined comparison of participant Uber Indices over time

While it is clear that while there are relative changes in Uber value over time, the magnitude of those changes is very small. The median value for all the responses is 14.29 with a standard deviation of only 1.3. This study determined whether Uber increased or decreased over the three year time span but the magnitude of the change was not used as a factor. The largest changes in Uber index were found with participants Terrance and Sally, each seeing increases of nearly four points. Through the data, there were only two instances of a decrease in value from year to year within the same problem domain. Magnitudes of Uber scores were not utilized in determining relative levels of increase and decrease because of the difficulty of making direct lexical comparisons between individuals in such in a small sampling. Creating a metric influenced by the magnitude of change was not seen to be justified considering the relatively small sample size.

5.3.2 Discussion

While it is possible that the magnitude of change in Uber index has real meaning in this study, some drawbacks of adopting that approach were discovered. The literature finding Uber to be a good representation of lexical diversity for variable length texts (Jarvis, 2002) did not specifically test using Uber as a direct comparison between texts showing extremely large differences in length, such as we saw in the study. It was found to be the most accurate way to describe the lexical diversity of an arbitrary length text, but these results tend to call into question use in our specific application. Nonetheless, the literature supporting the index is leads us to believe that it is still useful as a metric to supplement an assessment of student conceptual understanding but may appropriate to quantify it.

The values of Uber represent lexical diversity of a text and generates an effectively unit-less index, calculated from the the total number of lexical words (T) and the number of different lexical words used (N). The Uber index described in Equation 1 shows that the response to a linear variation of either variable produces a characteristic logarithmic curve. The Uber results for Brian will be show in an exploratory example. Brian’s example was chosen to illustrate some specific features that complicate using the magnitude of Uber as an evaluation metric.

Two data points were chosen based on the similarity of calculated Uber index to highlight some unexpected findings. The value calculated in the 2011 Pipe Network interview excerpt was 15.22 and the value calculated in the 2013 Pipe Network interview excerpt was 14.36. It might be tempting claim that the difference between the values are minute compared to their magnitudes, indicating that lexical diversity did not appreciable change between these two interviews. However, some factors complicate that conclusion. First, the lengths of the excerpts, or the total number of lexical words, (T), were quite different. The 2011 excerpt had 557 lexical words and the 2013 excerpt had 1920. The characteristic curves were generated using N as the independent variable and Uber value as the dependent Table 3, below:

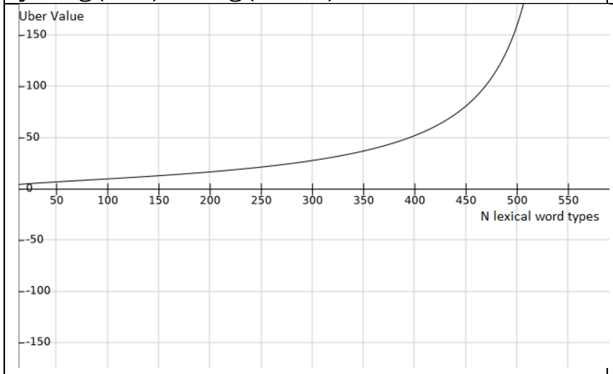
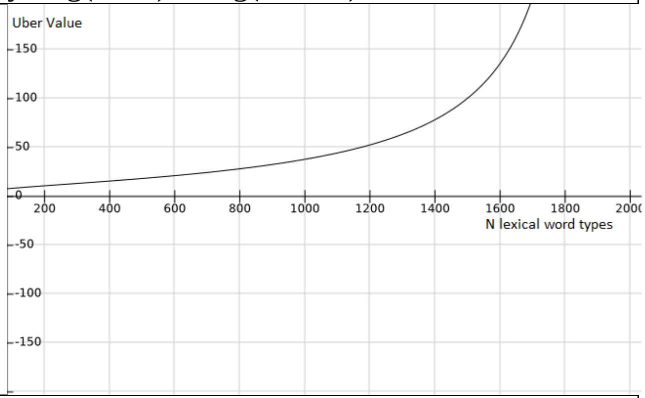
Equation: y=log(557)^2/log(557/x)	Equation: y=log(1920)^2/log(1920/x)
	
2011 Pipe Network T=557, N=132 Calculated Uber= 15.22	2013 Pipe Network T= 1920, N=260 Calculated Uber=14.36

Table 3: Exploration of Uber Index – Brian

These graphs represent a total possible range of Uber values (Y-axis) for text length 557 and text length 1920. As lexical diversity increases, the Uber value logarithmically increases toward an asymptotic value equal to the text length, T. The intents behind the linguistic derivation of the Uber value was intended to create a good representation of the lexical diversity of a typical population. In this case, the total length of text increases by a factor of 3.5 and the number of different lexical words only increases by a factor of 2 while the Uber value is only changed a small amount. While this would be a good indication of similarity in lexical diversity for an average respondent, we cannot definitively say our sample population is close to that average standard. This makes direct comparison of magnitudes between texts difficult.

5.4 Individual Case Studies

Individual analyses of the qualitatively derived and lexically quantified measures of conceptual understanding likely have differing utility in our particular dataset. However, it is not enough to judge overall efficacy of the approach entirely by comparing agreement between the measures. It is necessary to take a human centered approach and actually see how effective each measure is in illuminating the complexities of the interviews. The first case we will investigate is Phillip, where the results of change in conceptual understanding track well with the recorded change in Uber index. The following case of Roberta illustrates a case where these assessments disagree.

5.4.1 Case Study Phillip

Phillip's interview transcripts showed results that might have been predicted before completing this research. Conceptual understanding in both subject areas increased slightly over the three year span. Figure 11 details the number of codes generated in his response to the Pipe Network and L-Building problem over the course of the interview period. Phillip's highest conceptual proficiency was in the area of the L-building problem. The value of Uber index increased in both instances.

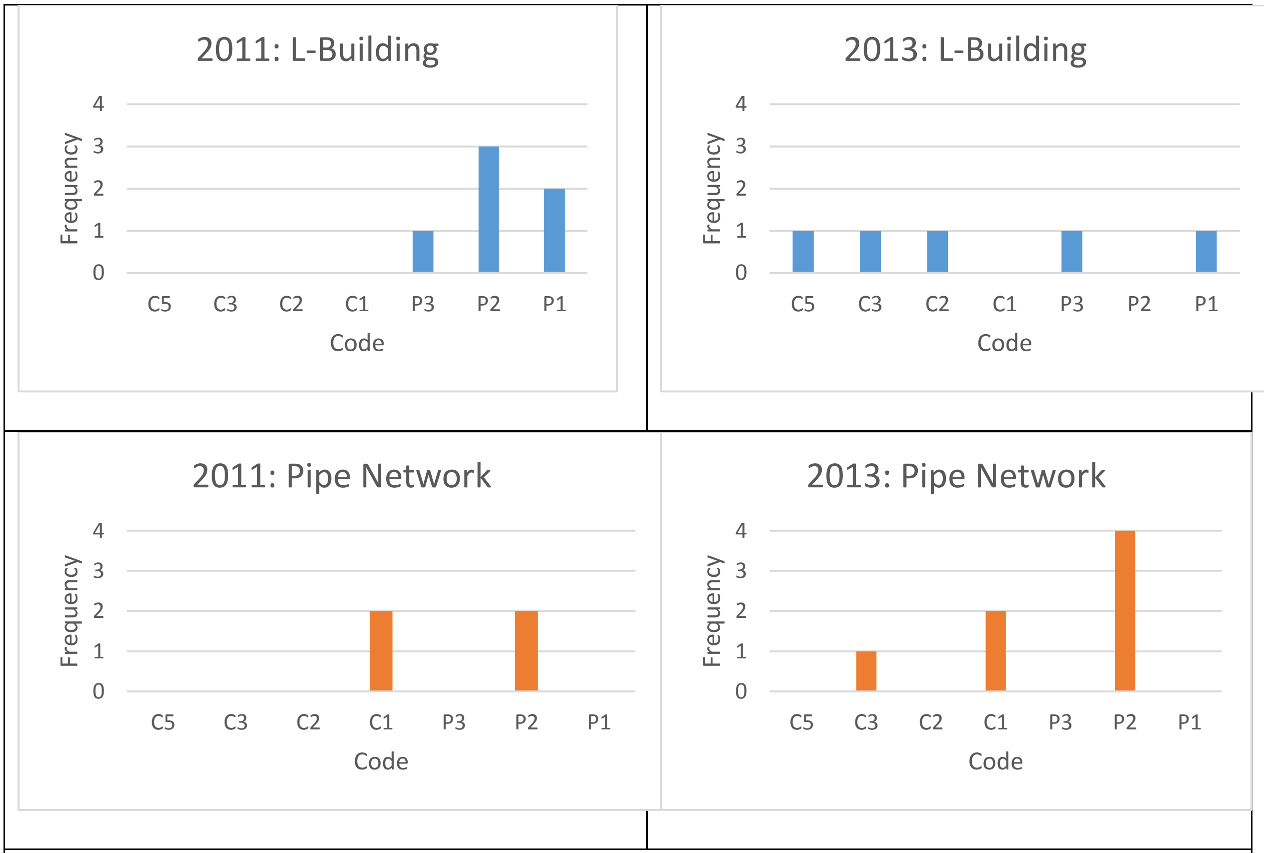


Figure 11: Comparison change in coding frequency over time for Phillip

The resulting codes from the 2011 interview showed a grouping of Procedural codes and a single low value Conceptual code. This would indicate a low conceptual understanding of the subject where solving techniques were limited to simple, surface level connections. By 2013, the resulting codes had changed to include more Conceptual codes of higher values in both cases. Even when given seemingly simple question prompts, Phillip was able to reflect on the context of answers and make connections to indirectly related concepts. His language complexity also increased, resulting in higher Uber values. This could be casually seen in his interview transcripts. Sentences became shorter and less ponderous. Various concepts were quickly raised and addressed. There was less verbal searching for an appropriate explanation.

The following interview excerpt demonstrates evidence of increasing levels of Phillip's conceptual understanding between the years 2011 and 2013. The interviewer is referencing the visual problem statement shown in Figure 7 that concerns wind loading on an L-shaped building. Phillip is asked if he can describe any general potential areas of concern when constructing such a building.

2011

Interviewer: Okay. Do you see any potential problems that could occur with the wind loading in the building?

Phillip: Well, yeah, especially if these are like open corridors or whatever in some area and some aren't, and just because they're open, you have no way to put shear walls and things like that, so you have nothing to take the load. Well, I guess more from the like the […] load and stuff, which the wind could affect, so it could affect their uplift, because it can get in the building. It can like push up. So cause, I guess, the diaphragms. That's technically what we call the uplift and twist and rip off and, plus building failure.

2013

Phillip: You don't need walls on a building, the only reason we have walls is if we like privacy. And elements, we like to be warm and, or cold in the summer. Really you just need the roof and the floor.

Interviewer: What could be done to counteract the problems?

Phillip: A lot of foundation haha. And, kind of design it in the way, like if the ocean is right here, which I know they designed it facing the ocean because of the views. That's

> not necessarily the best solution. You want to try to make it narrower, the narrow part where the wind is going to blow so it's easier to go around and not… nothing is built like that because that's not how architects want it.
>
> **Interviewer:** What about the window problem, is there anything you can do to counteract that?

In the example from 2011, Phillip demonstrates a step-by-step, or action sequence, solution strategy. Wind loading on a building will cause shear forces and shear forces are counteracted by the addition of shear walls. This would be indicative of a moderate to low level of conceptual understand where the analysis of a problem is simply solved by a standard if-then approach, which would not be out of place in any recent engineering graduate’s toolkit. His response in 2013, however, shows a relatively higher level of deep understanding of the problem domain. In this second excerpt, he tends to regard the walls as much smaller aspects of the previously seen problem, which now includes the motivation behind putting a building in a potentially high risk area in the first place. This underlying reason for constructing the building informs his acceptance that modifying the geometry to reduce wind loads might not be a feasible option, given that the requirements of the building likely include non-technical features of aesthetics. This takes into account requirements from the architects and the customers, as well as the engineer. The increase in generated Uber index looks to reflex this more nuanced problem solving strategy as well. This could potentially indicates a wider breadth of the subject matter he used to describe the problem.

Phillip did display high level Conceptual codes, which increased in both subjects. Interestingly, this did not reduce the number of Procedural codes found. It may be that Procedural answers were totally appropriate for procedural questions and did not indicate a low level of conceptual understanding. In this case, interpreting the Uber index might be valuable because it is theorized to increase with conceptual understanding.

5.4.2 Case Study Roberta

Contrary to the case of Phillip, the results from Roberta's analysis are slightly more contradictory. The collection of codes shown in Figure 12 relays that Roberta holds a much higher level of conceptual understanding in the L-Shaped Building problem than she holds in the Pipe network example.

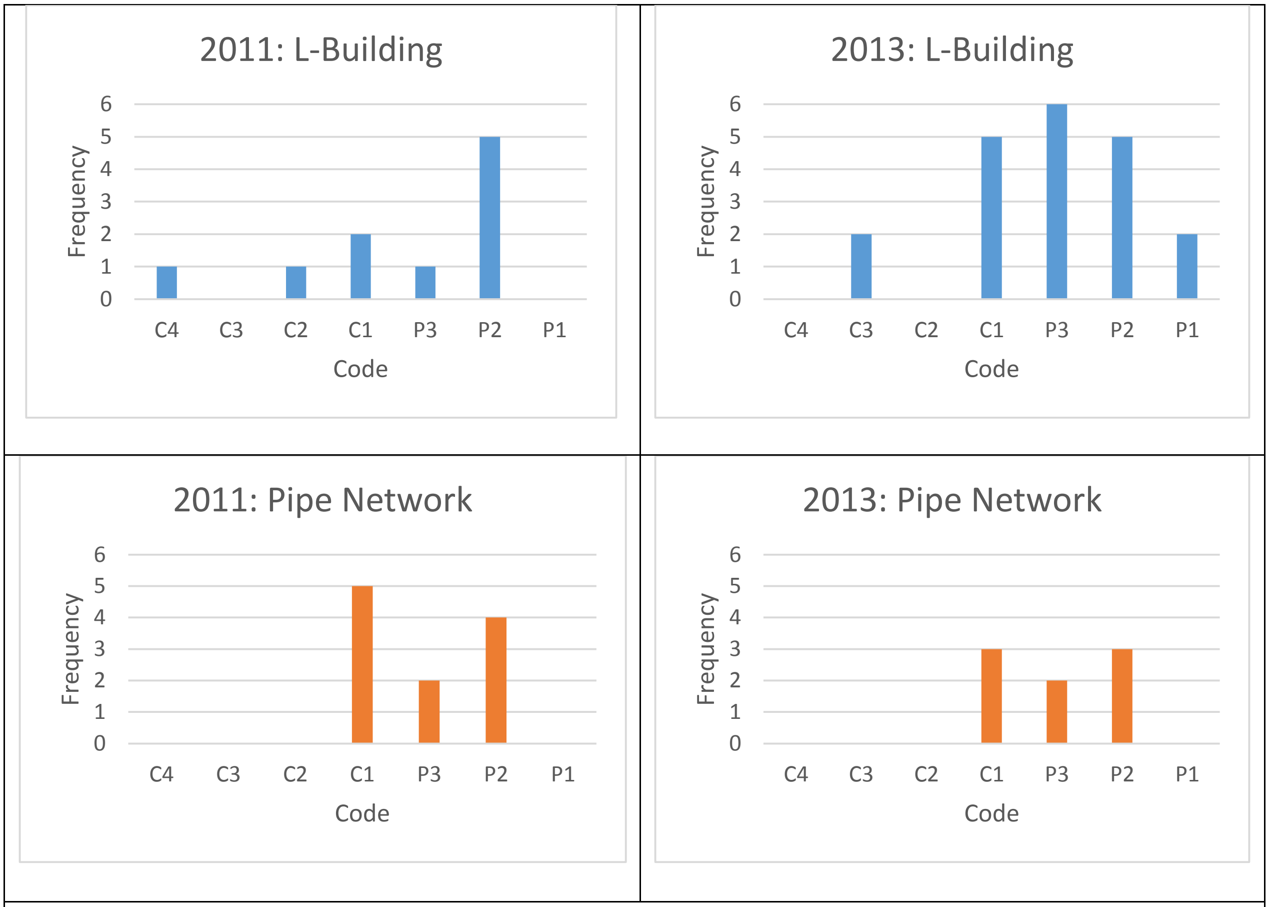


Figure 12: Comparison change in coding frequency over time for Roberta

A hallmark of the codes generated for Roberta was the relative lack of variability in one of the problem domains. The Pipe Network problem generated a similar number of codes with identical values over the three years. Showing the opposite effect, a moderate number of Conceptual codes (including a high valued one) were generated in the L-shaped building problem given in 2011 and seemingly degraded in value over time. Roberta's profession commonly involved analysis of buildings under construction, which led to the prediction that 2013 codes would be demonstrably higher. Instead, the overall values of the codes actually fell but the quantity of the Procedural ones increased significantly. When posed with a question in the L-building interview process, she was likely to give more rationales for answers but they tended to be less representative of strong conceptual understanding. It may be that the workplace environment she operated in supported simplistic concept connections as compared to the conceptual understanding developed in her engineering program.

When we compare this result to the change in Roberta’s Uber Index, we notice a discrepancy. While a researcher’s qualitative assessment of Roberta’s conceptual understanding did not appreciably change in the Pipe Network problem, the Uber index decreased, which indicates the opposite. A sample of the interview transcript is excerpted to explore this further. In this selection, the interviewer is probing the relationship she holds between volumetric flow rate, velocity, and pressure in the Pipe Network problem.

2011

Interviewer: Okay. How might you increase velocity?

Roberta: You could change the area.

Interviewer: Of the pipe?

Roberta: Yeah.

Interviewer: Okay. And which ones?

Roberta: And adding pressure could also kind of affect, I mean, your- no, ignore that. I would say initially my first guess would be area, area of the pipe.

Interviewer: Okay. And which, if you increase it, what happens to the velocity?

Roberta: If you increase the area of the pipe, it's going to decrease. That's what I would say.

Interviewer: Okay.

Roberta: There's a bigger area for the water to flow through and, if it's a small area, it's going to be more confined and it's going to want to flow faster, I think.

2013

Interviewer: Okay. Umm, okay. So we take away those ten houses. Now what if we change this six-inch to a twelve-inch pipe, how do you think that would change the system?

Roberta: And then what, what would these go to? [Referring to another pipe size within the problem]

Interviewer: They stay the same.

Roberta: They stay at two and four?

Interviewer: Mhmm.

Roberta: Oh, so it's just like a reducer. It would increase the pressure at these, both of them.

Interviewer: Okay. How come?

Roberta: Because now you have a bigger pipe here, so more water is going to be able to flow into here. So your flow rate here is going to be higher. Since your flow rate at this point has to equal the sum of those two, you know your flow rate there has to increase as well, in both of these lines. So, the increase in your flow rate, your pressure should increase too? I think? Isn't it? I don't remember that equation. Yeah, I don't remember but, that's what I would say.

In both years, we see that she is describing relationships between concepts that are definitely in the general problem domain but only a very fragile association exists between them. From this interaction we can infer that some level of procedural understanding exists; that when a concept such as flow rate is brought up, pressure or velocity likely has some role. However, not enough evidence exists to point to a moderate or high level of conceptual understanding. To the qualitative researcher, the responses in both years might be very similar. When the generated Uber Index is examined, we would expect to see constancy. However, the actual value decreased. It is possible that the varying length of the transcripts had an outsized effect on the Uber values. In general, Roberta's answers tended to be longer in 2013. The qualitative analysis suggests that similar levels of conceptual understanding exists in both years, which would be reflected in similar conceptual connections being demonstrated. The length of text may dominate the Uber analysis if the specific engineering vocabulary and justification don't appreciably change. In this case the qualitative assessment was deemed to be more reliable.

5.5 Comparison of Change in Conceptual Understanding to Change in Uber

We feel confident that noted change in conceptual understanding using qualitative analysis is accurate and reflective of the participants. This assessment was carried out by an experienced researcher and each code was given based on evidence found. This level of certainty is a mirror to the widespread adoption in the field. Determination of the Uber index was based on more assumptions that would be hard to assess without sifting through large amounts of procedurally generated data. This reason led to the impression that the Uber index was less reliable and informative. Various unaccounted for features such as total text length along with unverified

POS tags and lemmatization contributes to this. Despite this uncertainty, there seems to be a strong correlation between increases in conceptual understanding and increases in Uber index but Uber tended not to reflect the decreases in conceptual understanding when it was seen.

6 CONCLUSION

One of the most striking features of these results is the lack of agreement between these two methods, which occurs about half of the time. This indicates that we cannot immediately use values of Uber index as a surrogate for a qualitative analysis of conceptual understanding. There does seem to be a low level of experimental validation of literature linking lexical features and conceptual understanding, as would be expected based on the supporting literature. That low level of support is not enough to justify immediate application of this approach without further testing and refinement of the interview protocol and experimental procedure.

A positive outcome of this study was the development of a coding framework tailored to conversational-style responses of an engineering problem solving process. The framework provided an easily applicable structure that was able to classify most responses. The results of the individual assessments of participants' conceptual understanding appear to be repeatable and may be applied to different datasets. This not only allowed for a reliable benchmark to base this study on, but may serve as a validation for future refinements of the lexical analysis procedure.

Set the groundwork for further refinement of the study and validation of the process.

There is potential application for the approach of supplementing a qualitative coding with automatically generated Natural Language Processing in pre- and post-assessment across different subject areas.

6.1 Limitations of Current Approach

While individual comparisons of the Uber index to a qualitatively derived understanding of conceptual understanding shows merit, the lack of response variability in our sample is somewhat troubling. Uber index does seem to be a good indication of change in conceptual understanding but only when that change is large and the texts lengths are similar. When little conceptual change was seen, the resulting Uber values were not similarly correlated.

The experimentally determined lack of change in conceptual understanding was unanticipated and prevented a direct comparison to the Uber index in many cases. This resulted from the Uber index only being used to reflect a binary change in conceptual understanding. The confidence in determining lexical indicators for texts of widely varying length was not high enough to establish Uber as a high resolution instrument.

The qualitatively determined assessment of conceptual understanding was viewed as very strong indicator of change in conceptual understanding but some limitations of the interview procedure were not address when this study was implemented. Few precautions were in place to guard against the basic format of interview questions unduly influencing the response of interview participants. Short questions tended to get procedural answers. Without control of question type as an experimental parameter, there are few protections against the wide variety if interviewee responses.

6.2 Implications for Future Research

The surprisingly consistent Uber values derived in this study may have been accurate and reflective of the lexical diversity and conceptual understanding of the participants. This would be the case if the the participants in this study communicate incredibly similarly. The participants had graduated from the same engineering program at the same time, which they had all self-selected to participate in, indicating that to be a very real possibility. A broad base line of overall communication diversity needs to be established before the Uber findings can be strongly refuted or supported. This may be possible using a similar procedure and toolset but with differing metrics and literature support. Further experimentation should take care to select a more diverse participant pool.

The review of the techniques carried out in this research determined that this approach is likely to be more accurate if text length can be sufficiently managed and homogenized. This would be a strict control on a study designed to focus on largely spontaneous reasoning. Another approach would be to increase the dataset sufficiently such that differences in length can be statistically recognized and accounted for. However, this would make a complementary qualitative analysis very time consuming.

While the Uber index is well suited to short length responses, the style of the interview process might have been a hindrance. It is possible that a more robust representation of an engineer's conceptual understanding could be elicited through a different data collection method. Short written responses may serve to force the participants to better reflect on, and relay the understandings they hold. This may reduce the variability we found in Uber response due to text length in instances where demonstrated conceptual change is small.

BIBLIOGRAPHY

Barriball, K. L., & While, a. (1994). Collecting data using a semi-structured interview: a discussion paper. *Journal of Advanced Nursing*, *19*(2), 328–335. http://doi.org/10.1111/1365-2648.ep8535505

Bass, K. M., & Glaser, R. (2004). Developing Assessments to Inform Teaching and Learning CSE Report 628 Kristin M . Bass and Robert Glaser University of Pittsburgh May 2004 Center for the Study of Evaluation (CSE) National Center for Research on Evaluation , Standards , and Student Test, *1522*(310), 1–26.

Baxter, G. P., & Glaser, R. (1998). Investigating the Cognitive complexity of Science Assessments. *Educational Measurement: Issues and Practice*, (1), 37–45. http://doi.org/10.1111/j.1745-3992.1998.tb00627.x

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. *Text* (Vol. 43). http://doi.org/10.1097/00004770-200204000-00018

Borrego, M., Froyd, J. E., & Hall, T. S. (2010). Diffusion of engineering education innovations: a survey of awareness and adoption rates in U.S. engineering departments. *Journal of Engineering Education*, *99*(3), 185–207. http://doi.org/10.1002/j.2168-9830.2010.tb01056.x

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

Carcayry, M. (2011). Evidence analysis using CAQDAS: Insights from a qualitative researcher. *Electronic Journal of Business Research Methods, 9*(1), 10–24.

Case, J., & Light, G. (2011). Emerging research methodologies in engineering education research. *Journal of Engineering Education, 100*(1), 186–210. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2168-9830.2011.tb00008.x/abstract

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

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.

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

Chi, M. T. H., & Roscoe, R. D. (2002). The processes and challenges of conceptual change. In *Reconsidering Conceptual Change. Issues in Theory and Practice* (pp. 3–27). http://doi.org/10.1007/0-306-47637-1_1

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

Edward, F., Richard, N., Redish, E. F., & Steinberg, R. N. (1999). Teaching Physics : Figuring Out What Works Figuring out what works : Discipline- based education research, (2), 1–14.

Evans, D., Gray, G., & Krause, S. (2003). Progress on concept inventory assessment tools. *33rd Annual Frontiers in Education*, 2–9. http://doi.org/10.1109/FIE.2003.1263392

Galletta, A., & Cross, W. E. (2013). *Mastering the Semi-Structured Interview and Beyond: From Research Design to Analysis and Publication.* NYU Press.

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

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

Henderson, C., & Dancy, M. (2011). Increasing the impact and diffusion of STEM education innovations. *... Education Innovations, Available ...*. Retrieved from http://homepages.wmich.edu/~chenders/Publications/Henderson2011Diffusion of Engineering Education Inovations.pdf

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

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

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

Hruschka, D. J., Schwartz, D., St.John, D. C., Picone-Decaro, E., Jenkins, R. a., & Carey, J. W. (2004). Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research. *Field Methods*, *16*(3), 307–331. http://doi.org/10.1177/1525822X04266540

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

John-Steiner, V., & Mahn, H. (1996). Sociocultural Approaches to Learning and Development: A Vygotskyian Framework. *Educational Psychologist*. http://doi.org/10.1207/s15326985ep3103&4_4

Jones, M. L. (2007). Using software to analyse qualitative data. *Malaysian Journal of Qualitative Research*, *1*(1), 64–76. Retrieved from papers2://publication/uuid/F63D1D09-4501-4466-8AA0-0B9F70772ED9

Joshi, A. K. (2003). Natural Language Processing. *Annual Review of Information Science and*

Technology, *37*(1), 51–89. http://doi.org/10.1017/S0267190500001446

Kalyuga, S. (2009). Instructional designs for the development of transferable knowledge and skills: A cognitive load perspective. *Computers in Human Behavior*, *25*(2), 332–338. http://doi.org/10.1016/j.chb.2008.12.019

Lewins, A., & Silver, C. (2009). Choosing a CAQDAS Package What ranges and types of software support work with qualitative data? Choosing a CAQDAS Package Which software packages do we categorise as CAQDAS?, (April).

Litzinger, T., Lattuca, L., Hadgraft, R., & Newstetter, W. (2011). Engineering education and the development of expertise. *Journal of Engineering Education*, *100*(1), 123–150. http://doi.org/10.1002/j.2168-9830.2011.tb00006.x

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

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

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

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

Newcomer, J. L., & Steif, P. S. (2008). Student thinking about static equilibrium: Insights from written explanations to a concept question. *Journal of Engineering Education, 97*(4), 481–490. Retrieved from http://www.jee.org/2008/october/8.pdf

Norton, L. (2009). *Assessing student learning*. Bolton, MA: Anker Publishing Company, Inc.

Piaget, J. (1997). Development and learning. *Readings on the Development of Children*. http://doi.org/10.1080/14767333.2011.617145

Rittle-johnson, B., & Schneider, M. (2015). *Developing Conceptual and Procedural Knowledge of Mathematics*. *Oxford handbook of numerical cognition*. Oxford, UK: Oxford University Press. http://doi.org/0.1093/oxfordhb/9780199642342.013.014

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*. http://doi.org/10.1037/0022-0663.93.2.346

Sager, N. (1981). *Natural language information processing : a computer grammar of English and its applications*. Reading, Massachusetts: Addison-Wesely.

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

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

Settanni, M., & Marengo, D. (2015). Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. *Frontiers in Psychology*, *6*(July), 1–7. http://doi.org/10.3389/fpsyg.2015.01045

Šišková, Z. (2012). Lexical Richness in EFL Students ’ Narratives. *University of Reading Language Studies Working Papers, 4,* 26–36. Retrieved from http://www.readingconnect.net/web/FILES/english-language-and-literature/elal_LSWP_Vol_4_Siskova.pdf

Smith, J. I., & Tanner, K. (2010). The Problem of Revealing How Students Think: Concept Inventories and Beyond. *CBE-Life Sciences Education*, *9*, 10–16. http://doi.org/10.1187/cbe.09

Streveler, R., Litzinger, T., Miller, R. L., & Steif, P. S. (2008). Learning conceptual knowledge in the engineering sciences: Overview and future research directions. *Journal of Engineering Education*, *97*(3), 279–294. http://doi.org/10.1002/j.2168-9830.2008.tb00979.x/

Svinicki, M. D. (2010). A Guidebook On Conceptual Frameworks For Research In Engineering Education, 53.

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

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

Van Merrienboer, J. J. G., & Sweller, J. (2005). *Cognitive load theory and complex learning: Recent developments and future directions*. *Educational Psychology Review* (Vol. 17). http://doi.org/10.1007/s10648-005-3951-0

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.