Collaborative Review in Writing Analytics: N-Gram Analysis of Instructor and Student Comments

Alex Rudniy
Fairleigh Dickinson University
1000 River Rd
Teaneck, NJ 07666
1-646-684-5876
arudniy@fdu.edu

Norbert Elliot
New Jersey Institute of Technology
323 Dr. Martin Luther King Jr. Blvd
Newark, NJ 07102
1-856-952-7680
elliot@njit.edu

ABSTRACT
The purpose of this paper is to explore the use of n-gram analysis to analyze instructor and student comments elicited within *My Reviewers*, a web-based learning environment. Shown to be informative in a wide variety of applications, n-gram analysis is of interest in determining concept proliferation in topics, purposes, terminologies, and rubrics used in writing courses. As the present study demonstrates, unigram, bigram, digram, trigram, fourgram, and fivegram analytic methods reveal important information about instructor and student use of concepts; in turn, such analysis holds the potential to lead to precise and actionable revision behaviors.

Keywords
context informed linguistic analysis, *My Reviewers*, n-grams, web-based learning

1. INTRODUCTION
This study extends that of Aull [1] in context-informed corpus linguistics analysis. Defined as an approach that explores discourse "as realizations of socio-rhetorical contexts and as patterns across them," a context-informed approach to corpus analysis yields information that is useful in distinct educational settings (p. 52). This approach identifies linguistic patterns that may prove useful across settings in which commonality of student population and writing tasks are similar.

To extend the approach of Aull, we examined instructor and student comments posted on intermediate drafts within *My Reviewers*, a digital tool developed at University of South Florida (USF) to facilitate document reviews, peer reviews, team projects, and portfolios [2]. Within this web-based learning environment, document markup tools enable instructors to use a rubric to assess the primary coursework, including intermediate and final drafts.

In the present study, attention is given to a single course: English 1102: Rhetoric and Academic Research, a second-semester USF undergraduate writing course [3]. The course introduces students to rhetorical conventions and provides them with an opportunity to analyze, research, and compose arguments. Designed to improve academic writing, research, information literacy, and *critical thinking abilities, the course is unique in its focus on exploring the ways that writers gain agency—that is, credibility through argument, negotiation, and reasoning. In addition, the course incorporates projects using distinct print and digital genres. Because of its uniqueness (focus on writer agency) and variation (use of multiple genres), the course is ideal for exploring the usefulness of n-gram analysis in providing context-specific information regarding course specific information.

To lend specificity to the analysis, this study uses the term *course concept proliferation*. Generally speaking, first-year postsecondary writing courses simultaneously advance knowledge and skills as part of the cognitive domain of the course [4]. For example, the ability to think critically about a specific topic (how writers gain agency through evidence) is demonstrated through mastery of genre (how an essay is organized through claims). Analysis of instructor and student comments affords the analysis of proliferation of key course terms involving instruction and key trait terms involving assessment. As such, *concept proliferation* is defined as the degree to which course terms and assessment traits are present in comments—and what that presence suggests regarding instruction that unifies topic, purpose, terms, and rubrics for the benefit of students.

2. N-GRAM ANALYSIS
Because of its straightforward assumptions, n-gram analysis is ideal for a basic analysis of course concepts students should know and the evaluation of those concepts through rubric use.

2.1 Definition
An N-gram is defined as a sequence of n items as they appear in text—letters, words, phonemes, part-of-speech (POS) tags, or other elements. N in n-gram denotes the number of items in a sequence. Commonly, a single word is referred to as a unigram; two words are referred to as a bigram; three words constitute a trigram; four words constitute a four-gram; and five words constitute a fivegram [5,6].

2.2 Early Work
The history of n-gram model originates in Markov [7, 8]. N-grams are considered a version of the multi-order Markov model in which the probability of the Nth element depends on the previous N-1 elements and can be obtained from data [5]. Shannon [9] and Chomsky [10, 11] are known for applying n-grams for predicting subsequent elements within sequences (e.g., Shannon game) [12]. These elements can vary from a single character to a linguistic entity [8].

In the 1950s, 1960s and 1970s, n-gram models from one to five were used as a stand-alone research method in early works on natural language processing, in particular for hand-printing recognition and standardization, reading machines for the blind, and language computational analysis. Due to computational restrictions of that era, character n-grams were widely used in a large number of studies [13].

2.3 Contemporary N-Gram Applications

Bassil [8] designed an n-gram-based method for spelling corrections and evaluated it on the Yahoo! N-Grams Dataset 2.0 consisting of word n-grams of sizes from 1 to 5 [14]. Nadkarni et al. [5] describe the applications of character n-grams for auto-completion of words and phrases, spelling correction, speech recognition, and word disambiguation on the Google n-gram dataset for n=1..5, which was assembled from web data and the Google Books project [15]. The Google Books N-Gram Corpus is commonly used for analyzing cultural, social, and linguistic trends. It contains n-grams and their frequencies retrieved from books in several languages over the past five hundred years [16, 17]. Mayfield and McNamee [18] applied n-gram tokenization for stemming in a language-independent way. Gencosman et al. [19] describe character n-gram applications in speech recognition, optical character recognition, spelling correction, handwriting recognition, and statistical machine translation. In addition, Lecluze et al. [20] mention examples of character n-gram models for author and language identification, speech analysis, classification of multilingual documents, and information retrieval.

Rangarajan and Ravichandran [21] registered a US patent describing a system and a method for indexing and retrieval of stored documents using n-grams. While working on opinion extraction and classification tasks, Dave et al. [22] identified the n-gram model to be analytically competitive; specifically, trigrams demonstrated the best performance compared to bigrams and unigrams. Their work identified two major flaws related to product reviews: rating inconsistency when qualitative descriptions do not correlate with quantitative scores; and ambivalence and comparison when an overall conclusion contradicts a review body. Zhao [23] concludes that bag-of-n-grams-based methods achieve state-of-the-art results for sentiment classification of long movie reviews. Wang, McCalum and Wei [24] claim the importance of n-grams in multiple areas of NLP and text mining, especially for parsing, machine translation and information retrieval. The work by Bespalov et al. [25] determines that the n-gram model in conjunction with latent semantic analysis produce superior results for document-level classification tasks. N-grams were successfully used by Chaovalit and Zhou [26] for sentiment analysis. Lin and Hovy [27] demonstrated an n-gram-based method for automatic document summarization that outperforms human assessments in certain cases.

Ye et al. [28] have established influential research in data mining and classification, naming n-gram one of three most important approaches in text mining and sentiment classification. The n-gram method is known as the simplest and the most successful method in language modeling [29].

In writing analytics, n-gram models were used as a discriminator of different genres for corpus analysis and register variations [30]. This research domain was expanded by multiple analyses investigating n-grams variations between academic prose and conversation [31]; analysis of frequencies, structural types and functional categories of n-grams in textbooks [32]; student writings in history and biology [33]; L1 and L2 academic writing [34]; and n-gram frequencies in multiple registers [35]. Lately, n-grams are used in the preprocessing and feature-extraction stages while more advanced techniques are applied afterwards [36]. For example, N-gram frequencies serve as feature values used by data mining classification algorithms [6]. Jain et al. [37] applied a Markov model after extracting text features with bi- and tri-grams and their frequencies.

Justeson and Katz [38] used n-gram frequencies to identify technical terms in texts. After sorting by frequency, this method yielded noun phrases that were topically relevant to the documents of their corpus. More recently, Aull [1] has used n-gram analysis to distinguish first-year and expert writing by emphasizing the bigram "I will." Using such phrases, Aull found that expert writers draw attention to their involvement in, and control of, the socio-rhetorical subject matter of the text (e.g., "I will discuss"). In this way, the expert writers demonstrate their "text internal" presence and involvement within the unfolding argument and evidence. In contrast, first-year college writers adopted a more "text external" position in which they established themselves as more of a participant in the "real world" outside of the text (e.g., "I will always remember").

3. RESEARCH QUESTIONS

Baseline and descriptive, this study poses three questions:

1. How can n-gram analysis be used to examine concept proliferation of course terms students should know?
2. How can n-gram analysis be used to examine concept proliferation of assessment traits used to assess student work?
3. What type of n-gram analysis is best suited to examine concept proliferation?

4. METHOD

Instructor and student comments were retrieved from *My Reviewers* for ENC 1102 courses offered during the 2014 and 2015 academic years. The data were anonymized as required by federal regulations.

My Reviewers allows free-response textual comments and designation of numeric score on a 4-point scale employing 5 rubric traits: focus, evidence, organization, style, and format. The same essay draft is reviewed by several fellow students (peer review) and an instructor (expert review). To ensure inter-rater agreement, all comments in which instructor scores did not match peer scores were removed. Ten datasets were then constructed, two—one with instructor and one with peer comments—for each of the 5 rubric traits using intermediate drafts. The dataset is shown in Table 1.

Table 1. Sampling Plan: Datasets for Study

Dataset	Instructor Comments	Peer Comments
Dataset Trait 1. Focus	1,516	1,859
Dataset Trait 2. Evidence	2,976	3,809
Dataset Trait 3. Organization	1,219	1,682
Dataset Trait 4. Style	1,252	1,870
Dataset Trait 5. Format	2,549	4,084

Microsoft SQL Server was used for preparing the datasets. For text preprocessing and n-gram extraction, R, RStudio, and the TM package were employed. Following a common procedure for the pre-processing phase, text was converted to lower case; any non-word characters, numbers, and punctuation were removed. In this study, stemming was not applied since n-grams of word base

forms unnecessarily complicated analysis. Since we do not use any computer algorithms for subsequent text feature comparison, stemming brings extra complexity for interpreting n-grams. In future work, we plan to use stemmed n-grams as a preprocessing step for more sophisticated analysis using LSA. Similarly, adhering to common practice in text mining applications, the corpus was stripped of stop words, though there is evidence this operation may negatively affect results for certain tasks (e.g. plagiarism detection) [39]. Finally, whitespace such as line breaks and tabulation symbols was removed.

The corpus was tokenized into 1-, 2-, 3-, 4- and 5-gram models. N-gram frequencies were obtained with the help of a term-document matrix displaying the frequency of terms occurring in a collection of documents. The obtained models were used to build subsets of the most common n-grams, and n-grams used more than a hundred times per dataset. While analyzing corpus features, n-grams used across criteria by peers, instructors and both instructors and peers were identified.

5. RESULTS

Results will be presented in terms of the study questions. Interpretations will follow each result.

5.1 N-gram Analysis and Course Terms

Table 2 presents ENC 1102 course topics, their purpose, the genres used, and terms that students should know from each project. The dataset shown in Table 1 was assembled from each of the three projects.

Unique in this course is the use of constructed response tasks based on topics uniformly used across course sections. Equally unique is the clearly stated purpose of each topic, the variation in genre across essays, websites, and oral presentations, and identification of key course terms. Using the traits of focus, evidence and organization as sources of information about course knowledge, Table 3 presents a unigram analysis of each of the course projects with attention to terms students should know. Terms not mentioned in comments are listed with zero frequencies. Following each term, the number of instances of each term is used within the 100 most commonly used terms in the comments.

Table 2: Context: English Composition II

Topics	Purpose	Genre	Terms Students Should Know
Project 1: Analyzing Visual Rhetoric	"In Project One, you will learn how to identify one stakeholder's argument and analyze that stakeholder's use of visual and rhetorical strategies."	*Source-based essay*: identify one stakeholder's argument and analyze that stakeholder's use of visual and rhetorical strategies.	stakeholder, rhetorical appeals, ethos, pathos, logos, Kairos, visual rhetoric, visual fallacies
Project 2: Finding Common Ground	"In Project Two, you will learn how to present an unbiased analysis of two arguments created by stakeholders with seemingly incompatible goals about an issue or topic and create a feasible, objective compromise that would benefit both stakeholders."	*Source-based essay*: analyze two stakeholders with seemingly incompatible goals regarding the same issue or topic; identify common ground between stakeholders.	compromise, empathy, negotiation, Rogerian argument
Project 3: Composing Multimodal Assignments	"Project 3 brings all you have done full circle. You will use your understanding of the rhetorical situation to decide how to craft the most effective means of engaging your audience and empowering the audience to take the action you recommend."	*Multimedia Argument Website*: produce a complementary argument using the digital medium of a website to address these aims: educate an audience of non-engaged stakeholders about the issue or topic, engage the audience by convincing them that they should care about this issue or topic, and empower the audience to take action in some way. *Formal Essay*: produce a complimentary essay that addresses the website aims, *Presentation*: present their multimodal remediation (or a portion of it) for an audience of their peers. Individual instructors will dictate the specific requirements of these presentations.	multimodality, remediation, non-engaged stakeholder

Table 3. Unigram Analysis: Terms Students Should Know Used in 100 Most Frequent Comments

Projects 1, 2, and 3	Instructor	Student
Focus	stakeholder (571) rhetorical (454) ethos (0) pathos (0) logos (0) Kairos (0) visual (471) fallacies (0) compromise (603) empathy (0) negotiation (0) Rogerian (0) argument (481) multimodality (0) remediation (0) non-engaged (0)	stakeholder (571) rhetorical (278) ethos (0) pathos (0) logos (0) Kairos (0) visual (210) fallacies (0) compromise (536) empathy (0) negotiation (0) Rogerian (0) argument (331) multimodality (0) remediation (0) non-engaged (0)
Evidence	stakeholder (761) rhetorical (1011) ethos (0) pathos (470) logos (508) Kairos (0) visual (659) fallacies (477) compromise (633) empathy (0) negotiation (0) Rogerian (0) argument (927) multimodality (0) remediation (0) non-engaged (0)	stakeholder (740) rhetorical (502) ethos (0) pathos (0) logos (0) Kairos (0) visual (0) fallacies (0) compromise (436) empathy (0) negotiation (0) Rogerian (0) argument (998) multimodality (0) remediation (0) non-engaged (0)
Organization	stakeholder (223) rhetorical (180) ethos (0) pathos (0) logos (0) Kairos (0) visual (0) fallacies (0) compromise (313) empathy (0) negotiation (0) Rogerian (0) argument (248) multimodality (0) remediation (0) non-engaged (0)	stakeholder (326) rhetorical (234) ethos (0) pathos (0) logos (0) Kairos (0) visual (0) fallacies (0) compromise (306) empathy (0) negotiation (0) Rogerian (0) argument (0) multimodality (0) remediation (0) non-engaged (0)

5.1.1 Course Term Results

Distinct patterns emerge of congruence, disjuncture, and absence in Table 3. There is notable congruence among the terms that both instructors and students use. Regarding the trait of focus, stakeholder, rhetorical, visual, compromise, and argument are used in both instructor and student comments. Regarding the trait of evidence, stakeholder, rhetorical, compromise, and argument are used in both sets of comments. Regarding the trait identified as organization, the terms stakeholder, rhetorical, compromise, and argument are used in both sets of comments. There is also notable disjuncture. In terms of the trait of focus, instructors use the term visual twice as much as students. In terms of evidence, the term rhetorical is used twice more by instructors than by students; as well, while instructors use the term visual, students do not use that term at all. In terms of organization, instructors use the term while students do not. There is a notable absence of key

terms by both groups: ethos, pathos, logos, Kairos, fallacies, empathy, negotiation, Rogerian, multimodality, remediation, and non-engaged.

recurring patterns in writing comments through both the presence and absence of concepts.

Table 4. Rubric Terms: Trait Specifications

	Trait 1: Focus	Trait 2: Evidence	Trait 3: Organization	Trait 4: Style	Trait 5: Format
Terms in Rubric	critical thinking, thesis, ideas, analysis, assignment requirements	critical thinking, credible sources and supporting details, synthesis, visuals, personal experience, anecdotes, writer's idea, source's ideas	critical thinking, introduction, topic sentences, segues, transitions, conclusion	critical thinking, grammar, punctuation, point of view, syntax, diction, word choice, vocabulary	documentation style, MLA, APA, formatting, in-text citations, annotated bibliographies, works cited, document design

5.1.2 Course Term Interpretation

Patterns of congruence reveal that some of the course terms are being used in comments on intermediate drafts by both instructors and students. This pattern is praiseworthy and suggests a common referential frame. However, instructors appear to associate the use of visual artifacts as elements of evidence while students do not. Similarly, terms such as rhetorical are much more commonly used by instructors. In the case of terms from classical rhetoric—ethos, pathos, and logos—there is no use by either group; nor is there use of more contemporary rhetorical systems such as that developed by Carl Rogers [40]. And the presence of logical fallacies is not taken up by either group in the comments. Regarding use of such information, curricular strategies might be taken to ensure continued use of congruent terms, to investigate differing use of terms by instructors and students, and to probe more deeply into which terms are opaque or cosmetic and therefore unlikely to be used to advance student learning.

5.2 N-gram Analysis and Traits

Table 4 presents the 5 assessment traits used in ENC 1102 and their associated rubric terms.

Table 5 presents each of the rubric traits in for instructor comments in terms of unigram, bigram, trigram, fourgram, and fivegram analysis. Table 6 presents the same traits and analysis for student comments.

As is the case in the analysis of course terms, rubric traits also reveal distinct patterns of congruence, disjuncture, and absence.

5.2.1 Rubric Trait Results

Unigram and bigram analyses for instructor and students are largely congruent. For both groups, the presence of a thesis is associated with focus, just as evidence derives from sources, organization is understood as achieved through paragraphs, style is associated with correct grammar, and format is achieved through following specifications established by the Modern Language Association. Absent are terms related to organization. Regarding evidence, trigram analysis reveals some disjuncture. Instructors note that sources establish credibility; students, in contrast, note the presence and features of the works cited page—a format substitution for the complexities of establishing claims. Fourgram analysis reveals the presence of a writer, the innovator Jane Chen, while student comments remain vague in their reference to credible sources. Fivegram analysis continues to reveal specificity in instructor comments regarding evidence while students remain vague in noting that "quotes are really good." In terms of the rubric, absent are references to traits such as synthesis, personal experiences, anecdotes, segues, diction, and document design. Useful, n-gram analysis clearly exposes

5.2.2 Rubric Trait Interpretation

As is the case with course terms, patterns of congruence reveal that some rubric traits are being used in comments on intermediate drafts by both instructors and students. This pattern suggests a common referential frame often lacking across course sections. However, the traits are general and do not seem to accommodate multimodal genres; that is, while paragraphs are central to constructing an academic, source-based essay, the rubric does not address ways to achieve coherence in a website. Furthermore, rubric traits do not address the oral presentation genre associated with Project 3.

It must be noted that genres beyond the essay may not be evaluated within *My Reviewers* if instructors do not require that intermediate drafts be uploaded to the platform for review. This example demonstrates the complexities of capturing all student performance within a digital environment.

5.3 N-gram Analysis and Concept Proliferation

Tables 3, 4, and 5 reveal that various forms of n-gram analysis can be very useful in capturing key course terms and rubric traits as they are used in instructor and student comments. Implying metacognition, review comments suggest a deep and deliberate use of course concepts and evaluative frameworks. N-gram analysis reveals the presence of such words—and the directions that might be taken to examine their usefulness to students and their absence in areas where more specific guidance may be helpful to students.

Where unigrams and bigrams yield larger sample sizes, however, trigrams, fourgrams, and fivegrams reveal extremely small sample sizes. The benefits and costs of these smaller sample sizes, and the inferences drawn from them, should be taken into consideration before their use.

6. FURTHER RESEARCH DIRECTIONS

In her call for context-informed corpus linguistics analysis, Aull [1] has advanced connections between lexical analysis and classroom applications. In such pedagogically-based applications using bigram analysis, Forbes-Riley and Litman [40] have developed approaches for adapting student affect in intelligent tutoring dialogue systems. At the level of the student, this study confirms the possibility of connecting word-level patterns to curricular design. Real-time communication of such information to students and their instructors is the next step in advancing context-informed corpus linguistics analyses that are that are structured and actionable.

Table 5. Rubric Trait Analysis: Instructor Comments

Rubric Traits	Unigram	Bigram	Trigram	Fourgram	Fivegram
Focus	thesis (1015) paper (948) good (855) topic (755) specific (746)	thesis statement (194) good thesis (103) make sure (101) assignment requirements (81) call action (76)	assignment requirements met (32) make thesis specific (29) please write thesis (28) thesis answer question (28) aloud evaluate content (27)	aloud evaluate content word (27) evaluate content word flow (27) read aloud evaluate content (27) arguable thesis proposes compromise (25) build strong specific arguable (25)	aloud, evaluate, content, word, flow (27) read aloud evaluate content word (27) arguable thesis proposes compromise stakeholders (25) build strong specific arguable thesis (25) specific arguable thesis proposes compromise (25)
Evidence	sources (1946) evidence (1817) source (1742) use (1600) sure (1562)	make sure (221) use evidence (211) good use (185) final draft (175) relevant stuff (174)	smart relevant stuff (174) support paper s (110) sources establish credibility (106) introductions sources establish (104) article relevant research (89)	introductions sources establish credibility (104) article relevant research published (89) biochemist Jane Chen discusses (89) credible magazine biochemist jane (89) Jane Chen discusses significance (89)	article relevant research published credible (89) biochemist Jane Chen discusses significance (89) credible magazine biochemist Jane Chen (89) magazine biochemist Jane Chen discusses (89) published credible magazine biochemist jane (89)
Organization	paragraph (917) paragraphs (721) paper (709) topic (652) organization (639)	topic sentences (118) topic sentence (81) make sure (74) papers (56) body paragraphs (47)	thesis form required (37) easily followed writer (35) followed writer audience (35) form required organization (35) organization easily followed (35)	easily followed writer audience (35) followed writer audience reader (35) form required organization easily (35) organization easily followed writer (35) required organization easily followed (35)	easily followed writer audience reader (35) form required organization easily followed (35) organization easily followed writer audience (35) required organization easily followed writer (35) thesis form required organization easily (35)
Style	paper (691) issues (638) grammar (611) use (609) person (581)	word choice (129) sentence structure (100) third person (83) final draft (69) community comments (61)	read paper aloud (39) see notes page (38) person point view (33) use third person (33) continue develop writing (27)	continue develop writing style (27) develop writing style try (25) comments grammar style support (23) community comments grammar style (23) refer community comments grammar (23)	continue develop writing style try (25) community comments grammar style support (23) refer community comments grammar style (23) help strengthen word choice vary (21) revising way will help strengthen (21)
Format	page (1581) MLA (1512) cited (1421) works (1417) citations (1346)	works cited (937) cited page (489) text citations (395) MLA format (222) final draft (197)	works cited page (480) consult MLA style (122) MLA style guide (122) works cited list (110) draft consult MLA (107)	consult MLA style guide (122) draft consult MLA style (107) final draft consult MLA (107) Purdue owl help proper (101) community comments Purdue owl (96)	draft consult MLA style guide (107) final draft consult MLA style (107) comments Purdue owl help proper (95) community comments Purdue owl help (95) consult MLA style guide community (95)

Table 6. Rubric Trait Analysis: Student Comments

Rubric Traits	Unigram	Bigram	Trigram	Fourgram	Fivegram
Focus	paper (1213) thesis (1090) focus (1018) topic (983) good (975)	assignment requirements (195) thesis statement (140) throughout paper (104) focus paper (87) meets assignment (79)	meets assignment requirements (70) assignment requirements thesis (39) met assignment requirements (37) paper meets assignment (35) meet assignment requirements (31)	paper meets assignment requirements (31) meets assignment requirements thesis (20) great job staying topic (8) meet assignment requirements thesis (8) good job staying topic (7)	paper meets assignment requirements thesis (10) ad helps reflect goal message (4) ads reflect stuff touched d never (4) also focused logical manner centered (4) and me people disagree following (4)
Evidence	sources (2458) evidence (2312) paper (2309) good (2044) used (1932)	text citations (365) credible sources (219) make sure (202) sources used (160) throughout paper (155)	works cited page (92) use text citations (47) good use evidence (45) good use sources (37) just make sure (36)	fair selection credible sources (21) credible sources supporting details (17) selection credible sources supporting (10) ideas source s ideas (9) good use text citations (8)	selection credible sources supporting details (10) fair selection credible sources supporting (7) relationship thesis primary secondary sources (7) across backed paper just make (5) also really good quoted gave (5)
Organization	paper (1118) well (980) paragraphs (969) paragraph (958) good (893)	well organized (122) topic sentences (118) logical progression (71) organization paper (67) paper organized (60)	paper well organized (44) paper organized well (23) essay well organized (11) logical progression ideas (11) transitions topic sentences (11)	logical progression supporting points (7) well organized easy follow (7) paper well organized easy (6) essay hard figure rhetorical (5) figure rhetorical appeal addressing (5)	essay hard figure rhetorical appeal (5) hard figure rhetorical appeal addressing (5) paper well organized easy follow (5) parts essay hard figure rhetorical (5) additionally essay nice cohesive flow (4)
Style	paper (1246) grammar (1019) errors (1005) good (943) sentences (910)	word choice (256) grammatical errors (187) point view (187) grammar punctuation (124) make sure (110)	consistent point view (60) can easily fixed (27) grammar punctuation errors (26) person point view (26) point view consistent (26)	third person point view (13) point view throughout paper (11) consistent point view throughout (8) errors can easily fixed (8) grammatical errors throughout paper (8)	addressed three rhetorical appeals one (5) appeals one paragraph piece visual (5) away use words like everyone (5) commas missing stay away use (5) couple time commas missing stay (5)
Format	page (2692) cited (2336) format (2317) paper (2228) works (2133)	MLA format (1199) works cited (1033) cited page (925) text citations (545) make sure (307)	works cited page (745) work cited page (171) name page number (105) followed MLA format (102) last name page (99)	last name page number (86) works cited page needs (52) text citations works cited (51) citations works cited page (48) format works cited page (45)	text citations works cited page (39) MLA format works cited page (25) last name page number top (21) make sure works cited page (20) name page number top right (18)

7. ACKNOWLEDGEMENTS

This research is supported by the National Science Foundation under Award #1544239, "Collaborative Research: The Role of Instructor and Peer Feedback in Improving the Cognitive, Interpersonal, and Intrapersonal Competencies of Student Writers in STEM Courses." We wish to thank the support of our principal investigator, Joseph M. Moxley, as well as our fellow investigators Chris Anson, Christiane J. Donahue, Valerie Ross, and Suzanne T. Lane. We are thankful for the reviews of Laura Aull, Jill Burstein, and Dave Eubanks. Thanks also to Rafael Walker for expert manuscript editing.

8. REFERENCES

[1] Aull, L. (2015). *First-year university writing: A corpus-based study with implications for pedagogy*. London, UK: Palgrave Macmillan.

[2] Dixon, Z., and Moxley, J.M. (2013). Everything is illuminated: What big data can tell us about teacher commentary. *Assessing Writing* 18, 241-256

[3] University of South Florida (2016). First-year composition: University of South Florida. http://hosted.usf.edu/FYC/

[4] National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Committee on Defining Deeper Learning and 21st Century Skills, J.W. Pellegrino and M.L. Hilton, Editors. Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

[5] Nadkarni P.M., Ohno-Machado, L, and Chapman W.W. (2011). Natural language processing: an introduction. *J Am Med Inform Assoc*. 18, 544-551.

[6] Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., Chanona-Hernández, L. (2014). Syntactic N-grams as machine learning features for natural language processing. *Expert Systems with Applications* 41, 853–860.

[7] Markov, A.A., (1913). Essai d"une recherche statistique sur le texte du roman "Eugène Oneguine", *Bull. Acad. Imper. Sci*. 7, 153-162.

[8] Bassil, Y. (2012). Parallel spell-checking algorithm based on Yahoo! N-grams dataset. *International Journal of Research and Reviews in Computer Science* 3, 1429-1435.

[9] Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27, 379-423.

[10] Chomsky, N. (1957). *Syntactic structures.* Mouton: The Hague.

[11] Chomsky, N. (1956). Three models for the description of language. *IRI Transactions on Information Theory* 2, 113-124.

[12] Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journa*l 30, 50–64.

[13] Suen, C.N. (1979). N-Gram Statistics for Natural Language Understanding and Text Processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1,164-172.

[14] Yahoo! Webscope dataset Yahoo! N-Grams, ver. 2.0, http://research.yahoo.com/Academic_Relations

[15] Franz, A., and Brants. T. (2006). Google N-gram database (all our N-grams belong to you). http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html"

[16] Michel, J.B., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science* 331, 176-182.

[17] Kulkarni V. et al. (2013). Statistically significant detection of linguistic change. Proceedings of the 24th International Conference on World Wide Web 625-635.

[18] Mayfield, J., and McNamee P. (2003, July 28–August 1). Single n-gram Stemming. SIGIR'03, 415-416.

[19] Gencosman, B.C., et al. (2014). Character n-gram application for automatic new topic identification. *Information Processing and Management* 50, 821–856.

[20] Lecluze, C, et al. (2013). Which granularity to bootstrap a multilingual method of document alignment: Character n-grams or word n-grams? *Procedia - Social and Behavioral Sciences* 95, 473 – 481.

[21] Rangarajan, V, and Ravichandran, N. (1998, Jan. 6). System and method for portable document indexing using n-gram word decomposition. U.S. Patent.

[22] Dave K, et al. (2003, May 3-4). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. WWW2003, Budapest, Hungary, 519-528.

[23] Zhao, Z. (2016).Learning document embeddings by predicting n-grams for sentiment classification of long movie reviews. Accepted as a workshop contribution, ICLR.

[24] Wang, X., McCallum, A., and Wei, X., Topical n-grams: Phrase and topic discovery, with an application to information retrieval, Proceedings of the 7th IEEE International Conference on Data Mining, 697-702.

[25] Bespalov D, et al. (2011, October 24–28). Sentiment classification based on supervised latent n-gram analysis. CIKM'11, Glasgow, Scotland, 375-382

[26] Chaovalit, P., and Zhou, L. (2005). Movie review mining: A comparison between supervised and unsupervised classification approaches, HICSS, 2005, Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Proceedings of the 38th Annual Hawaii International Conference on System Sciences.

[27] Lin, C-Y, and Hovy, E. Automatic evaluation of summaries using n-gram Co-Occurrence Statistics . Proceedings of HLT-NAACL 2003, 71-78.

[28] Ye Q, Zhang A. and Law R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications* 36, 6527–6535.

[29] Huang, X., Peng, F., An, A., Shuurmans, D., and Cercone, N. (2003). Applying machine learning to text segmentation for information retrieval. *Information Retrieval* 6, 333–362.

[30] Tang, X., and Cao, J. (2015). *Procedia - Social and Behavioral Sciences* 198, 474 – 478.

[31] Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., and Quirk, R. (1999). L*ongman grammar of spoken and written English*. London/New York: Longman.

[32] Biber, D., et al. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25, 371-405.

[33] Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23, 97-423.

[34] Chen, Y. H., and Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. Language, Learning and Technology 14, 30-49.

[35] Gries, S. T. (2010). Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora. Proceedings of Corpus Linguistics 2009, University of Liverpool.

[36] Nassirtoussi, et al. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications* 41, 7653–7670.

[37] Jaina, K, et al. (2015). Chunked n-grams for sentence validation. *Procedia - Computer Science* 57 209 – 213.1

[38] Justeson, J. S., and Katz, S.M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1, 9-27.

[39] Stamatos, E. (2011, Oct. 24-28). Plagiarism detection based on structural information. CIKM'11. Glasgow, Scotland.

[40] Hairston, M. (1976). Carl Rogers's alternative to traditional rhetoric. *College Composition and Communication*, 27, 373-377.

[41] Kate Forbes-Riley and Diane J. Litman. Using bigrams to identify relationships between student certainness states and tutor responses in a spoken Dialogue Corpus. Proceedings of 6th SIGdial Workshop on Discourse and Dialogue, Portugal.