Semantic Features of Math Problems: Relationships to Student Learning and Engagement

Stefan Slater
Ryan Baker
Jaclyn Ocumpaugh
Teachers College Columbia University
525 W. 120th St
New York, NY 10027
{slater.research,
ryanshaunbaker,
jlocumpaugh}@gmail.com

Paul Inventado
Peter Scupelli
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
{pinventado,
scupelli}@cmu.edu

Neil Heffernan
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609
nth@wpi.edu

ABSTRACT

The creation of crowd-sourced content in learning systems is a powerful method for adapting learning systems to the needs of a range of teachers in a range of domains, but the quality of this content can vary. This study explores linguistic differences in teacher-created problem content in ASSISTments using a combination of discovery with models and correlation mining. Specifically, we find correlations between semantic features of mathematics problems and indicators of learning and engagement, suggesting promising areas for future work on problem design. We also discuss limitations of semantic tagging tools within mathematics domains and ways of addressing these limitations.

Keywords

Text mining, semantic analysis, problem features, engagement, learning, correlation mining, mathematics corpora

1. INTRODUCTION

As content is developed at scale for online learning systems, particularly systems that leverage content developed by large numbers of authors, it becomes important to distinguish between problems which are well-written and conducive to learning and those which are poorly worded or otherwise difficult to understand. Crowd-sourced content, where content is authored by a broader community [21], is a powerful and scalable method of content creation, which can be used to quickly develop and deploy new content and curricula ([46], [17]).

For this reason, it is critical that an equally scalable method of analyzing problem quality be developed, to prevent learning platforms that leverage crowd-sourced content from becoming dominated by ineffective content. In other platforms such as Wikipedia the quality of crowd-sourced materials is improved through substantial coordination between contributors [20]. However, there is relatively little work evaluating crowd-sourced learning content at scale. In contrast with more traditional educational measurement (from tests), where determining items' ability to discriminate student knowledge is a standard part of item analysis [11], there has been less attention to this problem for online learning systems. While some researchers have attempted to determine which hints are more effective [18], or which problems are associated with more learning [14], these efforts have focused on what, but not why, particular system features can impact students, limiting their degree of general use. A more theoretical approach was taken by [49] where a design space of over 70 features characterizing Cognitive Tutor lessons was distilled and correlated with an automated gaming the system detector. However, this work identified the characteristics of tutor lessons using hand-coding, a method that is infeasible for larger datasets, and was limited to the relatively narrow space of problems designed by professional educational developers.

An alternative method for the analysis of the design of content in large-scale educational systems is text mining. There is a considerable amount of small-scale research on linguistic features that impact reading in mathematical contexts [47], but as [16] point out, many of the traditional readability indices used to study language at scale are limited in the features they consider. As a result, many early studies did not find a relationship between readability and performance in mathematics word problems [48].

As more advanced linguistic tools have become available, large-scale investigations of mathematics language have become more fruitful. For example, [44] have used LIWC [37] and CohMetrix [15] to study the effects of linguistic properties of mathematics problems ([44], [45]). [45] found that third-person singular pronouns (e.g., he, she) are significantly associated with correct answers and fewer hint requests in Cognitive Tutor problems. They found positive correlations between the use of work-related terms and learning, and negative correlations between the use of terms related to social constructs and learning. These findings highlight the potential value of linguistic features for better understanding learning, as well as the need to explore a wider range of semantic categories in a broader range of mathematics content areas.

In this paper, we use a discovery with models approach, generating prediction labels from automated detectors of student learning and engagement that were developed for the ASSISTments online learning system ([2], [32]). We build on [46]'s approach of using text mining software and text elements, such as HTML tags and Unicode characters, to distill features from a corpus of mathematics problems. We then use correlation

EDM '16, June 29–July 2, 2016, Raleigh, NC, USA.

mining approaches to identify links between these features and our labels of student engagement and learning as a means for determining which combinations of linguistic features are associated with particularly effective problems.

1.1 ASSISTments

The current study uses data collected from the ASSISTments system. ASSISTments is an online intelligent tutoring system used by over 50,000 students annually for middle-school mathematics. It provides both formative and summative *assessment* as well as extensive student support (*assistment*) and detailed teacher reports. It also facilitates research using randomized controlled trials (RCTs) that allow researchers to conduct studies without interfering with instructional time [17].

Within the system, students are assigned problem sets that may vary on several dimensions. Problem sets can be differentiated in terms of how problems are assigned: (a) In *Complete All* problem sets, problem order may be randomized; students must correctly answer all of the questions assigned and cannot advance to the next problem unless they have answered correctly. (b) In *If-Then-Else* problem sets, students must correctly answer a specified percentage of questions correctly (default is 50%) in order to pass, or *else* they may be given additional problems. (c) Finally, in *Skill Builder* problem sets, students must get 3 consecutive correct answers in order to pass, thus allowing students who show mastery to move on quickly to new assignments while providing struggling students with extended practice.

The purpose of the current study is to evaluate the semantic properties and HTML metadata (which may carry semantic meaning) of problems authored in ASSISTments. Many have been vetted by the ASSISTments expert team, but others (76% as of 2014) were created by teachers themselves [17]. ASSISTments provides scripted templates, which allow teachers to customize problem sets for specific topics. Therefore, finding ways to identify meaningful differences in teachers' problem design is an important area of research.

2. DATA & METHODS

In this paper, we analyze 179,908 problems within the ASSISTments system, most developed by teachers. We study these problems using the features of the problems themselves, in combination with data from the log files of 22,225 students who used ASSISTments during the 2012-13 school year. We applied models from previous research on engagement and learning to these students' log files in order to determine how these constructs are associated with features of the design of the problems, developed through linguistic analysis and other data about the problems. In doing this, we excluded from consideration features that had been previously used within the learning and engagement models described below, to prevent overfitting.

2.1 Learning & Engagement Measures

Learning and engagement were assessed automatically, using detectors or models of these constructs.

2.1.1 Student Learning

Student learning was assessed by fitting the moment-by-moment learning model to the data [2]. The moment-by-moment learning model (MBMLM) attempts to infer the specific effect of each learning opportunity on a student's overall mastery. We used [2]'s look-ahead-two probabilistic approach, which assumes that learning can occur at multiple points along a student's trajectory of learning a skill, rather than [43]'s approach which assumes a single moment of learning. We also choose this formulation because it explicitly analyzes future performance, allowing us to focus on cases where students perform better than expected after encountering a particular problem. Using the MBMLM allows us to isolate the average learning associated with specific problems within the data and compare these averages to other problems that either lack or have particular features of interest.

2.1.2 Automated Detectors of Engagement

Detectors of student engagement were developed using data from *in situ* classroom observations, conducted by experts certified in the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP 2.0). The protocol is enforced by HART, an Android application designed specifically for the BROMP and freely available for non-commercial research [33], which enforces the protocol while facilitating data collection.

Upon completion of the observations, data mining techniques were then employed to provide models of each construct that were cross-validated at the student level. In this paper, affective models developed for three different populations of students were applied, matching urban, suburban, and rural models to student data based on the location of their schools, in order to ensure population validity [32]. A detailed description of the features and algorithms used in these detectors is given in [32] and [34].

2.1.3 Applying Across-Student Measures of Learning & Engagement to Individual Problems

In this paper, both the MBML model and the engagement models were used as indicators of problem effectiveness. This section describes how these models were aggregated across the 179,908 problems and 22,225 students in this study. The formulation of the MBMLM in [2] is calculated once for each problem, at the time of the first attempt, and there is only one estimate per problem. Therefore, MBML was estimated for each student based on the sequence in which the problem was seen. Problem-level measures were then produced by averaging the MBML values across all students who saw a given problem.

The affective models were applied by segmenting the data at 20-second intervals (matching the original approach used to develop the detectors), and then applying each model to each segment. Confidence values for each detector was averaged twice at the problem level: first for each student (in order to avoid biasing the estimates in favor of the affect experienced by students who spent longer working the problem), then across all students who had seen that problem. This resulted in five measures per problem (average boredom, confusion, engaged concentration, frustration, and gaming), which we used, along with MBMLM outcomes, as our dependent variables.

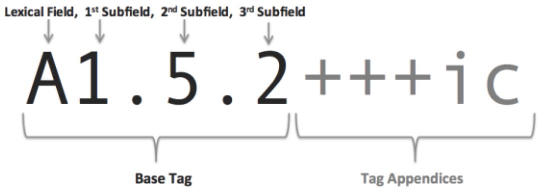
2.2 Feature Engineering

A number of different design features may influence student learning and engagement. In this paper, we explore features of both the problem text and its meta-text. Specifically, we look at word counts, lexical category features generated by a semantic tagger, and features generated from the metadata connected to the problem, which provides us with a separate source of semantic data (e.g., the use of mathematical notation which would not be captured by a semantic tagger) as well as with information about its use of tables, images, formatting, bolded or emphasized text.

2.2.1 Wmatrix Semantic Tags

The semantic content of ASSISTments problems was analyzed with Wmatrix [39], a corpus analysis and comparison tool that parses text at a word and multi-word level. As of 2004, this included 42,300 single word entries and over 18,400 multi-word expressions [38]. Wmatrix has been used in a number of analyses, including work to tag and identify lexical patterns in ontology learning [13] and work to study how students self-explain when learning science content [12]. Its semantic tagger uses a semi-hierarchical structure where all known words and multi-word units are classified into one of 21 lexical fields, represented with letters by its tagging system. These lexical fields may (or may not) be further subdivided in up to three different levels, which are represented in what we will refer to as the base tag.

Figure 1. WMatrix tagging system.



Within the lexical tag, we will refer to the lexical field (alphabetical) and the 1st, 2nd, and 3rd order subfields (numeric) as the *base tag*. Additional information about antonyms (*black* vs. *white*), comparatives (*better, worse, more confusing*, etc.), superlatives (*best, worst, most confusing*, etc.), gender (*masculine, feminine*, and *neuter*), and anaphoric status (i.e., contextual reference), may or may not be *appended* to a base tag. Wmatrix documents 234 distinct base tags, and represents a large number of additional possible labels through appendices

In the ASSISTments data, 442 distinct Wmatrix tags (base + appendices) were identified. These tags were most likely to fall under 7 lexical fields: General & Abstract Terms (A), Numbers & Measurement (N), Social Actions, States, & Processes (S), Psychological Actions, States, & Processes (X), Names & Grammatical Words (Z), Money & Commerce in Industry (I), and Time (T).

2.2.2 Accommodating Known Wmatrix Limitations

Although Wmatrix has been evaluated for its effectiveness in a range of genres, domains, and historical periods [38], semantic taggers can have a number of limitations when applied to highly specialized domains ([28], [24]; [36]; [30]; [27]). For example, research has shown that words which contain more than one unit of meaning create challenges for taggers that apply only one label per word [41]. As a result, semantic taggers which work specifically with scientific language have become an area of research interest ([1], [10]), but the language of mathematics has not yet been well-developed.

As such, features generated by Wmatrix must be carefully checked within this data set and may need to be supplemented by domain-specific tags. For example, we found several Wmatrix tags that erroneously tagged high-frequency items that appeared in ASSISTment's instructions to students, including problems that instructed students to *enter* fractions in a specific format in order to receive credit or which told students that they had 3 attempts *left.* Wmatrix treated many of these words (e.g., *enter* and *left)* as an indication of physical movement (M1, as in *entering a building* or *turning left*). A few erroneous tags also appeared to result from the development of Wmatrix as a tool for British English. For instance, ASSISTments users, who are primarily American English speakers, wrote a number of problems involving a person named *Randy*, whose name was automatically (and erroneously) tagged as involving sexual content.

To mitigate this issue, significant correlations were carefully inspected individually. This approach has been found to be useful in previous studies where semantic taggers were applied to new domains [12]. While the large size of the ASSISTments corpus limits our ability to address this problem completely, thorough efforts were made to examine and understand relationships discovered through the use of Wmatrix. In instances where Wmatrix applied a tag involving the wrong sense of a word for the context in which it was used, we have specifically noted this difference and what sense of a word or words the tag is capturing within ASSISTments.

2.2.3 Math Symbols and Other Textual Metadata

In addition to generating features with Wmatrix, we also generated features based on the metadata of each problem. We were primarily concerned with identifying Unicode characters that are semantically meaningful in mathematics contexts. In the ASSISTments corpus, we labeled 68 symbols, such as those for integrals, mean, standard deviation, and exponents. These domain-specific symbols present unique challenges to the teaching and learning of mathematics [40], but are not detected by most lexical analysis tools, which have not generally been developed for mathematics domains. In addition, we identified 14 HTML tags that were used to format ASSISTments problems, including tags used for boldface, italics, paragraph structure, and images. Because many of these functions can also alter the semantics of a problem, we also generated features that reflect these uses of HTML in problem metadata. These features were generated by counting the number of times that each HTML code was used in a problem, in parallel to the application of the Wmatrix tags discussed in previous sections.

3. RESULTS

To explore the relationship between these problem features and the BROMP-trained measures of engagement and learning, we correlated each problem feature to each predicted variable. We selected Spearman's ρ as our correlation coefficient because of its increased robustness when correlating non-normal data as compared to other parametric coefficients such as Pearson's R [50]. Additionally, with such a high number of comparisons being conducted it was necessary to adjust our significance criterion to account for the possibility of tests being incorrectly identified as significant. The Benjamini and Hochberg post-hoc procedure [4] was used to control for these false discoveries. A table of results by dependent variable is presented in Table 1, which also provides the average confidence level for each detector as a baseline measure for this data.

Table 1. N of significant features by outcome measure.

Outcome Measure	Avg Conf.	Total Sig	Sig w/ |ρ| > 0.05	Sig w/ |ρ| > 0.10
Bored	0.16	118	16	0
Engaged Concentration	0.46	251	62	14
Confusion	0.03	285	60	5
Frustration	0.04	216	36	7
Gaming the System	0.02	257	43	5

Of the possible 2730 correlations, 1127 (41.3%) were statistically significant after controlling for multiple comparisons using Benjamini & Hochberg's post-hoc control. More features were significantly correlated with confusion than any other outcome measure, but large numbers of features were also correlated with gaming the system, engaged concentration, frustration and MBML. Boredom was correlated with fewer features, overall, than either of the other outcome measures. These broad findings suggest the potential for finding semantic features that may help to provide templates for improving the design of word problems.

3.1 *Features associated with all outcome measures*

In the following sections, we examine the relationships between our features and the individual outcome measures, but in order to provide a broad summary of which types of features had the largest effects, the absolute value of Spearman ρ was averaged across all six outcome measures for each feature in this study. Among the 64 features that were signifcantly correlated with all six outcomes, the 10 with the highest ρ average (shown in Table 2) were drawn from 5 lexical fields: Grammatical Bin (Z), General Terms (A), Time (T), Speech Acts (Q), and Numbers & Measurement (N). One HTML tag (<p>, paragraph) was also significant.

Table 2. 10 largest correlated features by average sig. |ρ|

Tag	Avg |ρ|	MBM Learning	Boredom	Concentra-tion	Confusion	Frustration	Gaming
Z5	0.116	0.193	0.086	-0.165	0.084	0.105	0.060
Z5mwu	0.104	0.114	0.034	-0.040	0.135	0.162	0.140
A12-	0.101	0.114	-0.027	0.030	0.086	0.153	0.198
T3-	0.091	0.084	-0.034	0.055	0.074	0.144	0.153
Q2.2	0.080	0.043	0.083	-0.162	0.068	0.071	0.051
T1.1.2	0.076	0.076	-0.051	0.031	0.067	0.116	0.116
<p>	0.071	0.149	0.054	-0.127	0.015	0.064	-0.015
N1	0.069	0.061	0.076	-0.077	0.082	0.080	0.035
A5.4+	0.066	-0.028	0.059	-0.130	0.074	0.038	-0.069
Z6	0.056	0.108	0.020	-0.034	-0.077	-0.032	0.071

Spearman's ρ is also shown for individual outcome measures, allowing us to examine the effects of these features in greater detail. Table 2 shows that WMatrix's Speech Acts tag (Q2.2, e.g., *answer*, *account*, or *speak out*) is correlated with small increases in learning, but is also positively correlated with increased boredom and gaming and decreased concentration. The Wmatrix features described as *Grammatical Bin* (words such as *as, but, in order to*) are also correlated with increased learning, boredom, and gaming. Correspondingly, they are also negatively associated with engaged concentration, illustrating the complicated interactions at play in this data and the importance of considering multiple outcomes when exploring design effects.

4. Results by Outcome Measure

While some interactions are complicated, we also see many features correlate in logical patterns. For example, features that are positively associated with boredom are often also negatively associated with engaged concentration, and vice-versa. Likewise, features associated with confusion are also associated with frustration. The remainder of this section discusses these patterns in greater detail, pairing outcome measures that are conceptually related (e.g., boredom and engaged concentration as well as MBML and gaming the system, which have shown to be inversely related in the past). Specifically, we will examine the ten features that are most negatively associated and the ten that are most positively associated with each outcome measure, discussing commonalities across outcome measures.

4.1.1 *Learning & Gaming the System*

The Spearman ρ values for the top ten features range from -0.078 to 0.233 for MBML and from -.095 to 0.198 for gaming the system. Table 3 presents these results, highlighting features that correlate with both outcome measures.

Table 3. Features most strongly associated with MBML and gaming the system

LEARNING			GAMING		
TAG	**Semantic Description**	**ρ**	**TAG**	**Semantic Description**	**ρ**
A5.2+	True/False	-0.078	N5+	Quantities	-0.095
S9	Religion & the supernatural	-0.075	A10+	Open/Closed; Hiding/Hidden; Findin	-0.092
A11.1+++	Important/Significant	-0.066	X2.1	Thought/belief	-0.084
A6.1+	Similar/Different	-0.062	A2.1+	Modify, Change	-0.082
G2.2+	General Ethics	-0.059	S5+	Groups and affiliation	-0.074
N3.2+++	Measurement: Size	-0.059	N5.2+	Exceeding; waste	-0.070
A3-	Being	-0.058	A5.4+	Authenticity	-0.069
Z8mwu	Pronouns etc.	-0.054	T1	TIME GENERAL	-0.069
N1mwu	Numbers	-0.051	N5	Quantities	-0.067
X5.2+	Interest/boredom/excited/ennergetic	-0.049	T2+	Time: Beginning and ending	-0.067
A12-	Easy/Difficult	0.114	A7+mwu	Definite (+modals)	0.086
Z5mwu	Grammatical bin	0.114	X2.4mwu	Investigate/examine/test/search	0.087
Z99	Unmatched	0.114	N3.8+	Measurement: Speed	0.093
N3.3---	Measurement: Distance	0.115	Z8	Pronouns etc.	0.093
X2.2+	Knowledge	0.121	A12+++	Easy/Difficult	0.098
M7	Places (geographical & conceptual)	0.130	T1.1.2	Time: General: Present; Simultaneous	0.116
N3.8+	Measurement: Speed	0.142	X8+	Trying	0.140
<p>	HTML paragraph	0.149	Z5mwu	Grammatical bin	0.140
Z5	Grammatical bin	0.193	T3-	Time: Old, new and young; age	0.153
M1	Moving, coming, & going	0.223	A12-	Easy/Difficult	0.198

Although gaming is an infrequent behavior, previous research has shown that it is linked to poorer learning ([7], [34]). Therefore the findings in Table 3 are somewhat surprising. We should expect gaming's infrequency to limit overlap between the two categories, and expect them to show inverse relationships when present. Instead, A12- (words related to *difficulty*), Z5mwu (multiword grammatical units like *as far as* or *for example*), and N3.8+ (words related to *higher speeds*), are all associated with increased MBML **and** increased gaming behaviors. Likewise, semantically similar categories like N1mwu (multiword *numbers*) and N5+ (*large quantities*) are associated with lowered MBML **and** lowered rates of gaming behaviors.

These anomalies might be due to the existence of problems that support learning but can be gamed relatively easily, or might suggest that particularly challenging problems lead to learning but also inspire gaming behavior. For example, A5.2+ (words associated with *true*) demonstrates the lowest correlation with learning, a result that is consistent with literature on the ineffectiveness of true/false questions [42]. Likewise Z8mwu (multiword pronouns, e.g., *anything at all*) is correlated with lower MBML, while Z8 (single word pronouns, e.g., *it, my,* and *you*) is correlated with increased gaming. These findings align with research showing that pronouns can be difficult to process cognitively (taxing working memory), as they require readers to infer their antecedents (the words that give them their meaning) from context ([25], [8], [22], [6]). This suggests that pronouns could inhibit learning by drawing mental resources away from mathematics task, perhaps inspiring some students to try to succeed with minimal cognitive effort.

These findings highlight important considerations for researchers working to improve learning systems, including the need to consider multiple measures. For example, [44] found that pronouns are associated with correct answers and lowered hint use. It is highly likely that pronouns can have beneficial impacts on learning, particularly through [44]'s hypothesized mechanism of increased cohesiveness. However, if pronoun use in ASSISTments and Cognitive Tutor is comparable, our results suggest that some correct answers could have been achieved by guessing rather than by learning.

Furthermore, if students are more tempted to game the system when presented with challenging problems, even though these are exactly the sort of problems needed to improve learning, then further research should explore whether or not these findings reflect two distinct different groups of students. It may be that some students need additional cognitive scaffolding or a motivational intervention in order to complete these problems without gaming, allowing them to learn as well as other students who are working through the curriculum in a more appropriate way. However, research has also shown that in some cases high achieving students also game the system, and the independent application of these models could be picking up on that trend, where students guess something that they actually know, but then correct this behavior in subsequent problems, which could cause the MBML model to perceive learning.

4.1.2 Confusion & Frustration

Confusion and frustration show considerable overlap, in line with prior theory on the relationship between these constructs ([9], [26]). As Table 4 shows, half (10) of the semantic features most strongly associated with one are also strongly associated with the other, including N6mwu (*frequency of occurrence*) which is negatively associated with both confusion and frustration. This corresponds with [44]'s findings that clear demarcations of time in mathematics problems can improve student outcomes.

Table 4. Features most strongly associated with confusion and frustration

CONFUSION			FRUSTRATION		
TAG	**Semantic Description**	**ρ**	**TAG**	**Semantic Description**	**ρ**
X2.1	Thought/belief	-0.149	X2.1	Thought/belief	-0.110
Z6	Negative	-0.101	N5+	Quantities	-0.070
N3.4	Measurement: Volume	-0.097	A11.1+++	Important/Significant	-0.063
N3.3---	Measurement: Distance	-0.079	N3.4	Measurement: Volume	-0.061
N6mwu	Frequency of occurance	-0.079	A2.2	Cause, Connected	-0.056
A2.2	Cause, Connected	-0.077	N6mwu	Frequency of occurance	-0.052
A1.5.1	Using	-0.076	X4.2	Means, method	-0.051
N5+	Quantities	-0.070	T2++	Time: Beginning and ending	-0.050
I1.3	Money: price	-0.068	A2.1+mwu	Modify, Change	-0.049
O4.1	General Appearance/Phys'l Propert	-0.066	<font>	HTML font adjustment	-0.049
Q1.2mwu	Paper documents & writing	0.081	I3.1	Work & Employment: generally	0.089
N1	Numbers	0.082	X2.4mwu	Investigate/examine/test/search	0.092
I3.2	Work & Employment: professional	0.083	<span>	HTML span (grouping of items in or	0.092
Z5	Grammatical bin	0.084	N6+	Frequency of occurance	0.093
A12-	Easy/Difficult	0.086	Z5	Grammatical bin	0.105
<em>	HTML italics	0.087	T1.1.2	Time: General: Present; simultaneous	0.116
I3.1	Work & Employment: generally	0.094	T3-	Time: Old, new and young; age	0.144
S6+	Obligation and necessity	0.105	X8+	Trying	0.148
X8+	Trying	0.115	A12-	Easy/Difficult	0.153
Z5mwu	Grammatical bin	0.135	Z5mwu	Grammatical bin	0.162

Notable semantic features within this pairing include Z5 and Z5mwu. Both capture what are known as grammatical bin, which includes prepositions (*of, to, after, amid*), conjunctions (*and, or, but*), certain adverbs (e.g., *as, so, which, than, when*), the infinitival maker (*to* + verb), determiners (e.g., *a* and *the*) and certain auxiliary verbs (e.g., *do*). Previous research has suggested that the highly specific style of scientific language increases the use of these parts of speech, especially in the sort of definitional contexts that we might find in many learning contexts [3]. [29], for example, notes that students sometimes struggle with prepositions. In fact, this pattern is sometimes referred to as the *stylistic barrier hypothesis* [31], which suggests that differences between the language students use at home and the language used in the classroom may interfere with the learning process.

HTML features that that correlate with confusion and frustration match findings in the literature. For example, [35] suggest that italics are difficult to read, and our findings show that they are correlated with higher confusion. Changes in font size, however, are associated with lower frustration; it is possible that teachers are using changes in font size to clarify visual hierarchy and problem meaning.

Features associated with concreteness (N3.4, N3.3, A2.2, A1.5.1, N5+, I1.3, O4.1, T2++) correlate with lowered confusion and frustration, matching the literature on the *concreteness effect*, which shows that concrete words are not only processed faster than abstract words in many experimentally controlled studies [23], the two may operate in separate neurological pathways ([19], [5]). These findings are hypothesized to be an artifact of the word-to-word mapping system the brain uses to process language, where concrete words may have stronger ties to more basic concepts. Interestingly, [23] have found evidence for similar pathways for emotion words, which are acquired early and considered quite basic to the human experience. While several of the Wmatrix categories that might correspond with [23]'s account of emotion words do not appear in this list (E3, E4, X4.1), X2.1, described as *thoughts/beliefs*, has the strongest negative associations with both frustration and confusion.

Other features which correlate with increased confusion and frustration may reflect the sort of meta-instructions teachers use to support students working with complex mathematical problems. Consider, for example, the tags in the following examples:

(1) *You*_Z8mf *must*_**S6+** *show*_A10+ *your*_Z8 *work*_**I3.1**.
(2) *You*_Z8mf *have*_A9+ *three*_N1 *attempts*_**X8+**
(3) *Often*_**N6+** *it*_Z8 *helps*_S8+ *to*_**Z5** *write*_Q1.2[i1.2.1 *down*_Q1.2 [i1.2.2 *your*_Z8 *work*_**I3.1**.
(4) *Keep*_A9+ *trying*_**X8+**
(5) *Do*_**X8+**[i1.3.1 *your*_**X8+**[i1.3.2 *best*_**X8+**[i1.3.3
(6) *Do*_A1.1.1 *the*_**Z5** *difficult*_**A12-** *problems*_**A12-** *first*_N4

Several of these tags (as given in bold, above: I3.1 *work;* S6+ *must;* Z5 *to, the;* X8+ *attempts, trying*; A12- *difficult;* N6+ *often*) are correlated with increased confusion or frustration. This finding may reflect a preemptive scaffolding practice (e.g., teachers provide these additional instructions when students are working on problem types that they have struggled with in the past). However, it is important to rule out other possibilities. For instance, such additional instructions could distract or annoy the students. More seriously, it could also have priming effects.

4.1.3 Engaged Concentration & Boredom

Like confusion and frustration, we see considerable overlap in the features correlated with engaged concentration and boredom. However, unlike confusion and frustration, these two outcome measures are negatively associated with one another. Six of the features most negatively associated with concentration (N5-, N3.6, Z5, Q2.2, A4.1, and A5.4+) are among those most positively associated with boredom. Likewise, four of those most positively associated with concentration (A2.1+mwu, A6.1+++, T3, and A5.2+) are negatively associated with boredom.

Table 5. Features most strongly associated engaged concentration and boredom

ENGAGED CONCENTRATION			BOREDOM		
TAG	SEMANTIC DESCRIPTION	ρ	TAG	SEMANTIC DESCRIPTION	ρ
N5-	Quantities	-0.182	T1.1.2	Time: General: Present; Simultan'us	-0.051
N3.6	Measurement: Area	-0.178	A5.2+	True/False	-0.041
Z5	Grammatical bin	-0.165	X2	Mental actions & processes	-0.041
Q2.2	Speech Acts	-0.162	A2.1+mwu	Modify, Change	-0.034
A4.1	Generally/kinds/ groups/examples	-0.161	M6mwu	Location & Direction	-0.034
<em>	HTML italics	-0.144	T3-	Time: Old, new and young; age	-0.034
A6.3+	Variety	-0.143	A8	Seem/Appear	-0.030
A5.4+	Authenticity	-0.130	T2++mwu	Time: Beginning and ending	-0.028
<p>	HTML paragraph	-0.127	A11.1+++	Important/Significant	-0.027
Z7	If	-0.116	A6.1+++	Similar/Different	-0.027
A4.2+	Particular/general; details	0.068	A5.4+	Authenticity	0.059
N3.5	Measurement: Weight	0.069	Z8c	Pronouns etc.	0.061
N3.1	Measurement: General	0.074	A6.3+	Variety	0.063
S5+c	Groups and affiliation	0.074	N1	Numbers	0.076
A2.1+mwu	Modify, Change	0.075	S6+	Obligation and necessity	0.076
A6.1+++	Similar/Different	0.077	N5-	Quantities	0.078
T3	Time: Old, new and young; age	0.082	Q2.2	Speech Acts	0.083
A2.1+	Modify, Change	0.083	A4.1	Generally/kinds/ groups/examples	0.085
Y1	Science/technology general	0.112	Z5	Grammatical bin	0.086
A5.2+	True/False	0.115	N3.6	Measurement: Area	0.093

Interestingly, X2.1 (*thoughts/beliefs*) is not as closely related to boredom and engagement as it was to confusion and frustration, but two other features typically associated with language about humans show desirable associations with these two outcome measures. For instance S5+c (*groups & affiliation*) is associated with increased engaged concentration, while X2 (*mental actions/processes*) is associated with lowered boredom. Likewise A8, which tags words related to *seem* or *appear* (both mental processes typically ascribed to human subjects), also leads to lowered boredom.

These semantic features, along with several others that correlate with lowered boredom (T2++mwu *time demarcations* and M6mwu *location/direction*) may also be indicators that problems with greater narrativity improve student engagement. However, we must still be cautious about interpreting lower boredom as a desirable effect in and of itself, since A5.2+ (words associated with *true*) is also associated with lower boredom. This type of item is unlikely to bore students, since they can answer and pass it quickly. However, readers may recall that this feature is also correlated with lower learning, as one might expect based on previous research on True/False questions [42].

5. DISCUSSION AND CONCLUSIONS

Our analyses of the ASSISTments corpus complements previous research on the relationship between learning and the language of mathematics problems, but extends this line of inquiry by including educationally relevant behaviors and affective states as part of the learning outcomes measured. As discussed, a number of linguistic features (e.g., *pronoun*s, *mental states, time*, and *concretenes*s) have been found to be significant in previous work. However, we were also able to examine the degree to which these relationships reflect expectations about how behavior, affect, and learning are related.

For instance, some of the same features which were correlated with learning were also correlated with student frustration and gaming the system. While it might be hypothesized that frustrated students would be more likely to game the system, there is also evidence from within ASSISTments that frustration can be important for learning [26]. The MBML model used here is a look-ahead algorithm, which may optimize the opportunity to identify the problems that trigger learning even when learning process is causing student frustration. However, it's also possible that these problems are triggering strong but distinct reactions in different students (e.g., students who persist vs. students who game the system when they become frustrated). Future work will hopefully shed more light on this unusual relationship.

Overall, these results point to a number of promising avenues for further research within the ASSISTments system. One key future approach will be to conduct RCTs of the features identified in this study, re-designing problems to eliminate problematic features or incorporate positive features, in order to determine whether our findings can drive enhanced design. At the same time, it will be important to explore some of the interactions that may exist between different combinations of linguistic features, or between linguistic features and other behaviors or actions within the tutor.

We also found several unusual patterns in our data, such as some features being associated with increases in both learning and with gaming the system. We believe this may be due to our dataset containing two different populations of students – those who are persistent in the face of challenging and difficult problems and those who are frustrated by these problems and attempt to game the system to avoid working through them. We hope to understand this relationship in greater detail through RCTs (as discussed below). Ultimately, we hope to use our findings to construct guidelines for teachers creating their own content in the system, which can be embedded directly into the authoring tools teachers use, providing useful feedback on their problem design.

5.1 Randomized Controlled Trials

Having found a set of features that are associated with differences in student engagement and learning, our next step will be to conduct a set of randomized controlled trials (RCTs) to test whether the effects we found are genuinely causal, and whether re-designing problems based on these findings can improve student outcomes. By determining which of these features are causal, we can expand scientific understanding of learning and engagement in online learning systems. By developing methods for concretely improving math problems, we can develop better guidelines and recommendations for the many instructors (and others) developing problems for the ASSISTments platforms. In the longer-term, we hope to make all of the problems in the ASSISTments platform engaging and educationally effective for each of the growing number of students who use ASSISTments to learn mathematics and other subjects.

5.2 Continued Feature Engineering

Another important area of future work will be to conduct further feature engineering, particularly in terms of text features specific to the language of mathematics. One of the shortcomings of the current study is that the language of mathematics is poorly modeled in existing tools. In addition to challenges cause by domain or context-specific uses of certain words, many semantic taggers rely on syntactic probabilities that may be difficult to capture when math problems are interspersed with text. Simply developing taggers that can identify embedded mathematics formulas (e.g., labeling '3+2' as addition) could help to ameliorate this issue. We hope that, by developing more robust tools for the analysis of this particular corpus, we will be able to better predict and understand learning and engagement.

As research progresses, features derived from combinations of Wmatrix tags will also become important since many of the sub-categories within and across Wmatrix's lexical fields may be semantically similar enough, or co-occur frequently enough, to warrant combining them within ASSISTments data. For example, Wmatrix treats *deciding* as separate from *choosing*, *selecting*, and *picking*, but this division may not be useful in mathematics learning corpora. Likewise, feature combinations may help to contextualize Wmatrix categories that are prone to incorrectly categorizing high-frequency words. For example, since many features in this study are highly correlated with M1, combinations involving this tag may be used to differentiate its use in instructions to students (e.g., "You have 3 attempts *left*") from its use in physical descriptions related to geometry (e.g., "Jill turns *left* and walks 3 more miles.").

5.3 Directions for Future Work

In this paper, we discovered relationships between semantic elements of text in the ASSISTments system and learning, affective, and behavioral student outcomes. In doing so, this work contributes to the emerging body of research studying the design of mathematics problems at scale.

Our findings show that a large number of semantically meaningful relationships exist, some of which correlate with a wide range of learner outcomes. These features provide insights that will help to develop guidelines for effective problem designs in ITSs. However, the existing suite of tools available for large scale textual analysis may not be optimal for tagging the specialized language of mathematics found in the ASSISTments system. Thus an additional area for future work includes the development of semantic taggers that are more appropriate for mathematics corpora. These efforts will help us to better understand how the linguistic properties of math problems influence student success at scale. In turn, by exploring potential relationships between persistence and student perceptions of challenge, we can work to design mathematics problems that are both more informative and more engaging.

6. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (NSFDRL 1252297). Any opinions and findings expressed are the authors' and do not necessarily reflect the views of the NSF.

7. REFERENCES

[1] Adar, E., & Datta, S. Building a Scientific Concept Hierarchy Database (SCHBASE). *Ann Arbor, 1001*, 48104.

[2] Baker, R. S., Goldstein, A. B., & Heffernan, N. T. (2011). Detecting learning moment-by-moment. International Journal of Artificial Intelligence in Education, 21(1-2), 5-25.

[3] Barber, C. (1962). Some measurable characteristics of modern scientific prose. *Contributions to English syntax and philology*, 21-43.

[4] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Royal Stat. Soc. B: Methodological, 289-300.

[5] Binder, J., Westbury, C, McKiernan, K, Possing, E, Medler, D. (2005). Distinct brain systems for processing concrete & abstract concepts. *J. Cog. Neurosci.*, *17*(6), 905-17.

[6] Carriedo, N., Elosúa, M., & García-Madruga, J. (2011). Working memory, text comprehension, and propositional reasoning: A new semantic anaphora WM test. *The Spanish J. Psych.*, *14*(01), 37-49.

[7] Cocea, M., Hershkovitz, A., & Baker, R.S. (2009). The impact of off-task and gaming behaviors on learning: immediate or aggregate?

[8] Cook, A. E., Myers, J. L., & O'Brien, E. J. (2005). Processing an anaphor when there is no antecedent. Discourse Processes, 39(1), 101-120.

[9] D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning & Instruction*, 22(2), 145-157.

[10] Drouin, P. (2010). Extracting a bilingual transdisciplinary scientific lexicon. eLexicography in the 21st C: new challenges, new applications. Louvain-la-Neuve: Presses Universitaires de Louvain/Cahiers du CENTAL, 43-53.

[11] Ferketich, S. (1991). Focus on psychometrics. Aspects of item analysis. Research in Nursing & Health, 14(2), 165-168.

[12] Forsyth, R., Ainsworth, S., Clarke, D., Brundell, P., & O'Malley, C. (2006). Linguistic-computing methods for analysing digital records of learning. In Online Proceedings of the 2nd International Conf. on e-Social Science, 28-30.

[13] Gacitua, R., Sawyer, P., & Rayson, P. (2008). A flexible framework to experiment with ontology learning techniques. Knowledge-Based Systems, 21(3), 192-199.

[14] Gowda, S.M., Pardos, Z.A., & Baker, R.S. (2012). Content learning analysis using the moment-by-moment learning detector. In Intelligent Tutoring Systems (pp. 434-443). Springer Berlin Heidelberg.

[15] Graesser, A. McNamara, D, Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193-202.

[16] Graesser, A., McNamara, D., Louwerse, M. (2012). Sources of text difficulty: Across the ages and genres. Sabatini & Albro, Assessing reading in the 21st C.: Aligning & applying advances in the reading & measurement sciences. Lanham, MD: R&L Education.

[17] Heffernan, N., & Heffernan, C. (2014). The ASSISTments Ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *Inter'l J. Artificial Intelligence in Ed.*, *24*(4), 470-497.

[18] Heiner, C., Beck, J., & Mostow, J. (2004). Improving the help selection policy in a Reading Tutor that listens. In *InSTIL/ICALL Symposium*.

[19] Jessen, F., Heun, R., Erb, M., Granath, D., Klose, U., Papassotiropoulos, A., & Grodd, W. (2000). The concreteness effect: Evidence for dual coding and context availability. *Brain & Language, 74*(1), 103-112.

[20] Kittur, A., & Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: quality through coordination. Proc. of ACM Conf. on Computer-supported cooperative work 37-46.

[21] Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proc. of the ACM SIGCHI conference on human factors in computing systems* 453-456.

[22] Klin, C. M., Weingartner, K. M., Guzmán, A. E., & Levine, W. H. (2004). Readers' sensitivity to linguistic cues in narratives: How salience influences anaphor resolution. *Memory & Cognition*, *32*(3), 511-522.

[23] Kousta, S., Vigliocco, G, Vinson, D, Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. *J. Exp. Psych: General*, *140*(1), 14.

[24] L'Homme, M. C. (2003). Capturing the lexical structure in special subject fields with verbs and verbal derivatives. A model for specialized lexicography. *International Journal of Lexicography*, *16*(4), 403-422.

[25] Light, L., & Capps, J. (1986). Comprehension of pronouns in young and older adults. *Developmental Psych.*, *22*(4), 580.

[26] Liu, Z., Pataranutaporn, V., Ocumpaugh, J., & Baker, R.S. (2013). Sequences of Frustration and Confusion, and Learning. In *EDM*, 114-120.

[27] Loginova, E., Gojun, A., Blancafort, H., Guégan, M., Gornostay, T., & Heid, U. (2012). Reference lists for the evaluation of term extraction tools. *Terminology & Knowledge Engineering (TKE'12)*, 30.

[28] Martin, J. H. (1994). METABANK: A KNOWLEDGE-BASE OF METAPHORIC LANGUAGE CONVENTIONS. *Computational Intelligence*, *10*(2), 134-149.

[29] McGregor, M. (1991). Language, culture and mathematics learning. McGregor & Moore (Eds.), *Teaching mathematics in the multicultural classroom: A resource for teachers and teacher educators*, 5-25. U. of Melbourne, School of Mathematics & Science Education.

[30] Morin, E., & Daille, B. (2010). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, *44*(1-2), 79-95.

[31] O'Toole, J. M. (1998). Climbing the fence around science ideas. *Australian Science Teachers Journal*, *44*(4), 51.

[32] Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., Heffernan, C. (2014) Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology, 45* (3), 487-501.

[33] Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T., Salvi, A. van Velsen, M., Aghababyan, A., Martin, T. (2015). HART: The Human Affect Recording Tool. *Proc. of the ACM Special Interest Group on the Design of Communication (SIGDOC).*

[34] Pardos, Z. A., Baker, R. S., San Pedro, M., Gowda, S. M., & Gowda, S. M. (2014). Affective states and state tests: investigating how affect and engagement during the school year predict end-of-year learning outcomes. 1(1), 107-128.

[35] Paterson, D. G., & Tinker, M. A. (1946). Readability of newspaper headlines printed in capitals and in lower case. *Journal of Applied Psychology*, *30*(2), 161.

[36] Patterson, O. (2012). Automatic domain adaptation of word sense disambiguation based on sublanguage semantic schemata applied to clinical narrative (Dissertation, U. Utah).

[37] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007*. Austin, TX: LIWC. Net.

[38] Piao, S., Rayson, P., Archer, D., & McEnery, A. M. (2004). Evaluating lexical resources for a semantic tagger.

[39] Rayson, P. (2008). Wmatrix corpus analysis and comparison tool. Lancaster University.

[40] Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. Reading & Writing Quarterly, 23(2), 139-159.

[41] Schutz, N. (2013). How specific is English for Academic Purposes? A look at verbs in business, linguistics and medical research articles. *Language and Computers*, *77*(1), 237-257.

[42] Toppino, T. C., & Ann Brochin, H. (1989). Learning from tests: The case of true-false examinations. *The Journal of Educational Research*, *83*(2), 119-124.

[43] van de Sande, B. (2013). Measuring the moment of learning with an information-theoretic approach. In EDM 288-291.

[44] Walkington, C., Clinton, V., & Howell, E. (2013). The associations between readability measures and problem solving in algebra. Martinez. & Castro Superfine, Eds. *Procs 35th meeting of the N. Am. Ch. Inter'al Group Psych. of Mathematics Ed. (pp.* 86-89). Chicago, IL: U. Ill, Chicago.

[45] Walkington, C., Clinton, V., Ritter, S. N., & Nathan, M. J. (2015). How readability and topic incidence relate to performance on mathematics story problems in computer-based curricula. *Journal of Educational Psychology*, *107*(4), 1051.

[46] Weld, D. S., Adar, E., Chilton, L., Hoffmann, R., Horvitz, E., Koch, M., ... & Mausam, M. (2012). Personalized online education—a crowdsourcing challenge. In Workshops at the 26th AAAI Conference on Artificial Intelligence (pp. 1-31).

[47] White, S. (2012). Mining the Text: 34 Text Features That Can Ease or Obstruct Text Comprehension and Use. Literacy Research and Instruction, 51(2), 143-164.

[48] Wiest, L. (2003). Comprehension of mathematical text. Philosophy of mathematics education journal, 17, 458.

[49] Baker, R.S.J.d., de Carvalho, A.M.J.A., Raspat, J., Aleven, V., Corbett, A.T., Koedinger, K.R. (2009) Educational Software Features that Encourage and Discourage "Gaming the System". *Proceedings of the 14th International Conference on Artificial Intelligence in Education,* 475-482.

[50] Gibbons, J. D., & Chakraborti, S. (2011). *Nonparametric statistical inference* (pp. 977-979). Springer Berlin Heidelberg.