Using Natural Language Processing Tools to Develop Complex Models of Student Engagement

Stefan Slater, Jaclyn Ocumpaugh, Ryan Baker
Penn Center for Learning Analytics
University of Pennsylvania
Philadelphia, Pennsylvania, USA
{slater.research,
jlocumpaugh,ryanshaunbaker}@gmail.com

Ma. Victoria Almeda
Human Development
Teachers College Columbia University
New York, New York, USA
mqa@tc.columbia.edu

Laura Allen
Department of Psychology
Arizona State University
Tempe, AZ, USA
LauraKAllen@asu.edu

Neil Heffernan
Department of Computer Science
Worcester Polytechnic Institute
Worcester, MA, USA
nth@wpi.edu

***Abstract*—This paper examines the effect of different linguistic features (as identified through Natural Language Processing tools) on affective measures of student engagement using a discovery with models approach. We build on previous literature, using automated detectors that identify when a middle-school student using an online mathematics tutor is experiencing boredom, confusion, frustration, or engaged concentration, to identify which problems are most engaging (or not) at scale. We then apply previously validated NLP tools to determine the degree to which engagement findings may be related to the linguistic properties of word problems, contributing to a growing literature on the effects of language on mathematics learning.**

1. Introduction

Affective research in education has often focused on constructs that are thought to increase or inhibit learning, such as boredom, confusion, engaged concentration, and frustration [2; 15]. Researchers have proposed theoretical models of the transitions between these constructs, such as [7]'s model, which predicts that when students reach an impasse in problem solving they are likely to alternate between episodes of confusion and concentration. Comparatively less work, however, has focused on aspects of the learning experience that may lead to these differences in affect.

Much of the fine-grained focus on the impact of specific educational content/design has focused on the impact these details have on fairly direct measures of learning, including work examining the effectiveness of hints [11] and the learning associated with specific problems [9]. Other work has investigated the relationship between design features of learning systems and engagement [3] or affect [8]. However, the types of features found in that work – such as evidence that equation-solving problems lead to better affect and engagement than brief word problems – do not provide much scope for enhancing mathematics problems, since word problems are an established part of education.

In this study, we seek to better understand how finer-grained aspects of the design of learning content influence student affect during learning by utilizing two recent advances in computing research, natural language processing (NLP) [5] and interaction-based affect detection [4]. Namely, we extend recent research [23] which used NLP to examine the relationship between semantic categories and student affect while working on mathematics problems in the online tutor ASSISTments [10; 21], to look at these relationships in closer detail. In this study, we combine the semantic features studied in [23] with more sophisticated linguistic measures to develop multi-feature models of linguistic predictions of student affect, creating the potential for understanding how linguistic features may influence and moderate one another and how these relationships are associated with differences in student engagement.

2. Previous Research

One potential area for understanding and evaluating how the fine-grained aspects of learning content is related to affective responses is through natural language processing, or NLP. Languages are highly complex, and often exhibit non-compositional patterns such as idioms and metaphors.

Comprehensibility and structure of textual content may underlie differences in student engagement and learning, but relationships between the features of the language in learning content can be difficult to study at scale.

As linguistic tools have become more powerful and more sophisticated, researchers have used them to better understand the role of language in mathematics education. For example, [24] found that correct answers and fewer hint requests are associated with word problems using third-person singular pronouns (e.g., he, she). They also found relationships between specific semantic categories in problem content and learning.

Although there is a growing literature on the relationships between language and mathematics learning, there are fewer studies that examine the relationship between the language of the learning context and student engagement. At the same time, interest in the complex relationships between student engagement and learning continues to grow [19]. Research by [23] examined the correlations between 442 semantic tags from the linguistic analysis tool Wmatrix [20] towards understanding the relationship between word choice and student affect and behavior in the mathematics tutor ASSISTments. However, this correlation mining approach suffered from the inability to investigate the effects of multiple features used in tandem. It was also limited in the kinds of linguistic features it investigated (semantic categories), and other factors, including those related to readability, are likely to show relationships with student engagement.

3. Data

3.1. ASSISTments Math Problems

We used data from the ASSISTments intelligent tutoring system for this study. ASSISTments is designed to assess students' mathematics knowledge while using automated scaffolding and hint messages to assist in learning [10]. The ASSISTments ITS is used by tens of thousands of students nationally each year, concentrated mostly in the northeastern US. One important feature of the design of ASSISTments that makes it particularly well-suited for the analysis conducted here is that ASSISTments contains a large variety of mathematics problems, as it allows teachers to author their own mathematics problems and share them with other teachers [21]. As such, ASSISTments content has a much broader variation in design than most other online learning systems.

3.2. Learners

Data for this study was generated from the 22,225 unique students nationwide who completed mathematics problems through the ASSISTments system as part of their regular instruction during the 2012-2013 school year.

4. Methods

4.1. Data Selection/Aggregation Across Word Problems

Learners in this study completed 179,908 different mathematics problems within ASSISTments, however, problems were filtered based on their appropriateness for analysis. Exclusion criteria included problems with fewer than 10 words or which were completed by fewer than 50 students. This resulted in 114,893 different problems in the final dataset. Data were aggregated at the problem-level, such that an average value for each outcome measure was produced for each problem.

4.2. Measures of Engagement

Models constructed from *in situ* classroom observations of student engagement, developed using the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP 2.0), were applied to student log files to allow a for retrospective analysis. In this method, which has been used to study student engagement in over a dozen different learning systems, a BROMP-certified coder records observations on a handheld app [HART; 18]. The data are then synchronized with the log files of the students who were observed, allowing researchers to examine how patterns of student interactions with the software vary depending on the observed indicators of student engagement [4]. Models developed for ASSISTments and cross-validated for differences in subpopulations [17] were used in the present study.

TABLE I. MODEL-FIT PERFORMANCE OF AFFECT MODELS [17]

Affect Model	**Kappa**	**A'**
Boredom	0.19	0.67
Confusion	0.38	0.74
Engaged Concentration	0.27	0.63
Frustration	0.17	0.59

4.3. Tools for Linguistic Analysis

To generate features for use in linguistic analyses, we used Wmatrix [20] and the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) [13]. Wmatrix has been used in previous research on the language of ASSISTments word problems, which demonstrated that semantic features of mathematics problems correlated individually to engagement [23]. While TAALES has not yet been applied to ITS and other online learning contexts, it has been used to assess written essay quality on the Michigan English Language Assessment Battery (MELAB; [12]) and to model students' vocabulary knowledge [1].

Wmatrix is a linguistic analysis tool that provides tags and identifiers for semantic domains (e.g. words that share similar meanings, such as 'sailboat' and 'yacht') and grammatical categories (e.g. first-person and second-person pronouns). The tagger matches individual words to a bank of 42,300 single word entries and 18,400 multi-word expressions, and also classifies individual words to a hierarchical structure of 21 lexical fields, with 234 base tags. Additionally, words can be tagged as antonyms, comparatives, superlatives, gender, and anaphorics. Our analyses identified 442 distinct Wmatrix tags within the set of problems we examined in ASSISTments, and full documentation for Wmatrix tags is available through the UCREL Semantic Analysis System (USAS) website at http://ucrel.lancs.ac.uk/usas/usas_guide.pdf.

TAALES is a tool for the evaluation of linguistic sophistication. It provides information about word frequency, range, bigram and trigram frequency, academic language, age of exposure, and updated psycholinguistic norms, which were not included in other current linguistic tools such as Cohmetrix [14]. Previous studies have used TAALES to predict second language acquisition and assessment [13] and math performance in standardized tests [6]. In the ASSISTments data, we calculated 137 of 485 TAALES indices for each word problem in our corpus, and full documentation for all 485 TAALES indices is available by downloading the Index Description Spreadsheet at http://www.kristopherkyle.com/taales.html.

Together, Wmatrix and TAALES comprise a broad set of English-language features, but working with mathematics tutor data involves the identification of mathematics-specific language such as equations, symbols, and numeric expressions. To identify these features we used HTML data in the ASSISTments problems to identify common mathematical features, such as the symbol for degrees (&deg;) and square roots (&sqrt;). We also included multiple design features that have been previously highlighted in research on affect in online learning [17] – this included descriptive information about the number of hints and scaffolds associated with a problem, the type of answer expected by the system (e.g. multiple choice, fill in the blank), as well as averaged performance data on the problems such as the number of successful and failed attempts. These features allowed us to account for differences in non-linguistic problem construction, differences in the degree of support provided to students, and differences in general problem ease or difficulty.

4.4. Model Development

We constructed a set of four linear regression models predicting affect from problem design features, one for each affective state, using the machine learning software RapidMiner [16]. RapidMiner is a machine learning package that fits and validates a variety of models. For this research, we used forward feature selection processes to determine which features contributed most strongly to the prediction of our outcome variables. In forward feature selection, an algorithm chooses the one feature that makes the greatest contribution to the outcome, and adds this feature to the model. It then adds the feature that makes the second-greatest contribution, after taking the first feature into account, and tests the model improvement. We continued this iterative process until we failed to significantly improve the regression model, or we had included eight features, whichever happened first.

Three-fold student level cross validation was used to split training and testing sets for model validation. Forward feature selection was only conducted using the training set; each testing set was entirely held out from analysis.

5. Results

5.1. Overview of Model Performance

We inspect the model-fit of our regression models based on the following goodness metrics: RMSE, Squared Error, and Spearman ρ. Table II shows the performance of each of our affect models. In general, we find that the confusion model performed the best, with the highest correlation and lowest RMSE and squared error values. When considering the strength of the correlations, the next best performing model is that of boredom, followed by frustration. The concentration model performs the worst, with a considerably weaker correlation and higher error values relative to the other affect models.

TABLE II. CROSS-VALIDATED MODEL-FIT PERFORMANCE OF NLP-BASED AFFECT MODELS

Affect Models	Goodness Metrics		
	RMSE	***Squared Error***	***P***
Confusion	0.042	0.002	0.238
Boredom	0.138	0.019	0.203
Frustration	0.078	0.006	0.165
Concentration	0.441	0.195	0.079

To better understand how the linguistic content, semantic content, and mathematics-specific language relate to affect, we examine the features of each of our affect models. Because of the size of the dataset we will primarily be examining the features with the highest β value and coefficients, rather than those with the lowest p values.

5.2. Linguistic Features of Confusion

Table III shows the list of features for our confusion model, which drew from all three feature types. BNC Written Trigram Frequency Normed (word) was found to be the strongest predictor of confusion, when controlling for other features, $\beta = -1.503$, $p < 0.0001$. In particular, commonly written trigrams were associated with less confusion whereas commonly spoken bigrams were associated with more confusion. These findings are somewhat surprising as we would have expected the opposite pattern of results, with students experiencing less confusion during problems that using words more typical of spoken (rather than written) language. Other features associated with more confusion included the Wmatrix feature X9.2- (words associated with failure, e.g. incorrect) and whether the problem had a single hint associated with it.

TABLE III. FEATURES OF THE CONFUSION MODEL

Features	Coeff.	SE Coeff.	β	p
KF Ncats Content Words	0.001	0.0000	0.000	<0.0001
BNC Spoken Bigram Normed (bi) Freq	0.010	0.0004	0.007	<0.0001
One hint	0.009	0.0003	0.014	<0.0001
BNC Spoken Bigram Normed (bi) Freq Log	0.003	0.0002	0.001	<0.0001
Average failed attempts over the year	0.010	0.0002	0.017	<0.0001
X9.2- (terms about success or failure)	0.022	0.0006	0.088	<0.0001
BNC Spoken Trigram Normed (word) Freq	0.036	0.0011	0.018	<0.0001
BNC Written Trigram Frequency Normed (word)	-1.378	0.0340	-1.503	<0.0001

5.3. Linguistic Features of Boredom

Table IV summarizes the list of features predicting boredom. Unlike the findings from the confusion model, written and spoken trigrams, as well as written bigrams, were positively associated with boredom. In other words, students were more bored when common combinations of words were present in the problems. Students were less bored, however, when problems used language from the Academic Formulas List (AFL; [22]), which contains linguistic sequences that appear more frequently in academic writing (e.g. "such as" and "an example of"). Taken together, these two findings suggest that academic wordings may be more interesting to students than simpler, more colloquial phrasings.

TABLE IV. FEATURES OF THE BOREDOM MODEL

Features	Coeff.	SE Coeff.	β	p
Kuperman AoA Content Words	0.002	0.0002	0.001	<0.0001
Is a base problem	0.01	0.0012	0.005	<0.0001
BNC Written Bigram Freq Normed (word) Log	0.004	0.0014	0.001	0.0131
BNC Spoken Trigram Freq Log	0.001	0.0001	0.001	<0.0001
Answer is fill in the blank	0.002	0.0008	0.002	0.0038
BNC Written Trigram Freq Normed (tri) Log	0.005	0.0012	0.002	<0.0001
Has hint	0.011	0.001	0.008	<0.0001
All AFL Normed	-0.028	0.0227	-0.074	0.3105

Interestingly, none of the features selected for inclusion into the boredom model were drawn from the semantic tagger Wmatrix. Instead, 6 of 9 features were drawn from TAALES (text complexity) measures, while the remaining features had to do with other elements of the problem's design.

5.4. Linguistic Features of Frustration

As shown in Table V, the strongest predictor of frustration (WMatrix's A5.3- label) involves the semantic content of the problem text. This feature identifies words which concern the evaluation of accuracy, specifically inaccuracies ("wrong", "error", "mistake"), and is associated with more frustration, $\beta = 0.493$, $p < 0.0001$. This feature also appeared in instances of meta-text (e.g., instructions about mindset rather than about the operations necessary to complete the problem).

Semantics also drives the second strongest feature in the frustration model, WMatrix's T3-, terms about relating to age/ maturity, $\beta = 0.160$, $p < 0.0001$. However, after these semantic categories are entered into the model, all of the other features related to confusion are drawn from the TAALES measures of lexical sophistication. Features associated with word frequency were associated with less frustration in students. This finding complements the findings for boredom, where the model shows that common function and spoken words appeared are associated with higher boredom, suggesting that problems written in a simple, non-academic vernacular are less frustrating, but also less interesting.

TABLE V. FEATURES OF THE FRUSTRATION MODEL

Features	Coeff.	SE Coeff.	β	p
BNC Spoken Bigram Normed (bi) Freq Log	0.027	0.0005	0.010	<0.0001
KF Nsamp All Words	0.000	0.0000	0.000	<0.0001
A5.3- (terms about accuracy)	0.083	0.0014	0.493	<0.0001
SUBTLEXus Freq Content Words Log	0.008	0.0003	0.003	<0.0001
T3- (terms relating to age or maturity)	0.031	0.0009	0.160	<0.0001
Brown Freq Function Words	0.000	0.0000	0.000	<0.0001
BNC Written Freq Function Words Log	-0.012	0.0003	-0.006	<0.0001
BNC Spoken Freq All	-0.004	0.0001	-0.003	<0.0001

Words

5.5. Linguistic Features of Concentration

Lastly, Table VI summarizes a list of predictors for engaged concentration. As expected, we find that bigrams and trigrams are associated with more concentration. The frequency of bigrams and trigrams are based on the British National Corpus, which has extensively used in prior lexical research. Examples of the trigrams from BNC index are as follows: “one of the”, “I don’t know”, and “a lot of”. Examples of frequently used bigrams from the BNC index are: “of the” and “in the”. Interestingly, we also find that abstract words related to being (e.g., existence) or effort (e.g., trying) are likely to increase concentration, even when controlling for other commonly used combinations of words.

TABLE VI. FEATURES OF THE CONCENTRATION MODEL

Features	Coeff.	SE Coeff.	β	p
KF Nsamp Content Words	0.000	0.0000	0.000	0.0114
BNC Spoken Trigram Normed (tri) Freq	0.018	0.0037	0.019	<0.0001
BNC Written Freq FW	0.000	0.0001	0.000	0.1723
A3+ (abstract terms related to being)	0.006	0.0015	0.010	<0.0001
X8+ (terms depicting effort)	0.018	0.0042	0.044	<0.0001
BNC Spoken Bigram Normed (bi) Freq	0.005	0.0043	0.004	0.3832
SUBTLEXus Freq Content Words	0.000	0.0000	0.000	1.0000
BNC Written Trigram Freq Normed (word) Log	0.016	0.0125	0.008	0.2712

6. Discussion/Conclusions

In this study, we investigated the impact of linguistic features of mathematics problems on students’ affective states during work in an online mathematics tutor. While previous research has examined the language of word problems, this study expands that work in several dimensions, including the kinds of linguistic features that were investigated and the outcome measures that were considered. We examined these features by developing complex models of linguistic features that are associated with student affect.

Our findings show that features related to commonly used combination of words are associated with positive concentration and negative confusion (and vice versa). These findings are aligned with previous research, which suggest that concentration and confusion are conceptually related to each other. Likewise, there appears to be relationships between the linguistic features associated with frustration and boredom, which are also theoretically aligned [7].

In addition to this, our findings reveal that features related to semantic content (i.e., WMatrix) are associated with frustration. Of particular interest, terms related to accuracy, such as “error” or “make a mistake” are associated with increases in student frustration. It is possible that the authors of these problems realized their difficulty and incorporated meta-instructions geared towards helping students to manage such emotions, but further research is needed to determine whether such messages are helpful.

Lastly, we find that the features most associated with boredom are related to the academic formula list (AFL). In particular, the use of academic words in math problems appears to lead to less boredom in problem solving. Our findings suggest that incorporating academic words in math problems could help in decreasing the likelihood of students’ boredom, as well as help improve their skills in receptive language when problem solving.

In sum, our findings provide implications for improving the design of math problems by focusing on key features linked to student’s affective states. It is our goal, going forward, to use these findings to help guide teachers to create math problems that promote better learning through improved student engagement.

Acknowledgments

We gratefully acknowledge our funders (NSF, DRL #1252297) for financial support on this project.

We gratefully acknowledge our four anonymous ACII 2017 reviewers for valuable feedback and insight on an earlier version of this paper.

References

[1] Allen, L. K., & McNamara, D. S. (2015). You are your words: Modeling students’ vocabulary knowledge with natural language processing. In *Manuscript submitted to the 8th International Conference on Educational Data*.

[2] Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners’ cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223-241.

[3] Baker, R. S., de Carvalho, A. M. J. A., Raspat, J., Aleven, V., Corbett, A. T., & Koedinger, K. R. (2009, June). Educational software features that encourage and discourage “gaming the system”. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 475-482).

[4] Baker, R.S.J.d., Ocumpaugh, J. (2014) Interaction-Based Affect Detection in Educational Software. In R.A. Calvo, S.K. D'Mello, J. Gratch, A. Kappas (Eds.), *The Oxford Handbook of Affective Computing*. Oxford, UK: Oxford University Press.

[5] Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51-89.

[6] Crossley, S., Liu, R., & McNamara, D. (2017, March). Predicting math performance using natural language processing tools. In *Proceedings of the 7th international conference on learning analytics and knowledge (LAK'17)*.

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

[8] Doddannara, L., Gowda, S., Baker, R.S.J.d., Gowda, S., de Carvalho, A.M.J.B (2013) Exploring the relationships between design, students' affective states, and disengaged behaviors within an ITS. *Proceedings of the 16th International Conference on Artificial Intelligence and Education*, 31-40.

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

[10] Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, *24*(4), 470-497.

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

[12] Jung, Y.J., Crossley, S. A., & McNamara, D. S. (2015). Linguistic Features in MELAB Writing Task Performances. CaMLA Working Papers, 5,1-17.

[13] Kyle, K. & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly 49*(4), pp. 757-786. doi: 10.1002/tesq.194

[14] Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, *36*(2), 193-202.

[15] McQuiggan, S., Lee, S., & Lester, J. (2007). Early prediction of student frustration. *Affective Computing and Intelligent Interaction*, 698-709.

[16] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD International Conf. on Knowledge Discovery & Data Mining,* 935-940.

[17] 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.

[18] Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T. (2015) *Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual.*. Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

[19] 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. *test*, *1*(1), 107-128.

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

[21] Razzaq, L., Patvarczki, J., Almeida, S. F., Vartak, M., Feng, M., Heffernan, N. T., & Koedinger, K. R. (2009). The Assistment Builder: Supporting the life cycle of tutoring system content creation. *IEEE Transactions on Learning Technologies*, *2*(2), 157-166.

[22] Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied linguistics*, *31*(4), 487-512.

[23] Slater, S., Baker, R., Ocumpaugh, J., Inventado, P., Scupelli, P., & Heffernan, N. (2016). Semantic Features of Math Problems: Relationships to Student Learning and Engagement.

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