2017

Applying Natural Language Processing Tools to a Student Academic Writing Corpus: How Large are Disciplinary Differences Across Science and Engineering Fields?

Scott A. Crossley
Georgia State University

David R. Russell
Iowa State University, drrussel@iastate.edu

Kristopher Kyle
University of Hawaii

Ute Romer
Georgia State University

Applying Natural Language Processing Tools to a Student Academic Writing Corpus: How Large are Disciplinary Differences Across Science and Engineering Fields?

Abstract

• **Background:** Researchers have been working towards better understanding differences in professional disciplinary writing (e.g., Ewer & Latorre, 1969; Hu & Cao, 2015; Hyland, 2002; Hyland & Tse, 2007) for decades. Recently, research has taken important steps towards understanding disciplinary variation in student writing. Much of this research is corpus-based and focuses on lexico-grammatical features in student writing as captured in the British Academic Written English (BAWE) corpus and the Michigan Corpus of Upper-level Student Papers (MICUSP). The present study extends this work by analyzing lexical and cohesion differences among disciplines in MICUSP. Critically, we analyze not only linguistic differences in macro-disciplines (science and engineering), but also in micro-disciplines within these macro-disciplines (biology, physics, industrial engineering, and mechanical engineering).

• **Literature Review:** Hardy and Römer (2013) used a multidimensional analysis to investigate linguistic differences across four macro-disciplines represented in MICUSP. Durrant (2014, in press) analyzed vocabulary in texts produced by student writers in the BAWE corpus by discipline and level (year) and disciplinary differences in lexical bundles. Ward (2007) examined lexical differences within micro-disciplines of a single discipline.

• **Research Questions:** The research questions that guide this study are as follows:

1. Are there significant lexical and cohesive differences between science and engineering student writing? 2. Are there significant lexical and cohesive differences between micro-disciplines within science and engineering student writing?

• **Research Methodology:** To address the research questions, student-produced science and engineering texts from MICUSP were analyzed with regard to lexical sophistication and textual features of cohesion. Specifically, 22 indices of lexical sophistication calculated by the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015) and 38 cohesion indices calculated by the Tool for the Automatic Analysis of Cohesion (TAACO; Crossley, Kyle, & McNamara, 2016) were used. These features were then compared both across science and engineering texts (addressing Research Question 1) and across micro-disciplines within science and engineering (biology and physics, industrial and mechanical engineering) using discriminate function analyses (DFA).

• **Results:** The DFAs revealed significant linguistic differences, not only between student writing in the two macro-disciplines but also between the micro-disciplines. Differences in classification accuracy based on students' years of study hovered at about 10%. An analysis of accuracies of classification by paper type found they were similar for larger and smaller sample sizes, providing some indication that paper type was not a confounding variable in classification accuracy.

• **Discussion:** The findings provide strong support that macro-disciplinary and micro-disciplinary differences exist in student writing in these MICUSP samples and that these differences are likely not related to student level or paper type. These findings have important implications for understanding disciplinary differences. First, they confirm previous research that found the vocabulary used by different macro-disciplines to be “strikingly diverse” (Durrant, 2015), but they also show a remarkable diversity of cohesion features. The findings suggest that the common understanding of the STEM disciplines as “close” bears reconsideration in linguistic terms. Second, the lexical and cohesion differences between micro-disciplines are large enough and consistent enough to suggest that each micro-discipline can be thought of as containing a unique linguistic profile of features. Third, the differences discerned in the NLP analysis are evident at least as early as the final year of undergraduate study, suggesting that students at this level already have a solid understanding of the conventions of the disciplines of which they are aspiring to be members. Moreover, the differences are relatively homogeneous across levels, which confirms findings by Durrant (2015) but, importantly, extends these findings to include cohesion markers.

• **Conclusions:** The findings from this study provide evidence that macro-disciplinary and micro-disciplinary differences at the linguistic level exist in student writing, not only in lexical use but also in text cohesion. A number of pedagogical applications of writing analytics are proposed based on the reported findings from TAALES and TAACO. Further studies using different corpora (e.g., BAWE) or purpose assembled corpora are suggested to address limitations in the size and range of text types found within MICUSP. This study also points the way toward studies of disciplinary differences using NLP approaches that capture data which goes beyond the lexical and cohesive features of text, including the use of part-of-speech tags, syntactic parsing, indices related to syntactic complexity and similarity, rhetorical features, or more advanced cohesion metrics (latent semantic analysis, latent Dirichlet allocation, Word2Vec approaches).

Keywords
corpus linguistics, disciplinary differences, natural language processing, STEM Writing, writing analytics

Disciplines
Language and Literacy Education | Science and Mathematics Education

Comments
This article is published as Crossley, Scott, David Russell, Kristopher Kyle, and Ute Römer. "Applying Natural Language Processing Tools to a Student Academic Writing Corpus: How Large are Disciplinary Differences Across Science and Engineering Fields?." *Journal of Writing Analytics* 1 (2017).

Research Article

Applying Natural Language Processing Tools to a Student Academic Writing Corpus: How Large are Disciplinary Differences Across Science and Engineering Fields?

Scott A. Crossley, *Georgia State University*
David R. Russell, *Iowa State University*
Kristopher Kyle, *University of Hawai'i*
Ute Römer, *Georgia State University*

Structured Abstract

- **Background**: Researchers have been working towards better understanding differences in professional disciplinary writing (e.g., Ewer & Latorre, 1969; Hu & Cao, 2015; Hyland, 2002; Hyland & Tse, 2007) for decades. Recently, research has taken important steps towards understanding disciplinary variation in student writing. Much of this research is corpus-based and focuses on lexico-grammatical features in student writing as captured in the British Academic Written English (BAWE) corpus and the Michigan Corpus of Upper-level Student Papers (MICUSP). The present study extends this work by analyzing lexical and cohesion differences among disciplines in MICUSP. Critically, we analyze not only linguistic differences in macro-disciplines (science and engineering), but also in micro-disciplines within these macro-disciplines (biology, physics, industrial engineering, and mechanical engineering).

- **Literature Review**: Hardy and Römer (2013) used a multidimensional analysis to investigate linguistic differences

across four macro-disciplines represented in MICUSP. Durrant (2014, in press) analyzed vocabulary in texts produced by student writers in the BAWE corpus by discipline and level (year) and disciplinary differences in lexical bundles. Ward (2007) examined lexical differences within micro-disciplines of a single discipline.

- **Research Questions**: The research questions that guide this study are as follows:
  1. Are there significant lexical and cohesive differences between science and engineering student writing?
  2. Are there significant lexical and cohesive differences between micro-disciplines within science and engineering student writing?

- **Research Methodology**: To address the research questions, student-produced science and engineering texts from MICUSP were analyzed with regard to lexical sophistication and textual features of cohesion. Specifically, 22 indices of lexical sophistication calculated by the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015) and 38 cohesion indices calculated by the Tool for the Automatic Analysis of Cohesion (TAACO; Crossley, Kyle, & McNamara, 2016) were used. These features were then compared both across science and engineering texts (addressing Research Question 1) and across micro-disciplines within science and engineering (biology and physics, industrial and mechanical engineering) using discriminate function analyses (DFA).

- **Results**: The DFAs revealed significant linguistic differences, not only between student writing in the two macro-disciplines but also between the micro-disciplines. Differences in classification accuracy based on students' years of study hovered at about 10%. An analysis of accuracies of classification by paper type found they were similar for larger and smaller sample sizes, providing some indication that paper type was not a confounding variable in classification accuracy.

- **Discussion**: The findings provide strong support that macro-disciplinary and micro-disciplinary differences exist in student writing in these MICUSP samples and that these differences are likely not related to student level or paper type. These findings have important implications for understanding disciplinary

differences. First, they confirm previous research that found the vocabulary used by different macro-disciplines to be “strikingly diverse” (Durrant, 2015), but they also show a remarkable diversity of cohesion features. The findings suggest that the common understanding of the STEM disciplines as “close” bears reconsideration in linguistic terms. Second, the lexical and cohesion differences between micro-disciplines are large enough and consistent enough to suggest that each micro-discipline can be thought of as containing a unique linguistic profile of features. Third, the differences discerned in the NLP analysis are evident at least as early as the final year of undergraduate study, suggesting that students at this level already have a solid understanding of the conventions of the disciplines of which they are aspiring to be members. Moreover, the differences are relatively homogeneous across levels, which confirms findings by Durrant (2015) but, importantly, extends these findings to include cohesion markers.

- **Conclusions**: The findings from this study provide evidence that macro-disciplinary and micro-disciplinary differences at the linguistic level exist in student writing, not only in lexical use but also in text cohesion. A number of pedagogical applications of writing analytics are proposed based on the reported findings from TAALES and TAACO. Further studies using different corpora (e.g., BAWE) or purpose assembled corpora are suggested to address limitations in the size and range of text types found within MICUSP. This study also points the way toward studies of disciplinary differences using NLP approaches that capture data which goes beyond the lexical and cohesive features of text, including the use of part-of-speech tags, syntactic parsing, indices related to syntactic complexity and similarity, rhetorical features, or more advanced cohesion metrics (latent semantic analysis, latent Dirichlet allocation, Word2Vec approaches).

Keywords: corpus linguistics, disciplinary differences, natural language processing, STEM writing, writing analytics

1.0 Background

An early notion that underlined English for Academic Purposes (EAP) research was that a specific academic discipline (e.g., mechanical engineering or biology) could generally be associated with a narrow range of linguistic choices (e.g., grammatical and lexical choices). A narrow range of features associated with a discipline would support the idea of discipline homogeneity, which could transfer into pedagogical interventions based on detailed knowledge of that academic discipline. However, in practice, early studies did not report strong differences between academic disciplines beyond specific grammatical features, such as passives or conditionals (Ewer & Latorre, 1969).

Over the past several decades, as corpora of academic writing began to be considered and as research moved beyond grammar into discourse, lexical, and rhetorical patterns, numerous important disciplinary differences in academic discourse in English began to emerge. As a result, using linguistic features as quantified in text corpora to explore disciplinary differences in academic texts has become an important part of EAP research (Durrant, 2014, in press; Hyland, 2002; Hyland & Tse, 2007). This large body of research on disciplinary differences has shown that differences often extend to micro-disciplines (i.e., disciplines within disciplines), often in dramatic ways. For instance, Ozturk (2007) studied differences between the move structure in published research article introductions within the micro-disciplines of second language acquisition and second language writing. The two micro-disciplines seemed to display different and almost unrelated move structures, which he suggested reflected the difference between established and emerging fields. In a similar fashion, Hu and Cao (2015) reported large differences in metadiscourse use between qualitative and quantitative research paradigms in published papers from three social science micro-disciplines (education, applied linguistics, and psychology).

However, when compared to professional writing, disciplinary differences in student writing corpora have not received similar attention and scope. Apart from comparisons of student and professional academic writing (Cortes, 2004; Hyland, 2008) and speaking (Biber et al., 2004), there has been relatively little research on disciplinary variation in student texts. Recent studies have started to address this gap and have taken important steps towards better understanding student writing through corpus-based analyses based on corpora such as the British Academic Written English (BAWE) corpus and the Michigan Corpus of Upper-level Student Papers (MICUSP; Durrant, 2014, in press; Hardy & Römer, 2013; Nesi & Gardner, 2012). This line of research has extended to the analysis of differences in grammatical and lexical features of student texts across different disciplines.

2.0 Literature Review

Hardy and Römer (2013) used a multidimensional analysis to investigate linguistic differences across four general academic divisions represented in MICUSP: humanities and arts, social sciences, biological and health sciences, and physical sciences. In this study, MICUSP texts were analyzed using the Biber Tagger, which assigns grammatical and syntactic tags to words and phrases. The tagger also includes semantic markers and some local cohesion features. Using output from the tagger, Hardy and Römer (2013) found that the four academic divisions and the disciplines within those divisions varied linguistically in a number of different ways. Student papers written in philosophy and education courses tended to be more involved (e.g., included many verbs and first and second person pronouns). In contrast, student papers written in physics and biology courses tended to be more informationally dense (e.g., included nominalizations, attributive adjectives, and relatively long words).

Durrant (2014) analyzed texts produced by student writers for the 86 “discipline levels” contained in the BAWE corpus. The discipline levels consisted of combinations of disciplines (e.g., agriculture, business, mathematics) and four student levels. He created discipline-specific frequency word lists for each level[1] and examined the extent to which words were shared across the disciplines. Durrant found that only about 50% of the words used were generic, indicating the other 50% were discipline-specific. An analysis of how discipline-specific words grouped together found that various levels of each discipline clustered based on vocabulary, indicating that discipline-specific vocabulary is not very diverse, although there were a few exceptions. Most of these exceptions were at the post-graduate level indicating that, in some cases, post-graduate writing diverges lexically from undergraduate writing. However, overall, Durrant reported that students in different disciplines of the same level were homogenous in their vocabulary use.

In a second study, Durrant (in press) analyzed disciplinary differences in lexical bundles (i.e., four-word sequences or quad-grams) in the BAWE corpus. Durrant examined 285 authors in 24 different disciplines using discipline-specific quad-gram frequency lists for each level.[2] Comparing overlap between quad-gram use between writers across the queried disciplines within the BAWE corpus, Durrant found that almost all disciplines showed a higher level of overlap internally when compared with external disciplines, leading Durrant to claim that there was a high degree of homogeneity within disciplines. Durrant also reported greater homogeneity within some disciplines (e.g., physics, law, and economics) as compared to other disciplines (e.g., biological sciences, sociology, English). A follow-up analysis revealed differences in vocabulary use between soft sciences (e.g., law, English, classics) and hard sciences (engineering, chemistry, biological sciences).

[1] The frequency lists were specific to the BAWE corpus and were not based on reference corpus (i.e., they were domain dependent). The frequency lists were also not based on lemmas.

[2] Like Durrant (2014), the frequency lists were specific to the BAWE corpus and were not based on a reference corpus. The frequency lists were also not based on lemmas.

Ward (2007) examined lexical differences within micro-disciplines of a single discipline. He conducted a corpus study of collocations in textbooks from five engineering micro-disciplines and reported large differences among the micro-disciplines, raising the question of whether or not there is a common engineering vocabulary. His findings allowed him to suggest collocations appropriate to each micro-discipline as a basis for teaching, in a manner similar to Grabowski (2015), who examined key words and n-grams specific to pharmaceutical discourse.

3.0 Research Questions

The present study extends previous analyses that have focused on student texts and differences in macro- and micro-disciplines. In contrast to previous studies, we use natural language processing (NLP) tools which allow us to examine not only lexical sophistication, which has been shown to be an important indicator of academic writing (Coxhead, 2000), but also text cohesion, which is an important component of larger discourse structures (McNamara, Kintsch, Songer, & Kintsch, 1996). Its inclusion addresses Flowerdew's (2014) call to include linguistic features that go beyond lexis. Critically, we use corpora and NLP tools to not only analyze linguistic differences in macro-disciplines (science and engineering), but also in micro-disciplines within these macro-disciplines (biology, physics, industrial engineering, and mechanical engineering). Our goal is to examine if differences exist at both the macro- and micro-discipline level in a corpus of student writing.

The research questions that guide this study are as follows:

1. Are there significant lexical and cohesive differences between science and engineering student writing?
2. Are there significant lexical and cohesive differences between micro-disciplines within science and engineering student writing?

4.0 Research Methodology

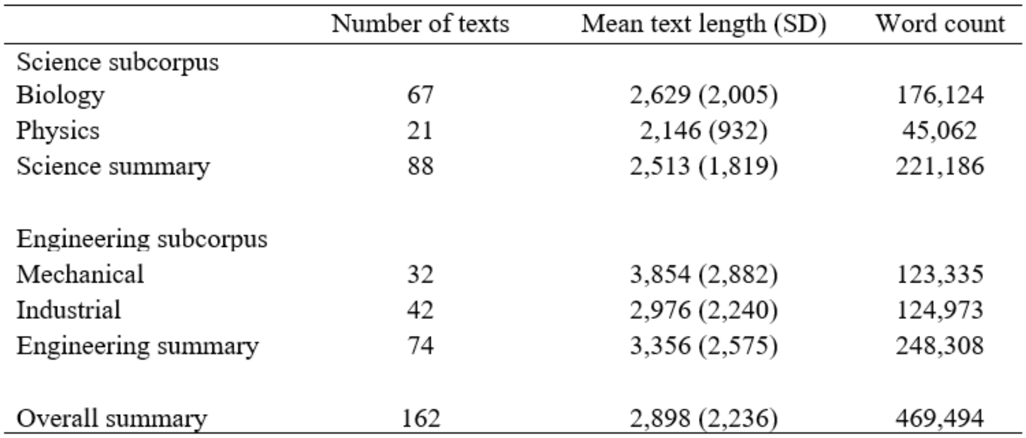
4.1 Corpus

For this analysis, we relied on the Michigan Corpus of Upper-level Student Papers (MICUSP; O'Donnell & Römer, 2012; Römer & O'Donnell, 2011). MICUSP is a corpus of proficient student academic writing samples collected at the University of Michigan. It consists of 829 A-graded papers, making up about 2.6 million words, submitted by students (both native and non-native speakers) from disciplines across four advanced levels of study: final year undergraduates, and first-, second-, and third-year graduate students. Writing samples come from sixteen different disciplines: biology, civil and environmental engineering, economics, education, English, history and classical studies, industrial and operations engineering, linguistics, mechanical engineering, natural resources and environment, nursing, philosophy, physics, political science, psychology, and sociology. Papers span a range of text types, including argumentative essay, creative writing, critique, report, research paper, research proposal, and response paper (see also Ädel & Römer, 2012).

From MICUSP, we selected science writing samples from biology (BIO) and physics (PHY) and engineering writing samples from mechanical engineering (MEC) and industrial and operations engineering (IOE). We selected these four disciplines because they represent two distinct areas of STEM research, natural sciences and engineering, and because, within each area, they provide clear distinctions between macro- and micro-disciplines. These four disciplines (BIO, PHY, MEC, and IOE) make up a MICUSP subcorpus of 162 papers and roughly 470,000 words. Table 1 provides an overview of our MICUSP science and engineering subcorpus and reports, for each selected discipline, the number of papers, word counts, and average text length (with standard deviation). Table 2 shows how the 162 MICUSP papers included in our analysis are distributed across paper types. As we would expect for science and engineering disciplines, the most common paper types students were asked to produce were report (69 of 162 texts) and research paper (63 of 162 texts).

Table 1

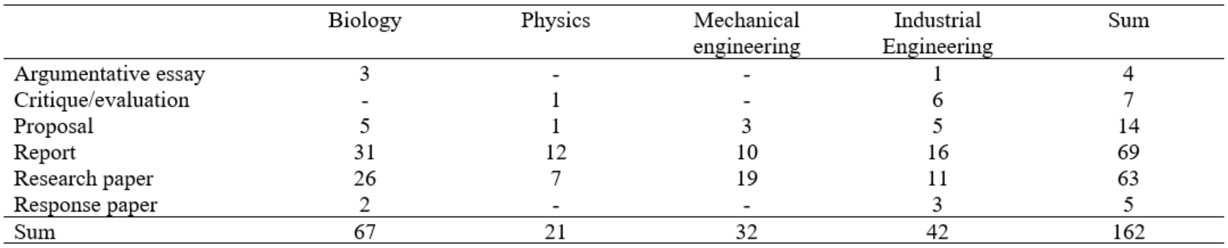
Details of the MICUSP Subcorpora Used in this Study



	Number of texts	Mean text length (SD)	Word count
Science subcorpus			
Biology	67	2,629 (2,005)	176,124
Physics	21	2,146 (932)	45,062
Science summary	88	2,513 (1,819)	221,186
Engineering subcorpus			
Mechanical	32	3,854 (2,882)	123,335
Industrial	42	2,976 (2,240)	124,973
Engineering summary	74	3,356 (2,575)	248,308
Overall summary	162	2,898 (2,236)	469,494

Table 2

Distribution of Papers Across Paper Types in MICUSP Subcorpus



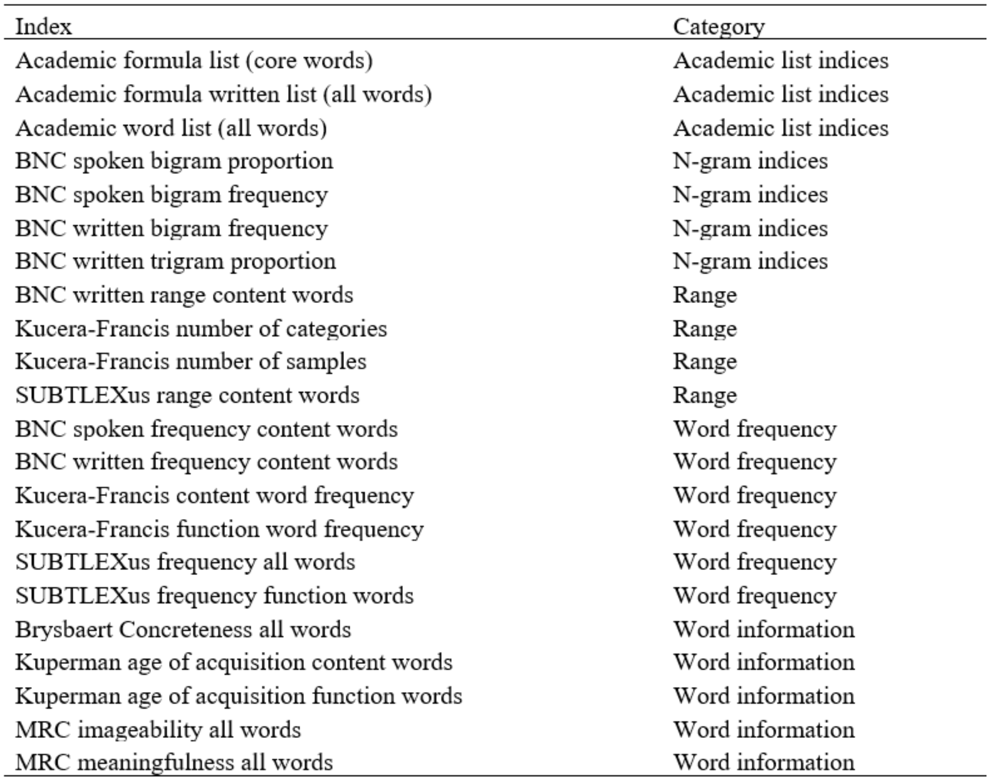
	Biology	Physics	Mechanical engineering	Industrial Engineering	Sum
Argumentative essay	3	-	-	1	4
Critique/evaluation	-	1	-	6	7
Proposal	5	1	3	5	14
Report	31	12	10	16	69
Research paper	26	7	19	11	63
Response paper	2	-	-	3	5
Sum	67	21	32	42	162

4.2 Analysis of Lexical Features

We used the Tool for the Automatic Assessment of Lexical Sophistication (TAALES; Kyle & Crossley, 2015) to assess lexical features. TAALES reports on 136 indices of lexical sophistication. In practice, though, most of these variables are extremely similar and differ only in the databases they pull from or the manner in which the indices are calculated. For instance, TAALES calculates 36 indices that measure word frequency. The frequency indices are derived from a number of different resources (i.e., the British National Corpus [BNC], the Brown Corpus, Kucera-Francis norm, Thorndike-Lorge norms, and the SUBTLEXus database). Each of these resources differs in terms of the corpora from which the frequency counts were derived. For instance, the BNC corpus is comprised of 100 million words of written (90 million words) and spoken (10 million words) English from Great Britain, while the Thorndike-Lorge frequency counts are based on Lorge's 4.5 million-word corpus of popular magazine articles compiled in the 1940s. In practice, the 36 frequency indices calculated by TAALES report on features that are construct-similar, and selecting all the frequency indices would lead to statistical and theoretical redundancy in any developed models. For these reasons, we pre-selected 22 indices from TAALES, focusing on five areas of lexical sophistication discussed briefly below. These areas were lexical frequency, range, n-gram frequency, academic vocabulary, and psycholinguistic word properties. All of these indices are domain independent (i.e., they are not based on data from MICUSP). We discuss these indices briefly below and provide an overview of the selected indices in Table 3. We refer the reader to Kyle and Crossley (2015) for further detail on the tool and how the indices are calculated.

Table 3

Selected TAALES Indices



Index	Category
Academic formula list (core words)	Academic list indices
Academic formula written list (all words)	Academic list indices
Academic word list (all words)	Academic list indices
BNC spoken bigram proportion	N-gram indices
BNC spoken bigram frequency	N-gram indices
BNC written bigram frequency	N-gram indices
BNC written trigram proportion	N-gram indices
BNC written range content words	Range
Kucera-Francis number of categories	Range
Kucera-Francis number of samples	Range
SUBTLEXus range content words	Range
BNC spoken frequency content words	Word frequency
BNC written frequency content words	Word frequency
Kucera-Francis content word frequency	Word frequency
Kucera-Francis function word frequency	Word frequency
SUBTLEXus frequency all words	Word frequency
SUBTLEXus frequency function words	Word frequency
Brysbaert Concreteness all words	Word information
Kuperman age of acquisition content words	Word information
Kuperman age of acquisition function words	Word information
MRC imageability all words	Word information
MRC meaningfulness all words	Word information

4.2.1 Word frequency and range indices. Words that are more frequent in natural language data are learned earlier and used more often than words that are less frequent in natural language data. Frequency has been shown to affect lexical decision times (Kuperman et al., 2012) such that high frequency words are processed more quickly than low frequency words. TAALES calculates frequency scores for all words, content words, and function words. TAALES also provides logarithmic transformations for each of these indices to control for Zipfian effects (Zipf, 1935), which are common in word frequency lists. TAALES computes indices for the following frequency lists: Thorndike-Lorge (Thorndike & Lorge, 1944), Kucera-Francis written frequency (Kucera & Francis, 1967), Brown verbal frequency (Brown, 1984), the British National Corpus (BNC; 2007), and SUBTLEXus (Brysbaert & New, 2009).

In addition to frequency information, TAALES includes a number of range indices, which account for how widely a word or word lemma is used, usually by providing a count of the number of documents in which that word occurs. TAALES calculates range indices for the spoken (915 texts) and written

(3,209 texts) subsets of the BNC, SUBTLEXus (8,388 texts), and Kucera-Francis (500 texts) corpora. TAALES also includes a range count based on Kucera & Francis' (1967) 15 text categories, which can be roughly described as genres.

4.2.2 N-gram indices. N-grams, as compared to single words, measure lexical chunks, common word combinations, and both syntagmatic and paradigmatic knowledge (Crossley, Cai, & McNamara, 2012). TAALES calculates n-gram indices based on bigram (e.g., *there is*) and trigram (i.e., *there is a*) frequencies from both written (90 million words) and spoken (10 million words) subsections of the BNC. In total, TAALES calculates five types of n-gram indices: non-normalized logarithm-transformed frequency counts, n-gram frequency by number of n-grams, n-gram frequency by number of words, the number of unique bi-grams and tri-grams per text, and n-gram proportion scores (by dividing the number of unique bigrams/trigrams in the text that are represented in the reference corpus by the number of words in the text).

4.2.3 Academic list indices. Academic word and formula lists are comprised of words and formulas that occur relatively infrequently in general language corpora, but occur frequently in academic texts (e.g., *analyze, method, reject*). These word lists have been shown to be important indicators of academic writing (Coxhead, 2000; Simpson-Vlach & Ellis, 2010). Academic list indices in TAALES are calculated based on the Academic Word List (AWL; Coxhead, 2000) and the Academic Formulas List (AFL; Simpson-Vlach & Ellis, 2010).

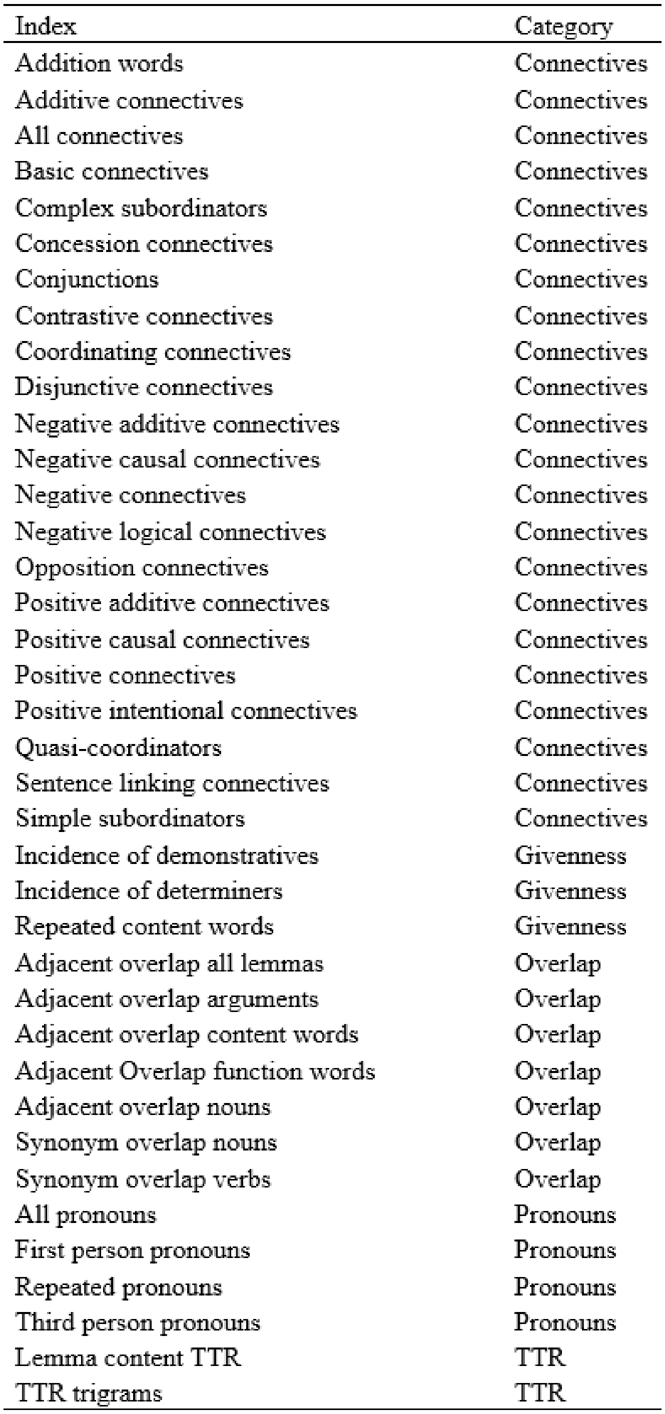
4.2.4 Word information indices. Word information indices measure psycholinguistic properties of words that can explain the variance in lexical decision times (e.g., Kuperman et al., 2012), lexical proficiency (e.g., Crossley et al., 2011a) and speaking proficiency (e.g., Crossley & McNamara, 2013). TAALES reports a number of word information scores that are derived from the MRC Psycholinguistic Database (Coltheart, 1981), Brysbaert, Warriner, & Kuperman (2014), and Kuperman, Stadthagen-Gonzales, Brysbaert (2012). Word information indices are calculated for all words (AW), content words (CW), and function words (FW). Word information indices were calculated from the following lists: familiarity (i.e., how familiar a word is; Coltheart, 1981), concreteness (i.e., how concrete a word is; Brysbaert et al., 2013; Coltheart, 1981), imageability (i.e., how imageable a word is; Coltheart, 1981), meaningfulness (i.e., how many associations a word has; Toglia & Battig, 1978), and age of acquisition (i.e., at what age is a word learned; Kuperman et al., 2012).

4.3 Analysis of Cohesion Features

We used the Tool for the Automatic Assessment of Cohesion (TAACO; Crossley, Kyle, & McNamara, in press) to assess text cohesion. TAACO reports on 146 indices of text cohesion. Like TAALES, though, most of these variables are extremely similar. For these reasons, we selected 38 indices from TAACO related to local, global, and text cohesion meant to measure text coherence. Cohesion is defined as the presence or absence of explicit cues in the text that allow the reader to make connections between the ideas in the text, while coherence is defined as the understanding that the reader derives from the text (McNamara et al., 1996). As with TAALES, there is overlap in the indices that TAACO calculates, leading to possible redundancies in the data. For instance, TAACO calculates 136 indices of lexical overlap at the sentence level as a result of features combinations that include measurements for all words, content words, function words, lemmas, part of speech tags (noun, verb, adjective, adverb, and pronoun), binary overlap, all overlap, and average overlap. To avoid redundancy, we pre-selected 38 indices for analysis (see Table 4 for selected indices). These indices included type-token ratios, sentence overlap, paragraph overlap, semantic overlap (both sentence and paragraph), givenness, and connectives. These are discussed briefly below. We refer the reader to Crossley, Kyle, and McNamara (2016) for further detail on the tool and the indices calculated.

Table 4

Selected TAACO Indices



Index	Category
Addition words	Connectives
Additive connectives	Connectives
All connectives	Connectives
Basic connectives	Connectives
Complex subordinators	Connectives
Concession connectives	Connectives
Conjunctions	Connectives
Contrastive connectives	Connectives
Coordinating connectives	Connectives
Disjunctive connectives	Connectives
Negative additive connectives	Connectives
Negative causal connectives	Connectives
Negative connectives	Connectives
Negative logical connectives	Connectives
Opposition connectives	Connectives
Positive additive connectives	Connectives
Positive causal connectives	Connectives
Positive connectives	Connectives
Positive intentional connectives	Connectives
Quasi-coordinators	Connectives
Sentence linking connectives	Connectives
Simple subordinators	Connectives
Incidence of demonstratives	Givenness
Incidence of determiners	Givenness
Repeated content words	Givenness
Adjacent overlap all lemmas	Overlap
Adjacent overlap arguments	Overlap
Adjacent overlap content words	Overlap
Adjacent Overlap function words	Overlap
Adjacent overlap nouns	Overlap
Synonym overlap nouns	Overlap
Synonym overlap verbs	Overlap
All pronouns	Pronouns
First person pronouns	Pronouns
Repeated pronouns	Pronouns
Third person pronouns	Pronouns
Lemma content TTR	TTR
TTR trigrams	TTR

4.3.1 Type-token ratio (TTR). TTR indices measure word repetition across text. TTR indices have demonstrated positive relations with measures of cohesion in previous studies (McCarthy & Jarvis, 2010) demonstrating that the texts with lower TTR values (i.e., more repetition) are more cohesive. However, TTR indices generally demonstrate negative relations with measures of text coherence (Crossley et al., in press). TAACO calculates a number of different TTR indices. These include simple TTR (the ratio of types to tokens), content word TTR (TTR using only content words such as nouns, verbs, adjectives, and adverbs), function word TTR (TTR using only function words such as pronouns, prepositions, and determiners), lemma TTR (TTR using word lemmas), content lemma TTR, and function lemma TTR. In addition to traditional word-based TTR indices, TAACO also calculates TTR for bigrams (TTR using the number of bigram types over the number of bigram tokens) and for trigrams (TTR using the number of trigram types over the number of trigram tokens).

4.3.2 Sentence overlap. Local cohesion overlap indices measure overlap between words at the sentence level. These indices have demonstrated positive relations with measures of cohesion in previous studies (McNamara, Louwerse, McCarthy, & Graesser, 2010), but generally demonstrate no negative relations with measures of coherence (Crossley et al., in press). TAACO calculates a number of sentence overlap indices. These indices compute lemma overlap between two adjacent sentences and between three adjacent sentences. TAACO calculates average overlap scores across a text for all lemma overlap, content word lemma overlap, and lemma overlap for POS tags such as nouns, verbs, adjectives, adverbs, and pronouns.

4.3.3 Paragraph overlap. Paragraph overlap indices measure overlap between words at the paragraph level. These indices have demonstrated positive relations with measures of text coherence in previous studies (Crossley et al., in press). TAACO calculates paragraph overlap indices between two adjacent paragraphs and between three adjacent paragraphs using the same features as the sentence overlap indices (e.g., average and binary lemma overlap, content word lemma overlap, and lemma overlap for part of speech tags).

4.3.4 Semantic overlap. Semantic overlap measures similarities between words at the sentence and paragraph levels. Semantic overlap indices have demonstrated positive relations with measures of cohesion (McNamara et al., 2010) and coherence in previous studies (Crossley et al., in press). Using the Wordnet database, TAACO calculates overlap between words and word synsets between sentences and between paragraphs. Unlike strict overlap indices, these indices will measure overlap between semantically related words (e.g., the synset for *jump* contains the related words *leap*, *bound,* and *spring*, among others). TAACO calculates semantic overlap between sentences and paragraphs for nouns and for verbs.

4.3.5 Anaphoric reference. Anaphoric reference refers to whether a previous noun is referred to using an indirect reference (i.e., a pronoun). TAACO calculates the incidence of a variety of referential pronoun types including first, second, and third person pronouns, subject pronouns, and quantity pronouns because pronouns can provide an indication of anaphoric reference (Crossley, Allen, Kyle, & McNamara, 2014). TAACO also calculates the ratio of nouns to pronouns.

4.3.6 Givenness. Givenness is an important element of measuring cohesion and reflects the amount of information that is recoverable from the preceding discourse. To assess givenness, TAACO counts the incidence of definite articles and demonstratives under the presumption that definiteness is associated with given information. Lastly, TAACO calculates the number and proportion of single content lemmas (i.e., how many lemmas occur only once in a text).

4.3.7 Connectives. Connectives are used to link segments of text together to create greater text coherence. TAACO contains a number of connective indices. Many of the connective indices are similar to those found in Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014) and are theoretically based on two dimensions. The first dimension contrasts positive versus negative connectives, and the second dimension is associated with particular classes of cohesion identified by Halliday and Hasan (1976) and Louwerse (2001), such as temporal, additive, and causative connectives. TAACO also reports on a number of indices based on how connectives operate rhetorically in written texts. The new lists include coordinating connectives, semi-coordinators, basic coordinators, quasi-coordinators, conjunctions, disjunctions, simple subordinators, complex subordinators, coordinating conjuncts, addition, sentence linking, order, reference, reason and purpose, condition, concession, and opposition.

4.4 Statistical Analysis

We conducted two analyses. The first examined differences between science (BIO and PHY) and engineering texts (MEC and IOE). The second examined differences for each micro-discipline. For each analysis, we first conducted a Multiple Analysis of Variance (MANOVA) to test if the selected indices that were normally distributed also demonstrated significant differences between the macro-disciplines and micro-disciplines. Next, we conducted a stepwise discriminant function analysis (DFA) using only the indices from each set that showed significant differences between the disciplines, but did not exhibit multicollinearity with other indices in the set. We set the threshold for multicollinearity at $r > .899$ (Tabachnick & Fidell, 2001). A discriminant function is generated by the DFA. This discriminant function produces an algorithm that can be used to predict group membership (i.e., the micro-disciplines of the texts).

We first conducted a DFA on the entire set of student writing samples. The model reported by this DFA was then used to predict group membership of the student writing samples using leave-one-out-cross-validation (LOOCV). LOOCV is a method designed to avoid overfitting a statistical or machine-learning model (Witten, Frank, & Hall, 2011). In this type of validation, a fixed number of folds equal to the number of observations (i.e., the 162 student writing samples) are selected. For each fold, one observation in turn is left out and the remaining instances are used as the training set (in this case the 161 remaining writing samples). We tested the accuracy of the model based on its ability to predict the discipline classification of the omitted instance. The LOOCV procedure allows testing of the accuracy of the model on an independent data set (i.e., on data that is not used to train the model). If the discriminant analysis model for both the entire set and the *n*-fold cross-validation set predict similar classifications, then the strength of the model to extend to external data sets is supported. In addition to using LOOCV to avoid overfitting the models, we ensured that the models had a minimum of 10 events per predictor variable (i.e., ten texts for each linguistic variable selected). Such a ratio is standard to control for overfitting in similar models (Concato, Peduzzi, Holford, & Feinstein, 1995; Freedman & Pee, 1989; Stevens, 2002). Thus, because we had a sample size of 162 texts, we limited the number of predictor variables (the indices from TAALES and TAACO) for each model to 16.

5.0 Results

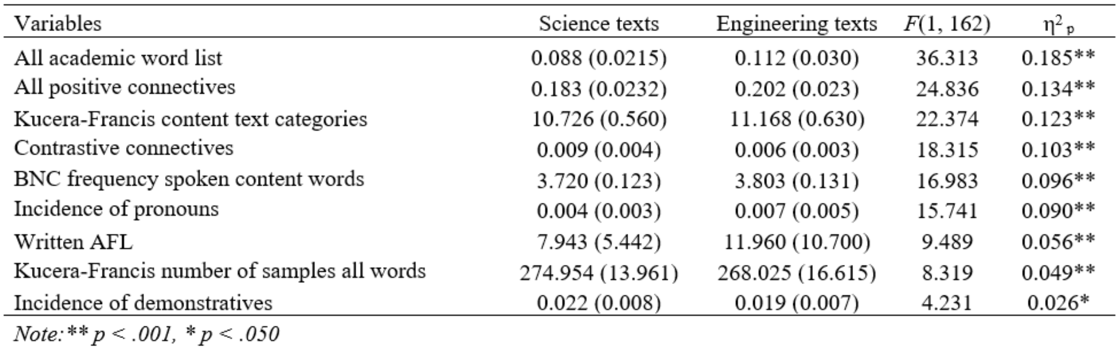
5.1 Macro-disciplinary differences

5.1.1 MANOVA. A MANOVA was conducted using TAALES and TAACO indices as the dependent variables and the text groupings of science and engineering as the independent variables. The 60 selected indices were first checked for normal distribution. All variables that were normally distributed and reported significant differences between the two disciplines were then assessed using Pearson correlations for multicollinearity (with a threshold of $r > .90$). Thirty-four indices demonstrated significant differences between science and engineering text while demonstrating normal distributions and not demonstrating multicollinearity with one another. These indices were used as predictor variables in the DFA.

5.1.2 Discriminant function analysis. We used a stepwise DFA to select the variables that best classify the grouping variable (text discipline). For our analysis, the significance level for a variable to enter or to be removed from the model was set at the norm generally adopted in applied linguistics: $p \leq 0.05$ (Larson-Hall, 2010). The stepwise DFA retained nine variables as significant predictors of discipline: *Kucera-Francis number of samples all words, All positive connectives* (e.g., after, in addition, therefore), *BNC frequency spoken content words, Contrastive connectives* (e.g., but, in contrast, conversely), *Incidence of pronouns, All AWL, Written AFL, Kucera-Francis categories content words,* and *Incidence of demonstratives*. Descriptive statistics and MANOVA results for these indices ordered by effect size are presented in Table 5.

Table 5

Means (Standard Deviations), F Values, and Effect Sizes for Science and Engineering Texts



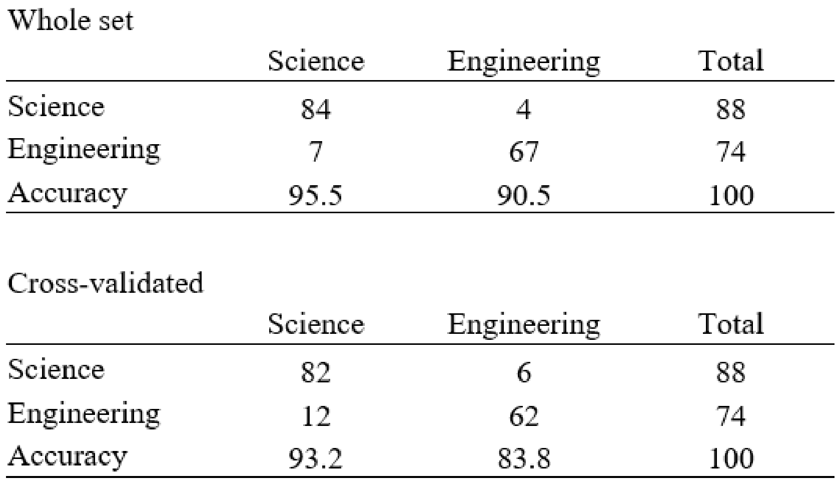
Variables	Science texts	Engineering texts	*F*(1, 162)	$\eta^2{}_p$
All academic word list	0.088 (0.0215)	0.112 (0.030)	36.313	0.185**
All positive connectives	0.183 (0.0232)	0.202 (0.023)	24.836	0.134**
Kucera-Francis content text categories	10.726 (0.560)	11.168 (0.630)	22.374	0.123**
Contrastive connectives	0.009 (0.004)	0.006 (0.003)	18.315	0.103**
BNC frequency spoken content words	3.720 (0.123)	3.803 (0.131)	16.983	0.096**
Incidence of pronouns	0.004 (0.003)	0.007 (0.005)	15.741	0.090**
Written AFL	7.943 (5.442)	11.960 (10.700)	9.489	0.056**
Kucera-Francis number of samples all words	274.954 (13.961)	268.025 (16.615)	8.319	0.049**
Incidence of demonstratives	0.022 (0.008)	0.019 (0.007)	4.231	0.026*

Note: ** $p < .001$, * $p < .050$

The results demonstrate that the DFA using the nine significant TAALES and TAACO indices correctly allocated 151 of the 162 writing samples in the total set, χ2 (df=1, *n*=162) = 120.747, p < .001, for an accuracy of 93.2% (chance for this analysis is 50% and baseline is 54%). For the leave-one-out cross-validation (LOOCV), the discriminant analysis correctly allocated 144 of the 162 writing samples for an accuracy of 88.9% (see the confusion matrix reported in Table 6 for results), indicating that the model is stable across the dataset. The measure of agreement between the actual discipline categorization and that assigned by the model produced a Cohen’s Kappa of 0.863, demonstrating an almost perfect agreement (Landis and Koch, 1977).

Table 6

Confusion Matrix for DFA Results: Science vs. Engineering Texts



Whole set

	Science	Engineering	Total
Science	84	4	88
Engineering	7	67	74
Accuracy	95.5	90.5	100

Cross-validated

	Science	Engineering	Total
Science	82	6	88
Engineering	12	62	74
Accuracy	93.2	83.8	100

5.2 Micro-Disciplinary Differences

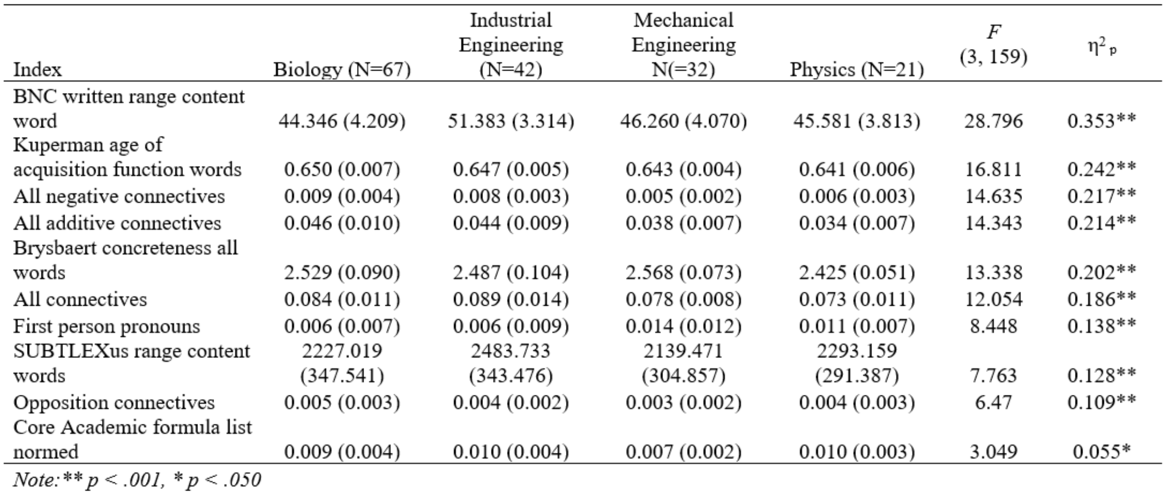
5.2.1 MANOVA. A MANOVA was conducted using TAALES and TAACO indices as the dependent variables and the text groupings of Biology, Physics, Industrial Engineering, and Mechanical Engineering as the independent variables. The 60 indices were first checked for normal distribution. All variables that were normally distributed and reported significant differences between the two disciplines were then assessed using Pearson correlations for multicollinearity (with a threshold of $r > .90$). Forty-three indices demonstrated significant difference between the micro-discipline texts while demonstrating normal distributions and not demonstrating multicollinearity with one another. These indices were used as predictor variables in the DFA.

5.2.2 Discriminant function analysis. We used a stepwise DFA to select the variables that best classify the grouping variable (text micro-disciplines). For our analysis, the significance level for a variable to enter or to be removed from the model was set at the norm generally adopted in applied linguistics: $p \leq 0.05$. The stepwise DFA retained 10 variables as significant predictors of discipline: *BNC written range content words, Kuperman age of acquisition function words, All negative connectives* (e.g., although, until, nonetheless), *All additive connectives* (e.g., also, and, actually), *Brysbaert concreteness all words, All connectives, First person pronouns, SUBTLEXus range content words, Opposition connectives* (e.g., but, however, yet), and *Core AFL normed.*

Descriptive statistics and MANOVA results for these indices ordered by effect size are presented in Table 7.

Table 7

Means (Standard Deviations), F values, and Effect Sizes for Biology, Physics, Industrial Engineering, and Mechanical Engineering Texts



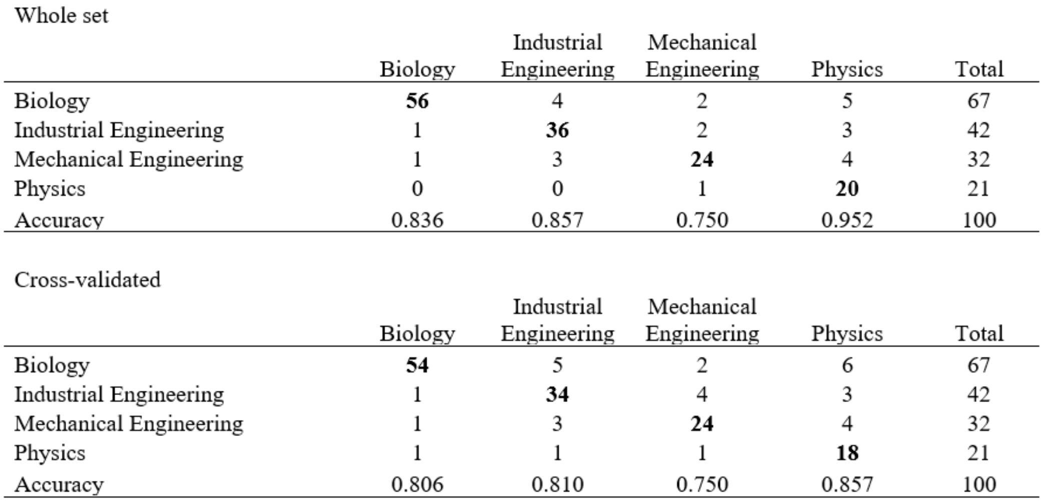
Index	Biology (N=67)	Industrial Engineering (N=42)	Mechanical Engineering N(=32)	Physics (N=21)	*F* (3, 159)	η^2 p
BNC written range content word	44.346 (4.209)	51.383 (3.314)	46.260 (4.070)	45.581 (3.813)	28.796	0.353**
Kuperman age of acquisition function words	0.650 (0.007)	0.647 (0.005)	0.643 (0.004)	0.641 (0.006)	16.811	0.242**
All negative connectives	0.009 (0.004)	0.008 (0.003)	0.005 (0.002)	0.006 (0.003)	14.635	0.217**
All additive connectives	0.046 (0.010)	0.044 (0.009)	0.038 (0.007)	0.034 (0.007)	14.343	0.214**
Brysbaert concreteness all words	2.529 (0.090)	2.487 (0.104)	2.568 (0.073)	2.425 (0.051)	13.338	0.202**
All connectives	0.084 (0.011)	0.089 (0.014)	0.078 (0.008)	0.073 (0.011)	12.054	0.186**
First person pronouns	0.006 (0.007)	0.006 (0.009)	0.014 (0.012)	0.011 (0.007)	8.448	0.138**
SUBTLEXus range content words	2227.019 (347.541)	2483.733 (343.476)	2139.471 (304.857)	2293.159 (291.387)	7.763	0.128**
Opposition connectives	0.005 (0.003)	0.004 (0.002)	0.003 (0.002)	0.004 (0.003)	6.47	0.109**
Core Academic formula list normed	0.009 (0.004)	0.010 (0.004)	0.007 (0.002)	0.010 (0.003)	3.049	0.055*

*Note:*** *p < .001,* * *p < .050*

The results demonstrate that the DFA using the ten significant TAALES and TAACO indices correctly allocated 136 of the 162 writing samples in the total set, χ2 (df=9, *n*=162) = 290.716, p < .001, for an accuracy of 84% (chance for this analysis is 25% and baseline is 41%). For the leave-one-out cross-validation (LOOCV), the discriminant analysis correctly allocated 130 of the 162 writing samples for an accuracy of 80.2% (see the confusion matrix reported in Table 8 for results), indicating that the model was stable across the dataset. The measure of agreement between the actual discipline categorization and that assigned by the model produced a Cohen’s Kappa of 0.778, demonstrating a substantial agreement (Landis and Koch, 1977).

Table 8

Confusion Matrix for DFA Results: Biology, Physics, Industrial Engineering, and Mechanical Engineering



Whole set

	Biology	Industrial Engineering	Mechanical Engineering	Physics	Total
Biology	**56**	4	2	5	67
Industrial Engineering	1	**36**	2	3	42
Mechanical Engineering	1	3	**24**	4	32
Physics	0	0	1	**20**	21
Accuracy	0.836	0.857	0.750	0.952	100

Cross-validated

	Biology	Industrial Engineering	Mechanical Engineering	Physics	Total
Biology	**54**	5	2	6	67
Industrial Engineering	1	**34**	4	3	42
Mechanical Engineering	1	3	**24**	4	32
Physics	1	1	1	**18**	21
Accuracy	0.806	0.810	0.750	0.857	100

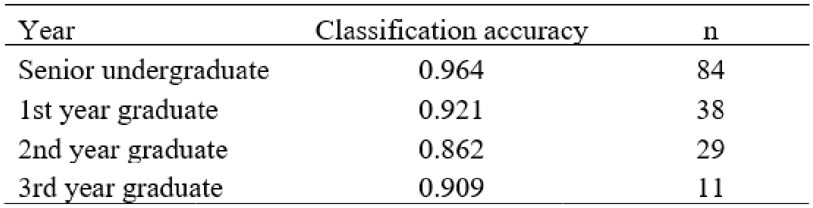
5.3 Post-hoc Analysis

We conducted post-hoc analyses of the DFA results for the macro-discipline and micro-discipline corpora. Our purpose in the post-hoc analyses was to ensure that year of study differences (i.e., differences between senior undergraduate, first- second-, and third-year graduate students) and paper types (e.g., proposals, reports, research papers) were not confounding the accuracy results reported by the DFA.

5.3.1 Year of study differences. For the macro-discipline analysis, the lowest accuracy was reported for the second-year graduate students (86%), and the highest accuracy was reported for the senior undergraduate students (96%, see Table 9 for classification accuracies). For the micro-discipline analysis, the lowest accuracy was reported for the senior undergraduate students (80%), and the highest accuracy was reported for the first-year graduate students (92%, see Table 10 for classification accuracies). The differences in accuracy hovered around 10%, providing some indication that year of study was not a confounding variable in classification accuracy.

Table 9

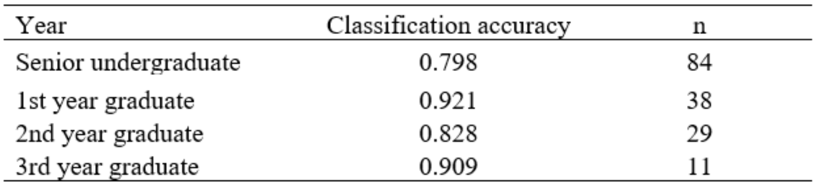
Differences in Accuracy Classification by Year of Study: Science and Engineering Texts



Year	Classification accuracy	n
Senior undergraduate	0.964	84
1st year graduate	0.921	38
2nd year graduate	0.862	29
3rd year graduate	0.909	11

Table 10

Differences in Accuracy Classification by Year of Study: Biology, Physics, Industrial Engineering, and Mechanical Engineering Texts

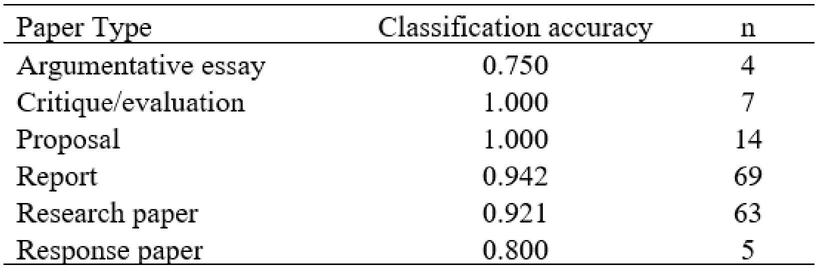


Year	Classification accuracy	n
Senior undergraduate	0.798	84
1st year graduate	0.921	38
2nd year graduate	0.828	29
3rd year graduate	0.909	11

5.3.2 Paper type differences. For the discipline analysis, the lowest accuracy was reported for the argumentative essays (75%), and the highest accuracy was reported for critique/evaluation and proposal (100%, see Table 11 for classification accuracies). For the micro-discipline analysis, the lowest accuracy was reported for the argumentative essays (50%), and the highest accuracy was reported for the critiques/evaluations (100%, see Table 12 for classification accuracies). The differences in accuracy were large for a few paper types, but the sample sizes for these paper types were very small (between 4-7 samples), calling into question the reliability of these differences. For those paper types that had larger sample sizes, the classification accuracies were similar, providing some indication that paper type was not a confounding variable in classification accuracy.

Table 11

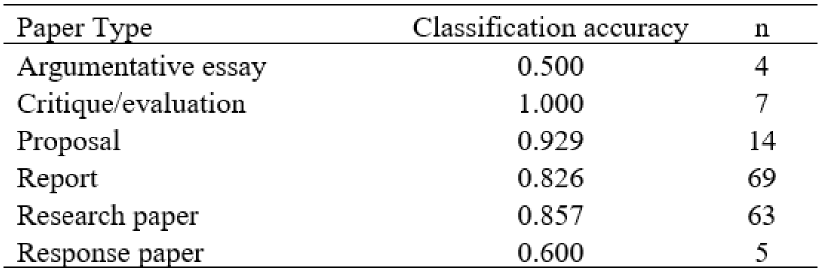
Differences in Paper Type: Science and Engineering Texts



Paper Type	Classification accuracy	n
Argumentative essay	0.750	4
Critique/evaluation	1.000	7
Proposal	1.000	14
Report	0.942	69
Research paper	0.921	63
Response paper	0.800	5

Table 12

Differences in Paper Type: Biology, Physics, Industrial Engineering, and Mechanical Engineering Texts



Paper Type	Classification accuracy	n
Argumentative essay	0.500	4
Critique/evaluation	1.000	7
Proposal	0.929	14
Report	0.826	69
Research paper	0.857	63
Response paper	0.600	5

6.0 Discussion

Our goal in this study was to systematically examine language differences in student writing from two macro-disciplines and four micro-disciplines to examine the potential for linguistic features to distinguish between macro-disciplines and micro-disciplines. We did this through the use of a corpus of student papers, a suite of natural language processing tools, and statistical analyses. While previous research has demonstrated discipline differences in student writing at the macro-discipline level (Durrant, 2014, in press; Hardy & Römer, 2013), our purpose was to investigate if such differences existed in student writing at both the macro- and micro-levels. In addition, previous research has mostly focused on grammatical and syntactic features (Hardy & Römer, 2013) and domain-dependent lexical features (i.e., frequency counts based on the corpus under investigation; Durrant, 2014, in press) to investigate macro-disciplinary differences. In contrast, this study examined domain-independent lexical features and cohesion features.

Overall, the findings provide strong support that macro-disciplinary and micro-disciplinary differences exist in student writing and that these differences are likely not related to student level. These findings have important implications for understanding disciplinary differences. We discuss relevant ideas below organized around the three central results (macro-disciplinary variation, micro-disciplinary variation, and year of study differences).

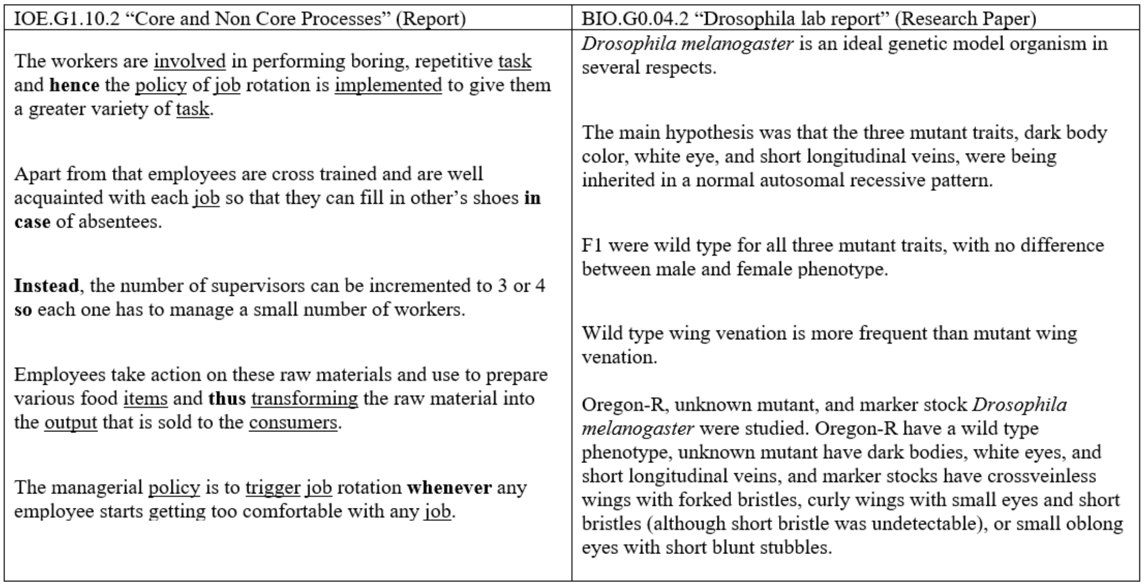
6.1 Macro-Disciplinary Variation

The statistical analyses applied in our study reveal very large differences between student writing in the two disciplines in terms of lexical features, confirming Durrant's (2015) finding that the vocabulary used by different disciplinary areas "is strikingly diverse" (352). Lexically, the results indicate that engineering writing samples contain more frequent words that occur in a greater number of text categories than science texts, while simultaneously containing a greater number of academic words and academic formulas. Moreover, our study extends the linguistic analyses conducted by Durrant (2015) and Hardy and Römer (2013) to include cohesion features. From a cohesion perspective, engineering texts contain more positive connectives and pronouns than science texts, while science texts contain more contrastive connectives and demonstratives. The results indicate that engineering texts use more frequent and more academic words, whereas science texts use more specialized vocabulary that is specific to a smaller range of texts and does not rely on traditional academic words or phrases. Engineering texts are also more additive in nature (e.g., they contain more connectives such as *after, in addition, therefore*) and depend more on pronominal reference, while science texts are more contrastive (e.g., contain more contrastive connectives such as *but, in contrast, conversely*) and provide more specific references.

As an example of this, we provide excerpts from an Industrial Engineering (IOE) text and a Biology text (Table 13). In the samples, the academic words are in bold and connectives are underlined. The excerpts illustrate the results reported in the statistical analyses in that the IOE text contains a greater number of academic words and connectives. The Biology text contains zero connectives and, while the text contains a number of discipline-specific words that are infrequent, the words are not in Coxhead's (2000) academic word list.

Table 13

Text Excerpts from Engineering and Science Texts



IOE.G1.10.2 "Core and Non Core Processes" (Report)	BIO.G0.04.2 "Drosophila lab report" (Research Paper)
The workers are involved in performing boring, repetitive task and **hence** the policy of job rotation is implemented to give them a greater variety of task. Apart from that employees are cross trained and are well acquainted with each job so that they can fill in other's shoes **in case** of absentees. **Instead**, the number of supervisors can be incremented to 3 or 4 **so** each one has to manage a small number of workers. Employees take action on these raw materials and use to prepare various food items and **thus** transforming the raw material into the output that is sold to the consumers. The managerial policy is to trigger job rotation **whenever** any employee starts getting too comfortable with any job.	*Drosophila melanogaster* is an ideal genetic model organism in several respects. The main hypothesis was that the three mutant traits, dark body color, white eye, and short longitudinal veins, were being inherited in a normal autosomal recessive pattern. F1 were wild type for all three mutant traits, with no difference between male and female phenotype. Wild type wing venation is more frequent than mutant wing venation. Oregon-R, unknown mutant, and marker stock *Drosophila melanogaster* were studied. Oregon-R have a wild type phenotype, unknown mutant have dark bodies, white eyes, and short longitudinal veins, and marker stocks have crossveinless wings with forked bristles, curly wings with small eyes and short bristles (although short bristle was undetectable), or small oblong eyes with short blunt stubbles.

These linguistic features were powerful predictors, classifying science and engineering texts with almost perfect accuracy (93.2%). This is in some ways unsurprising. Natural science (biology and physics in this corpus) and engineering (here mechanical and industrial) are linked together in ordinary understanding under the broad rubric of STEM disciplines, yet they are usually housed in separate, large academic units (e.g., colleges), and they are distinguished by a difference in orientation: "pure" science versus "applied" science, in common (Gieryn, 1983). Additionally, engineering, at least in the US, has a tradition of writing instruction specifically for it (Russell, 2002).

The findings suggest that the common understanding of the STEM disciplines as "close" bears reconsideration. In particular, upper level academic writing courses in the US are often taken by students in all STEM disciplines without differentiation by domain (Russell, 2002). Yet even at upper level undergraduate study, student writing based on discipline appears to be quite different. Thus, this study does not support the notion that macro-disciplines at all levels of writing are homogeneous in their use of lexical features (Durrant, 2015). In addition, this study suggests that macro-disciplines are also not homogeneous in terms of their use of cohesion features.

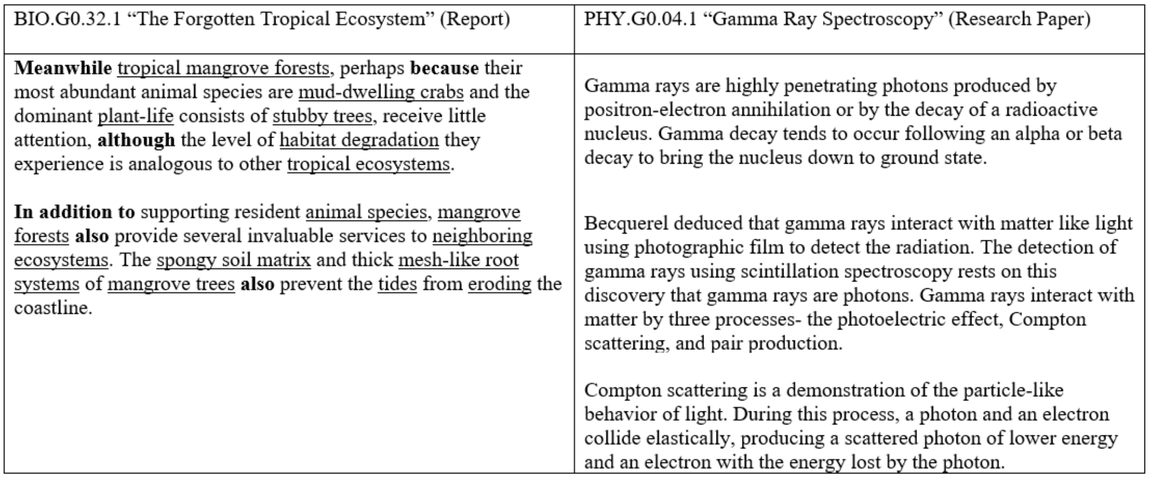
6.2 Micro-Disciplinary Variation

More notable than the disciplinary difference findings are the lexical and cohesion differences in student writing between micro-disciplines. These differences indicate a number of differences and similarities among micro-disciplines such that each micro-discipline can be thought of as containing unique linguistic features. For instance, lexically, biology writing samples contain function words that are thought to be acquired later by children and lower written range scores (i.e., more specific words). In addition, they contain more concrete words. Cohesively, biology writing samples contain more negative, opposition, additive, and overall connectives. Biology texts also have the lowest incidence of first person pronouns. Physics writing samples, in contrast, have lower scores for age of acquisition (function words), lower word concreteness, and higher range SUBTLEXus counts (i.e., less specific words). Cohesively, physics writing samples differ from biology texts in that they have fewer negative, additive, and overall connectives. They also contain a greater number of first person pronouns. Thus, strong differences seem to be apparent in two science micro-disciplines.

As an example of these differences, we present two text samples in Table 14. The first sample is from a Biology text in which low range words (i.e., less specific) are underlined and connectives are in bold. Both linguistic features are common in the Biology text whereas they are not represented in the Physics text even though both texts are from the same macro-discipline.

Table 14

Text Excerpts from Science Texts



BIO.G0.32.1 "The Forgotten Tropical Ecosystem" (Report)	PHY.G0.04.1 "Gamma Ray Spectroscopy" (Research Paper)
Meanwhile tropical mangrove forests, perhaps **because** their most abundant animal species are mud-dwelling crabs and the dominant plant-life consists of stubby trees, receive little attention, **although** the level of habitat degradation they experience is analogous to other tropical ecosystems. **In addition to** supporting resident animal species, mangrove forests **also** provide several invaluable services to neighboring ecosystems. The spongy soil matrix and thick mesh-like root systems of mangrove trees **also** prevent the tides from eroding the coastline.	Gamma rays are highly penetrating photons produced by positron-electron annihilation or by the decay of a radioactive nucleus. Gamma decay tends to occur following an alpha or beta decay to bring the nucleus down to ground state. Becquerel deduced that gamma rays interact with matter like light using photographic film to detect the radiation. The detection of gamma rays using scintillation spectroscopy rests on this discovery that gamma rays are photons. Gamma rays interact with matter by three processes- the photoelectric effect, Compton scattering, and pair production. Compton scattering is a demonstration of the particle-like behavior of light. During this process, a photon and an electron collide elastically, producing a scattered photon of lower energy and an electron with the energy lost by the photon.

From the engineering discipline, industrial engineering writing samples contain the highest BNC range scores (less specific words), lower word concreteness, and a greater number of academic formulas. From a cohesion perspective, industrial engineering writing samples contain more negative, additive, and overall connectives (in a manner similar to biology writing samples). They also have the lowest incidence of first person pronouns (along with biology writing samples). Mechanical engineering writing samples, in contrast, have lower BNC and SUBTLEXus range scores (i.e., more specific words), lower age of acquisition scores for function words, and higher concreteness scores. Cohesively, mechanical engineering writing samples have a lower incidence of negative, additive, opposition, and overall connectives (like physics writing samples) and the highest incidence of first person pronouns.

What we see, then, is a unique linguistic profile that arises for the samples of texts taken from each micro-discipline. We also see some similarities in the profile across disciplines such that industrial engineering texts share cohesive properties with biology texts and mechanical engineering texts have similar cohesive properties as physics texts. However, the differences among the micro-disciplines are quite strong and allow for a categorization accuracy of 84% across micro-disciplines. The effect size for this classification was robust, showing substantial agreement between the actual classification and the predicted classification. Thus, the NLP tools employed in this analysis allow us to distinguish student texts not only between macro-disciplines but also between micro-disciplines, and again, the differences are large (though not quite as large as between macro-disciplines).

The results suggest that there are important differences in disciplines that are perceived, from the outside, to be similar. This finding is interesting for a number of reasons. First, it problematizes previous research into differences between science and engineering texts in terms of abstract language use. For instance, in two studies, Biber (1988, 2006) reported that professional engineering texts contained more abstract information than science texts. However, the findings from this study indicate that a more nuanced interpretation may be necessary, at least in terms of learner texts. For instance, while industrial engineering texts may contain more abstract (i.e., less concrete) words, mechanical engineering texts contain more concrete words. Differences within macro-disciplines are also interesting in relation to the anecdotal and qualitative evidence that both faculty and students in the disciplines perceive the important—even critical—differences between their disciplines and closely related disciplines (Bazerman & Paradis, 1991). For insiders to a field, the differences are visible and important, while to outsiders they may be invisible or appear insignificant. Researchers have reported perceiving differences within a very narrow set of tolerances: a turn of phrase or framing of a problem that sets one subfield or sub-subfield apart from others (Bazerman, 1985; Harwood, 2006). Through their initiation into the discourse, students and experts within the micro-disciplines appear to reproduce these differences linguistically, implying that students recognize (consciously or not) disciplinary differences—even when disciplines are proximal. Of course, recognizing and reproducing differences are not the same as thinking and conceiving of disciplinary knowledge in ways that are inherent to one discipline over another, nor are they the same as writing successfully within a discipline. However, since the essays found in MICUSP are all highly successful essays (i.e., all received an A grade), it is likely that the use of the linguistic features that distinguish between macro- and micro-disciplines may relate to essay quality. However, additional studies examining a range of both low and high quality essays are needed to examine if the linguistic features that distinguish between micro- and macro-disciplines are also predictive of writing quality within those disciplines.

6.3 Year of Study Differences

Our last discussion section is in reference to the post-hoc analysis, which demonstrated that the differences discerned in the quantitative analysis are evident at least as early as the final year of undergraduate study. Moreover, the differences are relatively homogeneous across levels. This again confirms the findings of Durrant (2015) in terms of lexical features, but, importantly, extends these findings to include cohesion markers.

This raises the question of when and to what extent (at any given point) these disciplinary differences become evident in students' writing, particularly when fields are in the same general area (e.g., natural sciences). The fact that these are successful (A-graded) papers may in part explain this result, as the texts come from students who have more readily internalized the discourse of the discipline (and have been rewarded for it with higher grades). But the fact that the differences are so strongly evident as early as the last year of undergraduate education is particularly remarkable considering that in the US higher education system, students specialize later, overall, because they have two years of general education before taking a program of study primarily or exclusively dedicated to a discipline. Indeed, these results raise the question of when and under what circumstances large and very discipline-specific differences appear in student writing. In education systems where students specialize much earlier and devote their full attention to one discipline, these differences may appear in the first years of higher education or in secondary school (Krueger & Kumar, 2004; Osborne & Dillon, 2008).

7.0 Conclusion

The findings from this study provide evidence that macro-disciplinary and micro-disciplinary differences at the linguistic level exist in student writing. Moreover, these differences do not appear to be related to student level. Writing analytics focuses on the measurement of text features to better understand writing within education contexts, so the question remains as to how these findings might improve the teaching and learning of writing. In this regard, the findings of this study provide some guidance for teachers and students. Specifically, a contrastive rhetoric approach based on the reported findings would allow teachers to position writing within a discipline so that students would have the opportunity to understand or even examine differences between macro- and micro-discipline. For instance, teachers could provide students with discipline-specific writing guidelines for producing text samples that fit discipline expectations. Since the findings from this study indicate that discipline differences emerge as early as the last year of undergraduate education, it is likely that guidelines provided to students may match their already evolved tacit knowledge of the discipline. Thus, such guidelines would provide additional support for already developed writing expectations. More advanced students could use corpus analysis tools that generate word- or cluster-lists and allow for the visual examination of concordance lines to empirically analyze differences between macro- and micro-disciplines. Such approaches would allow students to better recognize the writing expectations of their specific discipline and provide concrete examples of discipline differences.

While these applications could prove helpful in the writing classroom, additional studies are necessary to overcome limitations within the current study. For instance, the size of the corpus and range of text types found within the current corpus are small. While appropriate for the analyses conducted, a larger corpus comprised of a greater number of writing samples, such as the BAWE corpus, would allow for greater generalizations to be made about the findings and provide greater confidence that the findings can be extended to a larger population. In addition, while the post-hoc analyses seem to indicate that paper type differences were not an intervening factor in the classification accuracy, the sample size for a few paper types (argumentative essay, critique/evaluation, response paper) was too small to completely allay concerns that paper type may interact with the linguistic features selected for this analysis. Additionally, the use of NLP tools to measure language use can be an imprecise metric. For instance, the tools used here can tell us about incidences of words and discourse markers as well as provide information about the words used. However, the tools cannot tell us if the words were used appropriately or if connectives were under- or over-used. Lastly, the disciplines examined in this research are heavily skewed toward the demonstration of knowledge and are unlike some other disciplines, such as those in the humanities, which focus on argumentation and critique. As a result, the findings from this analysis may be specific to knowledge-demonstrating disciplines only.

8.0 Directions for Future Research

Our corpus-based findings on macro-disciplinary differences seem to indicate that students are aware of disciplinary conventions and expectations, at least enough so that they are reflected in their writing, but more research is warranted to support this claim. Further NLP analyses of upper-level student writing corpora from different disciplinary families (e.g., social sciences, humanities, applied sciences other than engineering, etc.) would permit a map of disciplinary differences and similarities such that one could gauge the relative distances among the disciplines. This would help address concerns about potential differences between knowledge-demonstration disciplines (such as those used in this study) and disciplines that focus on argumentation and critique. An obvious first step would be to conduct similar NLP analyses with other subsets of MICUSP or other existing corpora of student writing. A likely first candidate is the British Academic Written English (BAWE) corpus (Nesi & Gardner, 2012). It is a larger corpus than MICUSP in terms of numbers of student texts. It includes some of the same disciplines as MICUSP, though they are grouped into different families. Nesi and Gardner (2012), who assembled the BAWE corpus, specifically call for a comparison between BAWE and MICUSP. As they point out, such an analysis would allow the comparison to include national differences in educational practices, within and across disciplines. Other large corpora should be developed to address data limitations in both MICUSP and BAWE, including coverage related to student level and language ability, student writing proficiency, and other individual difference features that may influence text production.

In addition, future studies should include NLP approaches that capture data that goes beyond the lexical and cohesive features of text. Such analyses could include the use of part of speech tags, syntactic parsing, and indices related to syntactic complexity and similarity. Additional NLP indices could investigate differences between disciplines in terms of rhetorical features (i.e., theses and arguments) or more advanced cohesion metrics computed using latent semantic analysis, latent Dirichlet allocation, and Word2Vec approaches.

Author Biographies

Dr. Scott Crossley is an Associate Professor of Applied Linguistics at Georgia State University. Professor Crossley's primary research focus is on natural language processing and the application of computational tools and machine learning algorithms in language learning, writing, and text comprehensibility.

David R. Russell has published widely on writing across the curriculum (WAC), international writing instruction, activity theory and genre theory. He is the author of *Writing in the Academic Disciplines: A Curricular History*, numerous articles, and co-editor of four collections. He also edits the *Journal of Business and Technical Communication.*

Kristopher Kyle is an Assistant Professor in the department of Second Language Studies at the University of Hawai'i. His research interests include second language acquisition and second language writing and speaking assessment. He is especially interested in applying natural language processing (NLP) and corpora to the exploration of these areas.

Dr. Ute Römer is an Associate Professor of Applied Linguistics at Georgia State University. Her research interests include corpus linguistics, phraseology, and second language acquisition. She serves on a range of editorial and advisory boards of professional journals and organizations, and is general editor of the Studies in Corpus Linguistics book series.

Acknowledgments

This research stems from an invitation to the *Symposium on Linguistic Perspectives on Discipline-Specific Language* hosted by the English Department at Iowa State University in 2014. We thank the organizers for developing and hosting the symposium.

References

Ädel, A., & Römer, U. (2012). Research on advanced student writing across disciplines and levels: Introducing the Michigan Corpus of Upper-level Student Papers. *International Journal of Corpus Linguistics, 17*(1), 3-34.

Bazerman, C. (1985). Physicists reading physics schema-laden purposes and purpose-laden schema. *Written Communication, 2*(1), 3–23.

Bazerman, C., & Paradis, J. (1991). *Textual dynamics: Historical & contemporary studies of writing in professional communities*. Madison: University of Wisconsin Press.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes, 5*(2), 97-116.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus*. TOEFL Monograph Series. Princeton, NJ: ETS.

Brown, G. D. A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavioural Research Methods Instrumentation and Computers, 16*, 502-532.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977-990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904-911.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, *14*(2), 30–49.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A*, 497-505.

Cortes, V. (2004). Lexical bundles in published and student writing in history and biology. *English for Specific Purposes*, *23*(4), 397-423.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D.S. (2014). Analyzing discourse processing using the Simple Natural Language Processing Tool (SiNLP). *Discourse Processes, 51*(5-6), 511-534.

Crossley, S. A., Cai, Z., & McNamara, D. S. (2012). Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In P. M.McCarthy & G. M. Youngblood (Eds.), *Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 214-219). Menlo Park, CA: The AAAI Press.

Crossley, S. A., Kyle, K., & McNamara, D. S. (in press). The Tool for the Automatic Analysis of Text Cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods.*

Crossley, S. A., & McNamara, D. S. (2009). Computationally assessing lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, *17*(2), 119-135.

Crossley, S. A., & McNamara, D. S. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology, 17*(2), 171-192.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learners using computational indices. *Language Testing, 28*(4), 561-580.

Durrant, P. (2014). Discipline- and level-specificity in university students' written vocabulary. *Applied Linguistics, 35*(3), 328-356.

Durrant, P. (in press). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics.*

Ewer, J. R., & Latorre, G. (1969). *A course in basic scientific English*. London: Longman.

Flowerdew, L. (2014). Learner corpus research in EAP: Some core issues and future pathways. *English Language and Linguistics*, *20*(2), 43-60.

Gieryn, T. (1983). Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. *American Sociological Review, 48*(6), 781–795.

Grabowski, L. (2015). Keywords and lexical bundles within English pharmaceutical discourse: A corpus-driven description. *English for Specific Purposes, 38*, 23-33.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hardy, J. A., & Römer, U. (2013). Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, *8*(2), 183-207.

Harwood, N. (2006). (In)appropriate personal pronoun use in political science: A qualitative study and a proposed heuristic for future research. *Written Communication, 23*(4), 424–450.

Hu, G., & Cao, F. (2015). Disciplinary and paradigmatic influences on interactional metadiscourse in research articles. *English for Specific Purposes, 39*, 12–25.

Hyland, K. (2002). Directives: Argument and engagement in academic writing. *Applied Linguistics, 23*, 215-239.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21.

Hyland, K., & Tse, P. (2009). The leading journal in its field: Evaluation in journal descriptions. *Discourse Studies, 6*, 703-720.

Krueger, D., & Kumar, K. B. (2004). Skill-specific rather than general education: A reason for US-Europe growth differences? *Journal of Economic Growth, 9*, 167-207.

Kucera H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Kuperman, V., Stadthagen-Gonzales, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods, 44*(4), 978-990.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757-786.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Louwerse, M.M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, *12*, 291–315.

McCarthy, P.M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381-392.

McNamara, D.S., Graesser, A.C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1–43.

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., & Graesser, A.C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*, 292-330.Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. New York, NY: Cambridge University Press.

O'Donnell, M. B., & Römer, U. (2012). From student hard drive to web corpus (part 2): The annotation and online distribution of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora, 7*(1), 1-18.

Osborne, J., & Dillon, J. (2008). *Science education in Europe: Critical reflections*. London: The Nuffield Foundation.

Ozturk, I. (2007). The textual organisation of research article introductions in applied linguistics: Variability within a single discipline. *English for Specific Purposes*, *26*(1), 25–38.

Römer, U., & O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*. *6*(2), 159-177.

Russell, D. R. (2002). *Writing in the academic disciplines: A curricular history*. Carbondale IL: Southern Illinois University Press.

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

Sinclair, J. M. (1999). The lexical item. In E. Weigand (Ed.), *Contrastive lexical semantics* (pp. 1–24). Amsterdam: John Benjamins.Swales, J. M. (1990). *Genre analysis. English in academic and research settings*. Cambridge, UK: Cambridge University Press.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics, fourth edition*. Needham Heights, MA: Allyn & Bacon.

The British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved from http://www.natcorp.ox.ac.uk/

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words.* New York: Teachers College, Columbia University.

Toglia, M. P., & Battig, W. R. (1978). *Handbook of semantic word norms*. New York: Lawrence Erlbaum.

Ward, J. (2007). Collocation and technicality in EAP Engineering. *Journal of English for Academic Purposes*, *6*, 18-35.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: The M.I.T. Press.