Novice Programmers Talking about Projects: What Automated Text Analysis Reveals about Online Scratch Users' Comments

Nicole Forsgren Velasquez, Deborah A. Fields, David Olsen, Taylor Martin, Mark C. Shepherd, Anna Strommer, Yasmin B. Kafai*

Utah State University
2830 Old Main Hill
Logan, UT 84322
+1 (435) 797-0571, +1 (435) 797-3479, +1 (435) 797-2342, +1 (435) 797-0814

*University of Pennsylvania
3700 Walnut Street
Philadelphia, PA 19104
+1 (215) 746-346

nicolefv@usu.edu, deborah.fields@usu.edu, david.olsen@usu.edu, taylor.martin@usu.edu, mark.shep@aggiemail.usu.edu, anna.strommer@aggiemail.usu.edu, kafai@upenn.edu

Abstract

In this paper we examine the possibilities of applying predictive analysis to users' written communication via comments in an open-ended online social networking forum: Scratch.mit.edu. Scratch is primarily used by youth ages 8-16 years to program software like games, animations, and stories; their social interactions take place around commenting, remixing, and sharing computer programs (called projects). This exploratory work contributes to work in educational data mining by broadly describing and comparing comments about projects versus other topics in Scratch. Referencing communication accommodation theory, we found that user comments about projects exhibited different linguistic cues than other comments, and these cues were successfully used to classify comment topic. Further, results also suggest that project comments embody richer language than other comments. This suggests several future avenues for research on youth's online comments about programming and other technical projects that may reveal educational opportunities in creating and sharing projects.

1. Introduction

New online communities are emerging where primary social activities consist of contributing and discussing projects that users have made themselves. Although these types of sites exist primarily for adults (i.e., Wikipedia, Instructables, Deviant Art), they are also quickly gaining ground with children and youth. Some reports indicate steady increases in teens' sharing of self-created online content over the past several years [e.g., 25; 27], and there are indications that sites specifically developed for *children* to share creations are increasing in number [16]. These websites include places where kids can share written stories or fanfiction (e.g., Fanfiction.net, Storybird.com), mods or adaptations of popular games (e.g., Little Big Planet, The Sims), and, as we focus on in this paper, computer programs that can take the form of video games, animations, stories, or art (e.g., Scratch, Kodu). Since the potential for learning when children design or make projects has been well documented for many years [7; 21], it is understandable why there is such excitement over this new phenomenon. Furthermore, sites such as Scratch are unstructured environments where engaged and authentic learning can occur, in contrast to constrained and structured learning environments such as cognitive tutors [4; 22].

Yet while websites where people post user-generated content (UGC) are increasing in number, we know relatively little about the quality of communication on these sites, especially on a large scale that can reveal trends across thousands of users. Further, we know even less about the quality of communication of young people on such websites. Given that children and youth are at different stages of development and that their language tends to differ from adults' language, it is important to study communication on websites dominated by them. Most importantly, communication itself is learning. Developing a better understanding of what children and youth are doing on these emerging types of websites will help us better grasp the opportunities in participating in such sites as well as the challenges to designing for richer communication on websites for children.

In addition, there may be value to understanding communication specifically about projects that children create. Although the benefits to learning by making projects are widely discussed [7], *sharing* and *discussing* creations may contain particular educational value. Heath [18] argued that sharing artistic creations in youth-based community organizations provided venues for receiving constructive criticism on projects as well as having a high risk, motivating goal of preparing a creation for a critical audience. As youth participated in these types of practices in specific arts communities, the language youth used changed: vocabulary developed, the structures of questions altered, and if-then conditional statements increased, the latter being a sign of seeing greater possibilities in the community's work (e.g., "if we do this, then…") [19]. Yet these findings relate to local, in-person communities. How does sharing one's creations online provide opportunities for constructive criticism, audience, and language development? Is language about projects children make any different than other language on the websites?

In this study we apply communication accommodation theory (CAT) to investigate the nature of comments on the Scratch website, an unstructured and engaging learning environment where the average age of participants is 12. In particular we analyze the nature of technical comments specifically about projects compared to other comments. We apply predictive analysis techniques to analyze a random sample of 8,000 comments to answer the following questions:

- Are comments about projects different from general social comments on the Scratch website?
- In what ways are they different and what future directions of study does this analysis suggest?

By analyzing these questions on a large, random sample of data collected from the Scratch website, we aim to contribute to our understanding of 1) the nature of youths' communication in this emerging genre of websites focused on open-ended user-generate content, and 2) communication specifically about the technical nature of children's projects, and any educational potentials this type of web-based communication may hold.

2. Background

2.1. Communication accommodation theory

Communication accommodation theory (CAT) seeks to understand the ways in which speech and communication patterns shift among the people in organizations and communities [13]. These communication pattern shifts can include convergence, where people alter communication patterns, accents, and nonverbal cues to match those within their community; and divergence, where people alter these communication patterns to distinguish themselves from others in the community [14]. In our current investigation of Scratch users, we focus on the convergent communication patterns seen in text comments (and their associated linguistic cues) posted to the site.

CAT is guided by four assumptions: first, speech similarities exist in all conversations. Second, communication within a community is shaped by one's perception of others' communication. Third, language patterns can communicate group membership. Fourth, community norms shape expectations about communication patterns [31]. Although the dynamic nature of communication accommodation cannot be captured in a cross-sectional study, the resulting language patterns of this accommodation will be present when examined at a point in time. We posit that Scratch youth users, as members of an online social networking community, will signal their Scratch community membership by aligning their comments to group norms and therefore communicate about Scratch projects in distinctive ways. Furthermore, we posit that comments about Scratch projects will differ from comments about other topics (such as relationships or community norms), and that these differences can be used to classify comment topics. In other words, we propose that there are different cultural norms for talk about projects and other forms of talk on the Scratch site, and that these two forms of talk will be distinguishable.

3. Context & data

3.1. Context: Scratch.mit.edu

The context for this research focuses on Scratch[1] (http://scratch.mit.edu), an online community and social networking forum focused on young people's computer programs. Since the site was launched in May 2007 from the MIT Media Lab, over 1.4 million registered members have collectively developed over 3 million projects using Scratch, a media-based programming language that allows for the creation of games, stories, and animations [24]. Catering to youth, primarily 8- to 16-year olds, Scratch implements an intuitive building-block approach to programming that reduces the possibility of syntax errors while still encouraging computational thinking

[1] Of note, our research concerns the Scratch 1.0 website. The new Scratch 2.0 website, released May 2013, contains significant design changes, including the ways in which users may communicate with each other.

[28]. To encourage personalized project development, programmed objects can be any two-dimensional graphic image, either hand-drawn or downloaded from the Web.

This unique web-based integration of a simplified programming language, highly-customizable project design, and social-network framework has encouraged the evolution of a diverse participatory community centered around sharing and discussing the projects created by its members, who continue to upload over 1,500 new projects per month [12]. Users contribute to site interactions by sharing their own projects, downloading and "remixing" friends' programs, adding "favorite" projects to their portfolios, and commenting and clicking "love-it" on other users' projects. Descriptive statistics listed below each project show the number of times a project has been viewed, downloaded, "favorite"-ed, and remixed; as well as the locations of any user-created galleries that currently host the project. Popular projects have a chance to make it to the Scratch "front page," a prized area for Scratch developers, as a project on the front page receives more views, downloads, and feedback. While the primary function of the Scratch site is programming, project creation and social networking are deeply intertwined through numerous forms of participation [12].

3.2. Data

The data for this study was taken from a random sample of 5,004 users (of over 20,000 active users) who logged into the Scratch site in January 2012. This sample had similar demographics to all Scratch users in regards to self-reported gender and age (69.20% male, 30.80% female). All comments generated by these 5,004 users during the month of January 2012 were collected, yielding 36,802 comments. As a first step in this initial textual analysis of kids' comments on Scratch, we selected a subset of this data for our analysis. Prior research revealed that a primary gateway activity on Scratch was sharing a project; users who shared projects online were more likely to leave comments, click "favorite" and engage in other kinds of social networking activities [20]. This prior research suggested to us that certain types of comments, namely comments *about projects*, would be particularly interesting to investigate.

It is worth noting the limitations of our particular dataset of user comments. The comments we collected are limited to 500 characters. The overall dataset was selected on 5,004 randomly selected users in order to study broad participation patterns on the Scratch site. While we have all of the comments for the selected users, they are completely decontextualized from the broader context because of the method used to collect the data. In other words, we can only evaluate what each individual comment says, not who someone is replying to or what project the comment was left on. We cannot trace who has left comments on whose projects (i.e., social network analysis) or study the broader context of those comments. This is certainly a limitation of the data. However, the advantage to the data is that we can study broad trends on the Scratch site as it relates to the kinds of comments users write and the kind of language used in those comments. This was a purposeful choice to look across participants on the Scratch site rather than selecting a narrower context that would provide more information about selected users. The latter has been done a number of times to study smaller, cohesive groups within the broader Scratch site, such as collaborative groups, role-playing groups, and leaders. Instead, our analysis provides the opportunity to investigate broader trends on a large website.

4. Analysis

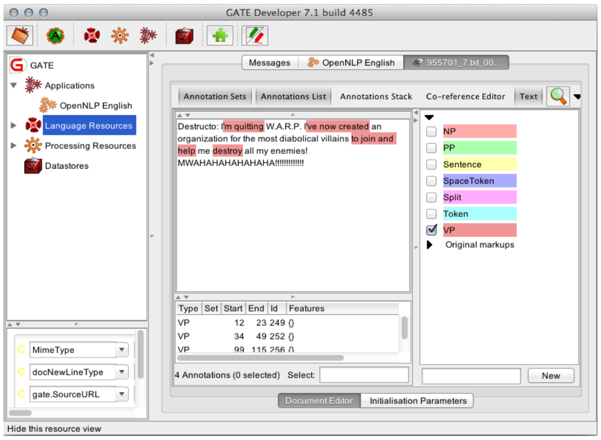
Educational data mining (EDM) methods can be utilized to inform educators about learning and communication on platforms such as Scratch [4]. To date, EDM methods have been used in many contexts [e.g., 2, 3, 6, 9]. However, the majority of the work has been focused on relatively constrained and structured learning environments like cognitive tutors [4, 22]. In contrast, Scratch presents an unstructured, authentic programming environment to support interest-driven learning.

New work utilizing analyses of speech and text has emerged as a way to study less structured environments, both local (in person) and online. For instance, such analyses have been applied to students' semi-structured interviews about astronomy [30] and to inquiry-driven science education in online environments [15]. While these studies carve new pathways for applying EDM methods to less structured environments, the Scratch website presents a particularly open online environment where interactions are not structured by prompts or directed activities but rather by users' broad interest in making and sharing projects in Scratch.

4.1. Message feature mining

We utilized message feature mining, a text analysis methodology, which can be used to classify messages based on linguistic features that are context- and content-independent [1]. This method is particularly good for this study because of the decontextualized nature of our data. When using this approach, two steps are followed: first, linguistic message cues are extracted over sections of text, and second, messages are classified using the extracted cues. In this study we utilize automated language coding to assist with cue extraction, and focus our analysis on language-based cues (LBC). LBCs separate (or "tokenize") linguistic information into text unit(s), also called terms [36]. These terms can include words, which are coded individually, and can be combined to make phrases, sentences or even entire messages, which are coded as a unit. Prior research suggests that certain LBCs can be detected with automated analysis methods [37], and that these features may be useful for text analysis across several contexts [32]. Examples of these cues are shown in Table 1 and include measures of quantity, complexity, diversity, and specificity [37].

LBC extraction of our user comments was conducted using GATE (General Architecture for Text Engineering), an open-source linguistic analysis tool that has been used in several projects that apply linguistic analysis methods [e.g., 32; 38]. GATE utilizes two types of resources: language resources, which contain the data to be analyzed, and programming resources, which are the types of analyses to be run on the data. To begin using GATE, language resources are identified. In the case of this analysis, 4,536 comments, all in individual text files, were loaded into the program (see below for explanation of this selection). Next, programming resources are loaded; these can include open source text analyzers, such as ANNIE [11], OpenNLP [5], or custom-programmed resources. Finally, an application is defined within the program, which consists of the language resource(s) to be analyzed and the programming resource(s) that are run on the text. GATE then executes this application, and annotates the text files. An example of GATE annotation of verb phrases (abbreviated VP) highlighted using the OpenNLP programming resource is shown below in Figure 1.



Destructo: I'm quitting W.A.R.P. I've now created an organization for the most diabolical villains to join and help me destroy all my enemies! MWAHAHAHAHAHAHA!!!!!!!!!!!!

Type	Set	Start	End	Id	Features
VP		12	23	249	{}
VP		34	49	252	{}
VP		99	115	256	{}

4 Annotations (0 selected) Select:

MimeType
docNewLineType
gate.SourceURL

Figure 1: GATE Annotation of Verb Phrases

Several programming resources within the GATE framework were used to extract LBCs. For example, one programming resource that is prepackaged with GATE was used to code each word individually according to its part-of-speech. These codes were then compiled to create word, verb, modifier, etc. counts for each comment. Another programming resource was used to code and count group references, such as "us," we," and "ours." In addition to part-of-speech and distinct word counts, dictionaries were used to identify and count cues like pleasantness (using an affect dictionary [33]) and spatial and temporal closeness [38]. The reader is referred to [38] for a complete discussion of the linguistic cues coded for this study; cues significant in this analysis are discussed in Section 5.2.

Classification of the messages follows cue extraction and involves: 1) manually classifying each message in the training set (described below), 2) selecting a classification method, 3) training and testing the model, and 4) evaluating the results. When selecting an appropriate classification method, the advantages and disadvantages of each must be considered [23]. For this study, we based our choice of classification method on the level of information available accompanying the results. While general-purpose neural networks show promise, they are limited in their explanatory power; that is, while they may reach a conclusion, no path to that conclusion can be traced. In contrast, decision trees outline logic that can be examined and further investigated. Therefore, we used the open-source J48 decision tree implemented in Weka [35] for the classification.

5. Results

Our reporting of results chronologically follows the steps we took in analysis. After randomly selecting the 8,000 comments from the larger dataset, we manually coded them into two primary categories: comments about projects and other comments (see section 5.1). Then we compared those two categories of comments using the linguistic based cues to see which cues were significantly different, and likely useful for classifying the comments (see section 5.2). Finally, we used those linguistic based cues where significant differences were found between project comments and other comments to conduct predictive analyses using the J48 decision tree (see section 5.3). Below we describe each of these processes and what they reveal about the comments youth leave on Scratch.mit.edu

5.1 Manual comment coding

Selected comments were manually coded to establish an accurate classification of each; this "ground truth" was later used to train and test the computer classification model. In other words, before we could apply automated analyses, we had to hand-code comments for a trustworthy dataset. For our analysis, we manually coded a random sample of 8,000 comments from the larger dataset to identify comments about projects and other types of comments. The sample size of 8,000 was chosen for two reasons. First, a cursory reading of the comments in our dataset suggested that one fourth addressed Scratch projects, and we were hoping to code a corpus of approximately 2,000 project comments. Second, the manual coding necessary to build and test classification models is time consuming. Therefore, 8,000 comments were randomly selected from the 36,802 available comments in the dataset using the RAND function in MySQL.

Our first step in this manual coding process was to create definitions for comment topic and provide examples of comments that fit into each category (see table below). The definitions for each category were discussed and decided upon by the research team. Once we had satisfactory, clear definitions, 800 comments (10% of our total 8,000 comments) were selected to establish reliability.

Table 1. Comment topic definitions

Category	Definition	Examples
Project	Talk clearly referencing projects, including making, critiquing, or playing a project. This includes responses to compliments ("Thanks!"), saying a project is amazing or praising a project ("Cool" "Awesome").	• I didn't make it I remixed! • Still good coloring and all. Appaloosa must be good at outline.
Other	Any comments that are not about projects and are not role play (common among youth comments). Includes community norms (crediting, remixing, socialization, etc).	• Thanks Paddle. I would be very grateful if you find the answer :) • That sure is unfortunate :(Don't worry though I still remember what your entry was like so it will still be counted.

Two researchers independently coded the set of 800 comments according to definitions presented above; analysis of the independently coded comments resulted in a Kappa of 0.94. Therefore, one researcher coded the remaining 8,000 comments, removing any non-English or role-playing[2] comments in the process. This coding identified 2,268 project comments and 5,259 other comments[3]. Because classification algorithms perform better with equal numbers of categorical data, 2,268 "other" comments were selected for the next stage of our analysis, resulting in a dataset of 4,536 comments.

5.2 Message feature mining results

We coded all messages (n = 4,536) based on 39 linguistic cues using GATE. We then compared the means of each LBC for project and other comments, and found that 14 were significantly different ($p < 0.01$). Overall, results suggest that youth are more thoughtful and engaged when commenting about projects. This can be seen in the means of the LBCs, which show higher quantity (i.e., overall words, verbs, and modifiers), expressivity (i.e., emotiveness), diversity (i.e., content and redundancy), specificity (i.e., spatial indicators and imagery), and affect (i.e., affect words, pleasantness, and activation). This is also manifest in lower levels of misspelling occurring in project comments. It is interesting to note that non-project comments have significantly more group (e.g., we, ours) and other pronouns (e.g., she, them), suggesting that those comments are more likely to contain references to collaborations or relationships.

Below, we present means and standard deviations for the LBCs that differ significantly

[2] Role-playing is a common practice in Scratch where users play act different characters, usually from popular media series. Because it was so common, yet unrelated Scratch programming or general site socializing, we sought to differentiate role play comments from comments about projects or other things.

[3] 349 comments were identified as "role playing" and 124 comments were not in English (the text analysis tools we used in our study only support English), and were excluded from the analysis.

between project and other comments, provide examples, and suggest possible interpretations. Note that all comments appear exactly as entered by the Scratch user; emphasis added.

Table 2: Mean (StdDev) of quantity LBCs

Linguistic Cues	Project	Other
Word Quantity	18.320 (19.719)	16.520 (19.113)
Verb Quantity	3.370 (4.196)	2.830 (4.094)
Modifier Quantity	2.400 (3.481)	1.890 (3.197)

Modifier quantity example, project comment: "I **really** liked how the background changed it made me feel like the guy was **actually** skydiving! The **flashing** stars are also **really** cool. The guys **skydiver** outfit is also **really** realistic!" (Modifiers: 6)

Overall word quantity, verb quantity, and modifier quantity (i.e., adjectives and adverbs) are all raw counts found in each comment. These values are higher for project comments than for other comments, indicating that project comments are longer and use more specific and descriptive language. Higher specificity could demonstrate a high quality of feedback on others' projects, such as in constructive praise or criticism (demonstrated in the example above). This is an area that has been specifically supported on the Scratch website [29], although its frequency in comments on Scratch projects has not yet been documented.

Table 3: Mean (StdDev) of expressivity LBC

Linguistic Cues	Project	Other
Emotiveness	0.270 (0.485)	0.210 (0.369)

Emotiveness is calculated as the total number of adjectives and adverbs (modifiers), divided by the total number of nouns and verbs. This is a measure of how expressive a comment is, with higher numbers indicating a higher proportion of modifiers used. Again, we see that project comments are more expressive than other comments, indicating that youth use more descriptive language when discussing projects.

Table 4: Mean (StdDev) of diversity LBCs

Linguistic Cues	Project	Other
Content Word Diversity	0.980 (0.066)	0.970 (0.101)
Redundancy	2.697 (3.871)	2.348 (3.537)

Redundancy example, project comment: "All **the** art **besides the** background **the** menu **and** curtains **are mine**. **I** made all **the** characters **and the** logo." (Redundancy: 0.550).

Diversity measures capture the mix of words used in comments. Redundancy is calculated as the number of function words divided by the total number of words, and content word diversity is calculated as the number of content words divided by the total number of words. In both cases, project comments score higher in diversity measures than other comments. Higher redundancy in project comments may indicate that users are more thoughtful when writing comments, being careful to always use articles, prepositions, etc. In turn, higher content word diversity may reflect the high use of verbs and modifiers in project comments, seen earlier. More research would be needed to verify these interpretations, but they remain consistent with the idea that comments about projects are more carefully written with increased use of verbs and modifiers.

Table 5: Mean (StdDev) of informality LBC

Linguistic Cues	Project	Other
Misspelled Words	0.100 (0.214)	0.210 (0.309)

Misspelled words is the ratio of misspelled words to total words per comment. In the case of the Scratch comments we analyzed, we see that project comments have half as many misspellings than other comments, which lends further support to the idea that users are more thoughtful and engaged when commenting about projects.

Table 6: Mean (StdDev) of specificity LBCs

Linguistic Cues	Project	Other
Spatial Far Ratio	0.030 (0.068)	0.020 (0.068)
Imagery	1.162 (0.650)	1.036 (0.733)

Spatial far example, project comment: "hey um do you know of a size for the sqaures that will make it equal? like see how the **last** row of blocks on mine are **outside** the **edge**? i need a size so they allign with the edges. (Im too leazy to do the math)" (Spatial far: 0.064)

Spatial far ratio is calculated by dividing the number of spatial far words (e.g., far, last, over; as defined by a dictionary [38]) by the total number of words in the comment. Imagery is the raw number of words that help paint a clear mental picture (e.g., red, bright, etc; as defined by a dictionary [38]).[4] In both

[4] Other specificity measures were coded, including ratios of spatial closeness (e.g., near, here), temporal immediacy (e.g., start, before) and temporal nonimmediacy (e.g., then, later), but these did not differ between project and non-project comments.

cases, specificity measures were higher for project comments than for other comments, indicating that commenters use much more descriptive and colorful language when talking about projects. Specificity could also play an important role in the quality of praise, criticism, or requests for help. The more specific a comment is about a project (as in this example) the more constructive the criticism, praise, or help might be.

Table 7: Mean (StdDev) of affect LBCs

Linguistic Cues	Project	Other
Affect Ratio	0.080 (0.191)	0.020 (0.069)
Pleasantness	1.559 (0.825)	1.341 (0.895)
Activation	1.381 (0.728)	1.209 (0.802)

Affect Ratio example, other comment: "Um... Ok? This is the first mean comment I've ever gotten so IDK how to react to this... If you don't have anything nice to say don't say anything and move on to the next page. That's the only way I can put it without sounding rude and **awful**. Have a nic" (Affect ratio: 0.019)

Affect measures are an indication of the level of emotive words used in comments. Affect ratio is calculated as the number of subjective emotion words divided by the number of total words per comment; pleasantness is the number of words that convey a feeling about an emotional state; and activation is then number of words that reflect emotional state dynamics [38]. In all measures of affect and emotion, project comments scored higher than other comments, indicating that Scratch users may display more emotion in comments about projects. This would not be surprising since projects are the primary areas for interaction and activity on the Scratch website [12] and there is a general effort amongst the designers to promote a positive online culture [10].

Table 8: Mean (StdDev) of pronoun LBCs

Linguistic Cues	Project	Other
Group References	0.001 (0.012)	0.004 (0.037)
Other References	0.004 (0.030)	0.010 (0.059)

Group reference example, "other" comment: "**We** all know you just want attention I dont blame you but this isn't the way" (Group reference: 0.063)

Group references and other references are coded as a ratio of group (inclusive) pronouns or group (other) pronouns divided by the total number of words in a comment; both were higher in other comments compared to project comments. This suggests that collaboration activities are more likely to be discussed outside the specific context of talk about projects, not surprising given that Scratch comments are limited to 500 characters. Singular pronouns (inclusive and other) were also coded, but did not differ significantly between project and other comments.

Below we discuss the ways that we used certain linguistic based cues to conduct predictive analyses about comments, to see whether a computer could automatically classify project and other comments. In the discussion we elaborate more on the differences between project and other comments and the future research that this analysis suggests might be productive.

5.3 Classification

We used the 14 LBCs that significantly differed between project and other comments in the J48 decision tree with ten-fold cross-validation in Weka; 2,906 comments were classified correctly and 1,630 comments were classified incorrectly, resulting in 64.07% accuracy (shown in Table 9).

Table 9: J48 classification results

	Classified as	
Actual	Project	Other
Project	1,462	806
Other	824	1,444
	Number	Percent
Correctly classified	2,906	64.07%

These results suggest that the significantly different LBCs identified by GATE can be used in a decision tree to classify project and other comments. The decision tree has 122 leaves (or decision end points) and the size of the three is 243. Compared to 50% accuracy from chance alone, the decision tree performed well. The root node of the tree (or initial decision point) is affect ratio, indicating that this LBC provides the most information alone.

6. Discussion

In this study we sought to understand whether there would be linguistic differences between Scratch users' comments about projects versus other topics. Communication accommodation theory suggests that people attune their language to cultural norms, and we hypothesized that there would be different cultural norms about project talk versus other kinds of talk (i.e., talk about relationships or small talk), and that these norms would be observable in linguistic cues. Indeed, our analyses using message feature mining revealed significant differences between project comments versus other kinds of comments. Furthermore, these LBCs that differed between project and other comments were successfully used to automatically classify comments with 64.07% accuracy. Our results show that project comments tend to be richer in many measures of language content: length, verbs, word content, adjectives, imagery, and even pleasantness. While earlier research by Heath [18, 19] and others found that in-person talk about making and creating things held educational promise for young people, our research suggests that online talk about projects may hold similar promise for engaging youth in rich language.

These results also provide evidence that youth craft their comments about projects to match other comments about projects. While the dynamic process of adapting comments to match the communication style of a community is not visible in this type of cross-sectional study, the similarity of project comments with other project comments and the dissimilarity to non-project comments provide evidence that the shift has happened. These differences were observable in LBCs with significant differences and in their combined use to classify comment type. These results provide support for communication accommodation theory in youth communication on Scratch.

One way that research with "big data" and analytics can contribute is to suggest areas for deeper investigations in more focused areas. In this study we analyzed a large set of user comments from across the Scratch site and found that project comments are not only qualitatively different from comments about other topics in ways that can be computationally identified by message feature mining. We also found that project comments exhibit more specific language richer in modifiers and affect, as well as sheer number of words and verbs. This suggests a specific focus for further analyses on project comments to elicit 1) the role of modifiers and specificity in project comments (i.e., what role do they play in constructive criticism or specific praise, and are there other functions of those linguistic cues), 2) the ways that affect is expressed and what role that plays in relations on the site. Other more quantitative analyses could study the diversity of users who engage in richer language in project comments. For instance, what percentage of comment writers engage in comments with higher values of modifiers or affect words? Or do certain users tend to be more verbose versus others?

With the growing availability and interest in online learning environments, research on language use and communication by youth is imperative, and several research directions should be considered. Similar analyses could be used to compare Scratch user comments to communication on other websites that also feature user-generated content to investigate the effects of site structure and norms on language use. Alternatively, communication styles on Scratch, which is largely unconstrained and interest-driven, could be compared to communication on other learning sites that provide much more structure to investigate differences in participation and language use. Future research should be extended to include investigations of collaboration and teamwork using computer-mediated communication among youth, a topic not well suited to Scratch because of the nature of the site, which supports primarily individual program development. While there has been comparable research on language use in online games like World of Warcraft [26] and Everquest [34], these sites are largely for adults and involve different kinds of tasks, though these research contexts provide opportunities to compare communication patterns and norms of youth and adults.

7. Acknowledgements

Blinded for review.

8. References

[1]M. Adkins, D.P. Twitchell, J.K. Burgoon, and J.F. Nunamaker, "Advances in Automated Deception Detection in Text-Based Computer-Mediated Communication", Proceedings of the SPIE Defense and Security Symposium, Orlando, Florida, 2004, pp. 122-129.

[2] S. Amershi, and C. Conati, "Combining unsupervised and supervised classification to build user models for exploratory learning environments", Journal of Educational Data Mining, 2009, pp. 18–71.

[3] R.S. Baker, A. T.Corbett, K.R. Koedinger, and A. Z. Wagner, "Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System", Proceedings of ACM CHI 2004: Computer-Human Interaction, 2004, pp. 383-390.

[4] R.S.J.D. Baker, and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions", Journal of Educational Data Mining, 2009, pp. 3-17.

[5] J. Baldridge, T. Morton, and G. Bierner, "The opennlp maximum entropy package", Technical report, SourceForge, 2002.

[6] M. Berland, T. Martin, T. Benton, P. Ko, and C. Petrick-Smith, "Using learning analytics to understand the learning pathways of novice programmers", Journal of the Learning Sciences, Manuscript in press, 2012.

[7] Bers, M.U.,New media and technology: Youth as content creators. Jossey-Bass/Wiley, San Francisco, 2011.

[8] Black, R.W., Adolescents and online fiction, Peter Lang, New York, 2008.

[9] P. Blikstein, "Using learning analytics to assess students' behavior in open-ended programming tasks" , Proceedings of the Learning Analytics and Knowledge conference (LAK11), 2011, pp. 110-116.

[10] K. Brennan, "Mind the gap: Differences between the aspirational and the actual in an online community of learners", The Future of Learning: Proceedings of the 10th International Conference of the Learning Sciences, ICLS, Sydney, Australia, 2011.

[11] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan,"GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications", Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, Philadelphia, 2002.

[12] D. A. Fields, M. Giang, and Y. B. Kafai, "Understanding collaborative practices in the Scratch online community: Patterns of participation among youth designers", To see the world and a grain of sand: Proceedings of the 10th International Conference of Computer Supported Collaborative Learning, International Society of the Learning Sciences, Madison, WI, 2013.

[13]Gallois C., T. Ogay, and H. Giles, "Communication Accommodation Theory: A look Back and a Look Ahead", In Gudykunst, William B. Theorizing About Intercultural Communication. Sage, Thousand Oaks, CA, 2005, pp. 121–148.

[14] Giles H., and P. Smith, "Accommodation Theory: Optimal Levels of Convergence", In Giles H., R. N. St. Clair, and N. Robert, Language and Social Psychology, Basil Blackwell, Baltimore, 1979.

[15]J.D. Gobert, M. A. Sao Pedro, R.S.J.s. Baker, E. Toto and O. Montalvo, "Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds", Journal of Educational Data Mining, 2012.

[16] S. Grimes, and D. Fields, Kids online: A new research agenda for understanding social networking forums, The Joan Ganz Cooney Center at Sesame Workshop, New York, 2012.

[17] G. Gweon, M. Jain, J. McDonough, B. Raj, and C.B. Rosé, "Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation, To see the world and a grain of sand: Proceedings of the 10th International Conference of Computer Supported Collaborative Learning, International Society of the Learning Sciences, Madison, WI, 2013.

[18] S. B. Heath, "Three's not a crowd: Plans, roles, and focus in the arts", Educational Researcher, 2001, pp. 10-17.

[19] Heath, S. B., "Working through language", In S. Hoyle & C. T. Adger (Eds.), Kids talk: Strategic language use in later childhood, Oxford University Press, New York, 1998.

[20] Y. B. Kafai, D. A. Fields, R. Roque, W.Q. Burke, and A. Monroy-Hernández, "Collaborative agency in youth online and offline creative production in Scratch", Research and Practice in Technology Enhanced Learning, 2012, pp. 63-87.

[21] Y.B. Kafai, and K. A. Peppler, "Youth, Technology, and DIY Developing Participatory Competencies in Creative Media Production", Review of Research in Education, 2011, pp. 89-119.

[22] K. Koedinger, K. Cunningham, A. Skogsholm, and B. Leber, "An open repository and analysis tools for fine-grained, longitudinal learner data", Educational Data Mining, 2008, pp. 157–166.

[23] D. M. Mackie, M. C. Gastardo-Conaco, and J. J. Skelly, "Knowledge of the advocated position and the processing of in-group and out-group persuasive messages", Personal Social Psychology Bulletin, Sage, 1992, pp. 145-151.

[24] J. Maloney, K. Peppler, Y. Kafai, M. Resnick, and N. Rusk, "Programming by Choice. Urban Youth Learning Programming with Scratch", Paper presented at the SIGCSE 2008 Conference, Portland, Oregon, 2008.

[25] A. Lenhart, "Teens and Sexting: How and why minor teens are sending sexually suggestive nude or nearly nude images via text messaging" , Report prepared for the Pew Internet & American Life Project, an initiative of the Pew Research Center, 2009.

[26] B.A. Nardi, S. Ly, and J. Harris, "Learning conversation in World of Warcraft", Proceedings, HICSS, 2007.

[27] National School Boards Association, Creating & connecting: Research and guidelines on online social and educational networking, National School Boards Association, Alexandria, VA, 2007.

[28] M. Resnick, J. Maloney, A.M. Hernández, N. Rusk, N, E. Eastmond, K. Brennan, A.D. Millner, E. Rosenbaum, J. Silver, B. Silverman, and Y.B.Kafai, "Scratch: Programming for everyone", Communications of the ACM, 2009, pp. 60–67.

[29] R. Roque, Y.B. Kafai, and D. A. Fields, "From tools to communities: Designs to support online creative

collaboration in Scratch", Proceedings of IDC 2012, Bremen, Germany, 2012.
[30] B. Sherin, "Computational studies of commonsense science: An exploration in the automated analysis of clinical interview data", In press/under review.

[31] Turner, L. H., and R. West, "Communication Accommodation Theory". Introducing Communication Theory: Analysis and Application (4th ed.), McGraw-Hill, New York, 2010.

[32] D. Twitchell, D.P. Biros, N. Forsgren, J. Burgoon, and J. F. Nunamaker Jr, "Assessing the veracity of criminal and detainee statements: a study of real-world data", International Conference on Intelligence Analysis. 2005.

[33] C. Whissell, M. Fournier, R. Pelland, D. Weir, and K. Makarec, "A dictionary of affect in language: IV. Reliability, validity, and applications", Perceptual and Motor Skills, 1986, pp. 875-888.

[34] D. Williams, N. Contractor, M.S. Poole, J. Srivastava, and D. Cai, "'The Virtual Worlds Exploratorium: Using Large-Scale Data and Computational Techniques for Communication Research", Communication Methods and Measures, 2011, pp. 163 — 180.

[35] Witten, I. H. and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java, Morgan Kaufmann, San Francisco, 2000.

[36] L. Zhou, Q. E. Booker, and D. Zhang, "ROD-toward rapid ontology development for underdeveloped domains", Proceedings of the 35th Annual Hawaii International Conference, IEEE, 2002.

[37] L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell, "Automating Linguistics-Based cues for detecting deception in Text-Based asynchronous Computer-Mediated communications", Group Decision and Negotiation, 2004, pp. 81-106.

[38] L. Zhou, J. K. Burgoon, D. P. Twitchell, T. Qin, and J. F. Nunamaker Jr., "A comparison of classification methods for predicting deception in computer-mediated communication", Journal of Management Information System*s*, 2004, pp. 139-166.