9

Examining Lexical and Cohesion Differences in Discipline-Specific Writing Using Multi-Dimensional Analysis

Scott A. Crossley, Kristopher Kyle, and Ute Römer

Introduction

English for Specific Purposes was founded on the simple notion that the texts that comprise specific academic disciplines (e.g., engineering or biology) were generally associated with specific linguistic features (i.e., lexical or grammatical choices). Thus, texts within disciplines were thought to be homogenous in their linguistic structures. Such homogeneity would allow teachers to develop pedagogical interventions based on detailed analyses of similarities within disciplines.

Early studies appeared promising with grammatical features such as the incidence of passives and conditional reporting differences between academic disciplines (Ewer and Latorre 1969). Since then, the variety and types of linguistic features identified and analyzed have widened to include not just grammatical features, but also language features related to discourse, lexical, and rhetorical patterns. This research has strengthened techniques that use linguistic features to explore disciplinary differences (Durrant 2014, 2015; Hyland 2002; Hyland and Tse 2009) and has provided evidence to support pedagogical interventions (de Chazal 2013). While much of this research has focused on differences at the macro level of disciplines (i.e., text differences in larger fields of study such as math and biology), some of the research has expanded to micro-disciplines (i.e., disciplines within disciplines). For instance Hu and Cao (2015) found differences in the metadiscourse used between qualitative and quantitative research paradigms in published papers from three social science micro-disciplines. Similarly, Ozturk (2007) examined differences between the move structure in published research article introductions found in second language studies (specifically within the micro-disciplines of second language acquisition and second language writing). Ozturk reported that the two micro-disciplines displayed different and mostly unrelated move structures, suggesting potential difference between an emerging field (second language writing) and a more established field (second language acquisition). Lastly, Ward (2007) conducted a corpus study of collocations in textbooks from five engineering

micro-disciplines. He found large differences among the micro-disciplines, raising the question of whether or not there is a common engineering vocabulary.

However, the majority of this research has focused on professional writing and has not examined potential disciplinary differences in student writing. The purpose of this study is to examine differences and similarities in student writing within two academic macro-disciplines (science and engineering) and among four academic micro-disciplines (biology, physics, industrial engineering, and mechanical engineering) using linguistic features related to lexical sophistication and text cohesion. We also examine potential differences in student writing in terms of gender, year of study (e.g., undergraduate and graduate students), and paper types (e.g., argumentative essays, proposals, and research papers).

Student writing

The majority of ESP studies examining disciplinary differences have focused on professional academic writing. While there is interest in examining student disciplinary writing, such research is still in its infancy and the majority of the studies available target differences in student and professional academic written (Biber et al. 2004; Cortes 2004) and spoken texts (Biber et al. 2004). However, recent research has moved beyond comparing student and professional texts and begun to examine disciplinary variation in student writing using corpus-based research. Much of this research is predicated on the development and availability of extensive new student writing corpora such as the British Academic Written English corpus (BAWE; Alsop and Nesi 2009; Nesi and Gardner 2012) and the Michigan Corpus of Upper-level Student Papers (MICUSP; Durrant 2014, 2015; Flowerdew 2015; Hardy and Römer 2013). These corpora are subdivided based on a number of academic disciplines affording researchers the opportunity to analyze discipline differences based on the grammatical and lexical features found in the student texts.

For example, Hardy and Römer (2013) used the Biber tagger, which assigns grammatical and syntactic tags to words and phrases, to investigate differences across four general academic divisions represented in MICUSP: humanities and arts, social sciences, biological and health sciences, and physical sciences. Hardy and Römer reported that the disciplines within these divisions varied linguistically in a number of different ways. For instance, student papers written in physics and biology courses tended to be more informationally dense (e.g., included nominalizations, attributive adjectives, and relatively long words) than those written in philosophy and education while philosophy and education papers tended to be more involved (e.g., included more first and second person pronouns and a greater number of verbs).

In a larger scale analysis, Durrant (2014) analyzed the words produced by student writers in the eighty-six discipline levels found in the BAWE corpus. The discipline levels consisted of combinations of disciplines (e.g., agriculture, business, mathematics) and was available for four student levels. Durrant created discipline-specific unlemmatized frequency lists specific to the BAWE corpus for each level. Using these lists, he examined the extent to which the words in the lists were shared across the disciplines.

Durrant found that about 50 percent of the words were discipline-specific while the remaining 50 percent of the words used were generic (i.e., were common across all disciplines). A subsequent analysis of how discipline-specific words grouped together yielded evidence that various levels for each discipline clustered based on vocabulary promoting the notion that discipline-specific vocabulary is not diverse, although there were exceptions. The majority of these exceptions were found for postgraduate level writing providing evidence that, in some cases, postgraduate writing diverges lexically from undergraduate writing. However, overall, Durrant concluded that students writing at different levels within the same discipline were homogenous in their vocabulary use.

In a second study, Durrant (2015) examined disciplinary differences in the BAWE corpus using lexical bundles (i.e., four-word sequences or quad-grams). This study focused on 285 authors from 24 different disciplines using discipline specific quad-gram frequency lists for each level. Like the previous study, these frequency lists were specific to BAWE and not lemmatized. When compared with external disciplines, Durrant found that almost all disciplines showed a higher level of overlap internally among the quad-grams for writers across the queried disciplines within BAWE. This finding led Durrant to argue that, within disciplines, there was a high degree of homogeneity in terms of quad-grams. 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 post hoc analysis also revealed differences in quad-gram use between hard sciences (engineering, chemistry, biological sciences) and soft sciences (e.g., law, English, classics).

The current study

The present study extends previous analyses that have investigated differences in macro- and micro-disciplines in student texts. We move beyond simple lexical features and instead examine features related to lexical sophistication, which has been shown to be an important indicator of academic writing (Guo et al. 2013; Kyle and Crossley forthcoming). These features include word properties, corpus frequency, academic language use, and the use of multi-word phrases. In addition, we investigate the role that textual cohesion, which is an important component of larger discourse structures (Loxterman et al. 1994; McNamara et al. 1996), may play in predicting differences between macro- and micro-disciplines. The inclusion of text cohesion indices addresses Flowerdew's (2014, 2015) call to include linguistic features that go beyond lexis.

For this study, we use linguistic features within a Multi-Dimensional (MD) Analysis to investigate differences in macro-disciplines (e.g., science and engineering) and in micro-disciplines within these macro-disciplines (e.g., biology, physics, industrial engineering, and mechanical engineering). We also investigate potential effects of gender, year of study (e.g., undergraduate and graduate students), and paper types (e.g., argumentative essays, proposals, and research papers) on student writing. Our main goal, however, is to examine if differences exist at both the macro- and micro-discipline level in a corpus of student writing. Fleshing out differences at the micro

and macro level may provide important information about differences in student disciplinary writing that may have important implications for discipline-specific pedagogy and student writing expectations.

Method

Corpus

For our analysis, we culled subsets of texts from MICUSP (O'Donnell and Römer 2012; Römer and O'Donnell 2011). MICUSP is a multidisciplinary corpus of successful (all A-graded) student academic writing samples collected at a large American research university, the University of Michigan. The corpus consists of 829 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 the following sixteen 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 and 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). The four selected disciplines (BIO, PHY, MEC, and IOE) make up a MICUSP sub-corpus of 162 papers and just under 470,000 words. Table 9.1 provides an overview of our corpus and its subsets. For each selected discipline, the table reports the number of papers, word counts, and average text length (with standard deviation). Our science sub-corpus (BIO and PHY) consists of 88 papers and about 221,000 words; our engineering sub-corpus (MEC and IOE)

Table 9.1 Details of the MICUSP sub-corpora used in this study

	Number of texts	Mean text length (SD)	Word count
Science sub-corpus			
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 sub-corpus			
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 9.2 Distribution of papers across paper types in MICUSP sub-corpora

	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

contains 74 papers and around 248,000 words. Table 9.2 shows how the 162 MICUSP papers included in our analysis are distributed across paper types. The majority of papers in our sub-corpora have been classified as either reports or research papers.

Parameters

The essays in MICUSP were collected under a variety of different conditions. All identified situational and functional parameters are discussed below.

1. Macro-disciplines. Two macro-disciplines are represented in the corpus: science ($n = 88$) and engineering ($n = 74$).
2. Micro-disciplines. The selected texts from MICUSP can also be divided into four micro-disciplines: biology ($n = 67$), physics ($n = 21$), mechanical engineering ($n = 32$), and industrial engineering ($n = 42$).
3. Gender. Another potential parameter in the MICUSP data is writer gender with 43 percent of the papers ($n = 69$) written by female students, and 57 percent of the papers ($n = 93$) written by male students.
4. Year of Study. There are four different years of study represented in MICUSP as well: final year undergraduate students ($n = 84$), first-year graduate students ($n = 38$), second-year graduate students ($n = 29$), and third-year graduate students ($n = 11$).
5. Paper Type. Six different types of papers are represented within our MICUSP subset. These include argumentative essays ($n = 4$), critiques/evaluations ($n = 7$), proposals ($n = 14$), reports ($n = 69$), research papers ($n = 63$), and response papers ($n = 5$).

Natural language processing tools

We selected two natural language processing tools from the Suite of Automatic Linguistic Analysis Tools (SALAT) developed by Crossley and Kyle. These two tools, discussed in greater detail below, automatically measure linguistic properties related to lexical sophistication and text cohesion.

Tool for the automatic analysis of lexical sophistication

TAALES (Kyle and Crossley 2015) incorporates over one hundred lexical sophistication indices. In this study, we used TAALES 1.3, which includes indices related to word and n-gram frequency (i.e., how many times an item occurs in a reference corpus), lexical range (i.e., how many documents in a reference corpus an item occurs), psycholinguistic word information (e.g., concreteness, familiarity, meaningfulness), and academic language (i.e., items that occur more frequently in an academic corpus than in a general use corpus) for both single words and multi-word units (n-grams, such as bigrams and trigrams).

The frequency and range indices are based on a number of classic and recently compiled corpora (Brown 1984; Brysbaert and New 2009; Kucera and Francis 1967; Svartvik and Quirk 1980; Thorndike and Lorge 1944). Psycholinguistic word information indices draw on the Medical Research Council (MRC) psycholinguistic database (Coltheart 1981), which includes word scores for familiarity, concreteness, imageability, and meaningfulness (how many associations a word has). TAALES also includes indices for recently collected psycholinguistic word information norms including concreteness (Brysbaert et al. 2014) and age of acquisition (i.e., at what age a word is estimated to be learned; Kuperman et al. 2012). Academic language indices are based on the Academic Word List (Coxhead 2000; Pawley and Syder 1983) and the Academic Formulas List (Simpson-Vlach and Ellis 2010). TAALES has been used to examine longitudinal lexical development (Crossley et al. 2016a), and to measure the relationship between lexical sophistication and writing quality (Jung et al. 2015; Kyle and Crossley 2016), speaking proficiency, and written lexical proficiency (Kyle and Crossley 2015).

Tool for the automatic analysis of cohesion

TAACO (Crossley et al. in press) incorporates over 150 classic and recently developed indices related to text cohesion. Included are indices related to type-token ratios, sentence overlap, paragraph overlap, and a variety of connectives (e.g., logical and causal connectives). For a number of indices, the tool incorporates a part of speech (POS) tagger from the Natural Language Tool Kit (Bird et al. 2009) and synonym sets from the WordNet lexical database (Miller 1995). The POS tagger affords examining content words (i.e., nouns, verbs, adjectives, and adverbs) as well as function words (i.e., determiners, propositions). TAACO has been used in a number of studies to model writing quality (Crossley et al. 2016b, c) and creativity (Skalicky et al. 2016).

TAACO provides linguistic counts for both sentence and paragraph markers of cohesion and incorporates WordNet synonym sets. Specifically, TAACO calculates type-token ratio (TTR) indices (for all words, content words, function words, and n-grams), sentence overlap indices that assess local cohesion for all words, content words, function words, POS tags, and synonyms, paragraph overlap indices that assess global cohesion for all words, content words, function words, POS tags, and synonyms, and a variety of connective indices such as logical connectives (e.g., *moreover*, *nevertheless*), conjuncts (*however*, *furthermore*), and causal connectives (e.g., *because*, *therefore*).

Statistical analysis

This study replicates and extends the methods used in Biber's (1988) MD Analysis (see Egbert and Staples this volume; Cantos-Gomez this volume). Following Biber's approach, we first entered the indices reported by TAALES and TAACO into a factor analysis (in this case, a principal component analysis; PCA) using a promax rotation. We selected a PCA over a factor analysis because we were interested in reducing the correlated observed variables reported in factor analysis into a small set of independent composite variables to be used in subsequent statistical analyses. Thus, we used PCA to cluster the indices that co-occurred frequently within the texts into groups, allowing for a large number of variables to be reduced into a smaller set of derived variables (i.e., the components). We used a promax rotation because it is an oblique rotation method that assumes that derived factors are correlated (e.g., lexical components will likely be interrelated). One potential problem with our analysis was corpus size (N = 162). Such a small sample size might be considered problematic because it is under the minimum sample size (N = 300) advocated for a factor analysis by some and the ratio of texts to variables is low (Tabachnick et al. 2012). However, other statisticians have argued that component loadings are more important than sample size. For instance, Guadagnoli and Velicer (1988) maintained that if a component had four or more loadings greater than .6, it was reliable regardless of sample size. Similarly, MacCallum et al. (2001) reported that if all features loaded above .6, then small sample sizes were acceptable. In this study, we follow the suggestion of MacCallum et al. (2001) and set a conservative cutoff for the eigenvalues of $\lambda \geq .60$. This ensured that only salient linguistics indices would be included in the analysis.

Our goal in conducting the PCA was to identify underlying functional interpretations that explain the co-occurrence factors among the indices that load into each component, which might represent common functions of the texts. We presume that the underlying functional interpretations will be related to the parameters discussed above (e.g., micro- and macro-disciplines, gender, year of study, and paper type). To confirm our hypotheses, we computed factor scores for each component. These factor scores allow group comparisons for each factor based on the parameters of interest and can be used to visually demonstrate which parameters load high and low in the dimensions. The factor score was calculated by subtracting the mean of all scores for an index from the score for that index on a specific essay. This value was then divided by the standard deviation of the index across all essays. For each essay, the average for all the indices in the factor was then calculated providing a factor score for the essay. The final factor scores for each essay were then averaged based on the essay conditions discussed above affording the opportunity to interpret the factors in consideration of situational writing parameters.

The components reported by the PCA were initially interpreted using a qualitative analysis of the linguistic indices that clustered in each dimension. To complement the factor scores, we also conducted confirmatory statistical analyses by conducting Analyses of Variance (ANOVAs; see Cantos-Gomez this volume) followed by stepwise discriminant function analyses (DFA; see the chapters by

Cantos-Gomez and by Veirano Pinto this volume) as done initially in Crossley, Varner, and McNamara (2014). Our dependent variable in these analyses were weighted component scores for the indices that loaded into each factor, which is common approach to developing aggregated feature sets (Burstein et al. 2013; Enright and Quinlan 2010). The weighted component scores allowed us to examine a small number of independent components that consisted of correlated features reported by TAALES and TAACO. To create a weighted component score, each index in a component is multiplied by its *eigen* weight. Each multiplied index score in a component is added together to create an overall weighted score. This weighted score is different than the factor loading in that it is an aggregated score across all variables contained within a factor. One value of these weighted scores is that they can easily be computed for new texts not in the original study (in the case of a replication study). Our independent variables were the parameters discussed above. The ANOVAs examined if statistical differences were reported for the TAALES and TAACO indices based on the parameters available.

The DFAs were used to provide evidence that the linguistic indices that load onto each dimension could be used to discriminate the essays based on the parameters selected. Such an approach was first used in Crossley, Varner, and McNamara (2014), but is generally uncommon in MD Analysis. However, a DFA can provide important information about the accuracy of interpreting the factors derived from the factor analysis. Thus, if a dimension was interpreted as representing macro-disciplines, we would conduct a follow-up DFA that treated the macro-discipline as the independent variable and the weighted component score as the dependent variable. In doing so, we can use the DFA to predict the degree of accuracy with which the weighted component score can predict the independent variables affording us the opportunity to confirm if the TAALES and TAACO indices that load into each factor can reliably distinguish the MICUSP parameters.

Results

Statistical assumptions

Prior to conducting the factor analysis, a number of statistical assumptions had to be tested for TAALES and TAACO indices to be included in the analyses. First, the indices needed to demonstrate normal distribution. We tested this by examining the skewness (≤ 2) and kurtosis (≤ 3) of all the selected indices. Those that did not demonstrate normal distribution were removed. We next checked for multicollinearity (i.e., collinearity among variables) for the remaining variables ($r \geq .900$). If multicollinearity was reported among any variables, only one of the variables was included in the data analysis. After controlling for normal distribution and multicollinearity, we were left with forty-eight variables, which helped address any subject to variable ratio problem in the PCA. Twenty-six were from TAALES and twenty-two were from TAACO. From these forty-eight variables, we derived z-scores (see Egbert and Staples this volume) for each variable so all variables were on similar scales.

Table 9.3 First ten eigenvalues from the principal component analysis

Factor number	Eigenvalue	Percent of variance	Cumulative variance
1	8.34	17.375	17.375
2	6.148	12.809	30.184
3	4.548	9.475	39.659
4	2.913	6.069	45.728
5	2.741	5.71	51.438
6	2.228	4.642	56.08
7	1.898	3.953	60.033

Table 9.4 Factor 1 loadings: Ease of function words

Indices	Loadings
SUBTLEXus function word frequency	.853
BNC function word frequency (written)	.778
SUBTLEXus function word range	.759
BNC bigram frequency (written)	.721
Kuperman age of acquisition (function words)	−.696
MRC meaningfulness (function words)	−.652
BNC bigram frequency (written logarithm)	.621

Factor analysis

The eigenvalues for the first seven factors are reported in Table 9.3. In total, the first seven factors accounted for about 60 percent of the shared variance, with component 1 explaining 17 percent of the variance alone. An analysis of the scree plot (see Egbert and Staples this volume), which can be used to find a characteristic break that indicates at which point additional factors explain little additional variance in the analysis (Biber 1988), demonstrated that the clearest break occurred between the fifth and sixth factors, indicating that a five-component/dimension solution is the best interpretation.

The final factor pattern for the five-factor solution included twenty-six indices selected from TAALES and TAACO. The factor loadings for each of the linguistic features in each of the four factors are presented in Tables 9.4 through 9.8.

Dimension analysis

Each factor was loaded into a dimension using factor scores and then interpreted. We then developed component scores based on these factors and used ANOVA and confirmatory DFA analyses to assess the strength of these interpretations. These five dimensions are discussed below.

Table 9.5 Factor 2 loadings: Text simplicity

Indices	Loadings
BNC trigram proportion (written)	.849
BNC frequency (written, all words)	.815
Function word overlap	.756
BNC trigram frequency (spoken)	.671
Positive causal connectives	.627
Kuperman age of acquisition (all words)	−.610

Table 9.6 Factor 3 loadings: Content word frequency

Indices	Loadings
SUBTLEXus Frequency (content words)	.845
Kucera-Francis samples (all words)	.829
Brown Frequency (content words)	.807
Thorndike-Lorge frequency (content words)	.696
MRC familiarity (content words)	.658

Table 9.7 Factor 4 loadings: Word overlap

Indices	Loadings
Adjacent overlap lemmas (all words)	.785
Adjacent overlap lemmas (content words)	.723
Adjacent overlap lemmas (adjectives)	.661
Adjacent overlap lemmas (function words)	.618

Table 9.8 Factor 5 loadings: Function word repetition

Indices	Loadings
Function word TTR	.877
Repeated pronouns	−.826
Bigram TTR	.770
Number of function types	−.711

Dimension 1: Ease of function words

The first factor comprised seven TAALES indices related to the lexical sophistication of function words. However, the factor was reversed such that sophisticated function words loaded lower and less difficult function words loaded higher. The factor explained 17 percent of the total variance (see Cantos-Gomez this volume; Egbert and Staples this volume). When the factor scores were computed for this dimension, the most appropriate

interpretation for the dimension was in distinguishing micro-discipline differences because the dimension separated biology and industrial engineering texts from mechanical engineering and physics texts. The dimension is presented in Figure 9.1.

The linguistic indices that separated MICUSP texts based on these domain differences can be grouped into function word sophistication because the indices include function word and bigram frequency, function word age of acquisition and meaningfulness (negatively correlated), and function word range indices.

An ANOVA and a confirmatory DFA were conducted on the weighted component score developed from the factor. In this analysis, the weighted component score for the dimension was used as the dependent variable to classify the essays according to micro-disciplines. The ANOVA results ($F(3,158) = 17.654$, $p < .001$, see Table 9.9 for descriptive statistics for the weighted component scores) demonstrated a significant difference between micro-disciplines. Pairwise comparison demonstrated that biology texts differed from mechanical engineering and physics texts and industrial

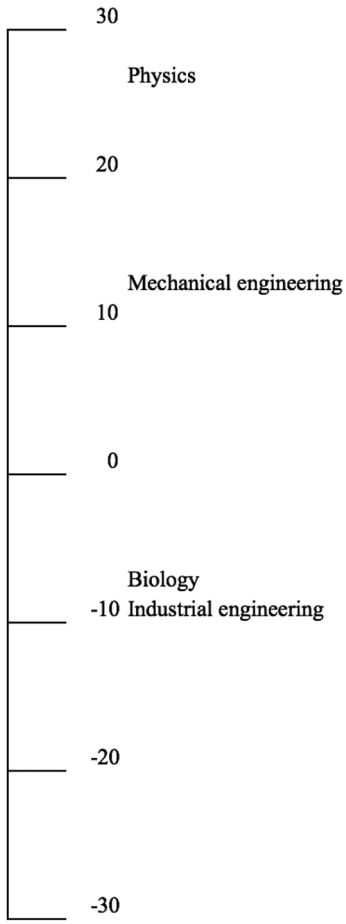

Figure 9.1 Dimension 1 loadings: Ease of function words.

Table 9.9 Descriptive statistics for weighted component 1

Parameter	Mean	Standard deviation
Biology	−1.670	3.460
Industrial engineering	−.618	3.191
Mechanical engineering	2.039	3.153
Physics	3.456	3.423

Table 9.10 Classification results for Dimension 1: Ease of function words

Predicted Group Membership		Biology	Industrial engineering	Mechanical engineering	Physics
Count	Biology	43	9	6	9
	Industrial engineering	21	6	10	5
	Mechanical engineering	5	4	8	15
	Physics	1	2	6	12

engineering texts differed from mechanical engineering and physics texts. Inferential statistics indicated that biology and industrial engineering texts contained more sophisticated function words than physics and mechanical engineering texts. The stepwise DFA retained the weighted component score as significant predictors of micro-disciplines. The results demonstrated that the DFA correctly allocated 69 of the 162 essays in the total set, $\chi 2$ (df = 9, n = 162) = 47.171, $p < .001$, for an accuracy of 42.6 percent (the chance level for this analysis is 25 percent, see the confusion matrix reported in Table 9.10 for results). The measure of agreement between the actual text type and the text type assigned by the model produced a weighted Cohen's Kappa of .370, demonstrating a moderate agreement. The confusion matrix indicates that the ease of function words component was most successful at distinguishing biology texts from other texts.

Dimension 2: Text simplicity

The second component comprised four TAALES indices related to the production of expected trigrams and lexical sophistication of all words and two TAACO indices related to overlap of function words between sentences and the use of causal connectives. The factor explained 13 percent of the total variance. When the factor scores were computed for this dimension, the most appropriate interpretation for the dimension was in distinguishing micro-discipline differences because the dimension separated biology and mechanical engineering texts from industrial engineering and physics texts. The dimension is presented in Figure 9.2.

The linguistic indices that separated the MICUSP texts based on these domain differences tend to lead to more texts that are easier to read. Thus, texts that load high on this dimension include the use of more common trigrams, more frequent words

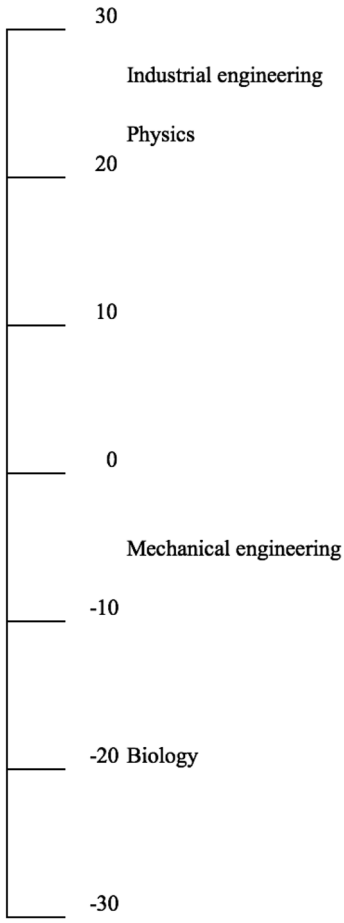

Figure 9.2 Dimension 2 loadings: Text simplicity.

and trigrams, more overlap of function words, more causality, and fewer sophisticated words (i.e., words with lower age of acquisition scores).

An ANOVA and a confirmatory DFA were conducted on the weighted component score developed from the second factor. The ANOVA results ($F(3,158) = 8.786$, $p < .001$, see Table 9.11 for descriptive statistics for the weighted component scores) demonstrated a significant difference between micro-disciplines. Pairwise comparison demonstrated that biology texts differed from industrial engineering and physics texts, and industrial engineering texts differed from mechanical engineering indicating that industrial engineering and physics texts included more features related to text simplicity than mechanical engineering and biology texts. The stepwise DFA retained the weighted component score as significant predictors of micro-disciplines. The results demonstrated that the DFA correctly allocated 76 of the 162 essays in the total set, $\chi 2$ (df = 9, $n = 162$) = 32.382, $p < .001$, for an accuracy of 46.9 percent (the chance level for this analysis is 25 percent, see the confusion matrix reported in Table 9.12 for results). The measure of agreement between the actual text type

Table 9.11 Descriptive statistics for weighted component 2

Parameter	Mean	Standard deviation
Biology	−1.219	3.637
Industrial engineering	1.810	3.271
Mechanical engineering	−.452	2.577
Physics	.959	1.716

and the text type assigned by the model produced a weighted Cohen's Kappa of .251, demonstrating a fair agreement. The model was best at categorizing biology and industrial engineering texts. A post hoc DFA categorizing mechanical engineering and biology texts against physics and industrial engineering texts correctly allocated 107 of the 162 essays in the total set, $\chi 2$ (df = 1, n = 162) = 15.148, $p < .001$, for an accuracy of 66 percent (the chance level for this analysis is 50 percent). The measure of agreement between the actual text type and the text type assigned by the model produced a weighted Cohen's Kappa of .304, demonstrating moderate agreement. These findings indicate differences at the macro level (i.e., science and engineering texts) based on text simplicity.

Dimension 3: Content word frequency

The third component comprised five TAALES indices related to frequency (both content and all words) and word familiarity. The factor explained around 10 percent of the total variance. When the factor scores were computed for this dimension, the most appropriate interpretation for the dimension was in distinguishing micro-discipline differences because the dimension separated mechanical engineering texts from all other domains. The dimension is presented in Figure 9.3.

The linguistic indices that separated the mechanical engineering texts can loosely be grouped into frequency metrics related to frequency including frequency of spoken words found in SUBTLEXus and written frequencies reported in Kucera-Francis and Thorndike-Lorge. The component also included Kucera-Francis samples and an MRC familiarity index.

An ANOVA and a confirmatory DFA were conducted on the weighted component score developed from the third factor. The ANOVA results ($F(3,158) = 6.124$, $p < .001$, see Table 9.13 for descriptive statistics for the weighted component scores) demonstrated a significant difference between micro-disciplines. Pairwise comparison demonstrated that mechanical engineering texts demonstrated significant differences with all other domains such that mechanical engineering texts contained lower scores for word frequency indices. The stepwise DFA retained the weighted component score as significant predictors of micro-disciplines. The results demonstrated that the DFA correctly allocated 46 of the 162 essays in the total set, $\chi 2$ (df = 9, n = 162) = 18.674, $p < .050$, for an accuracy of 28.4 percent (the chance level for this analysis is 25 percent, see the confusion matrix reported in Table 9.14 for results). The measure of agreement between the actual text type and the text type assigned by the model produced a weighted Cohen's Kappa of .075, demonstrating little agreement (i.e., the content word frequency component was not a strong discriminator at the micro-level).

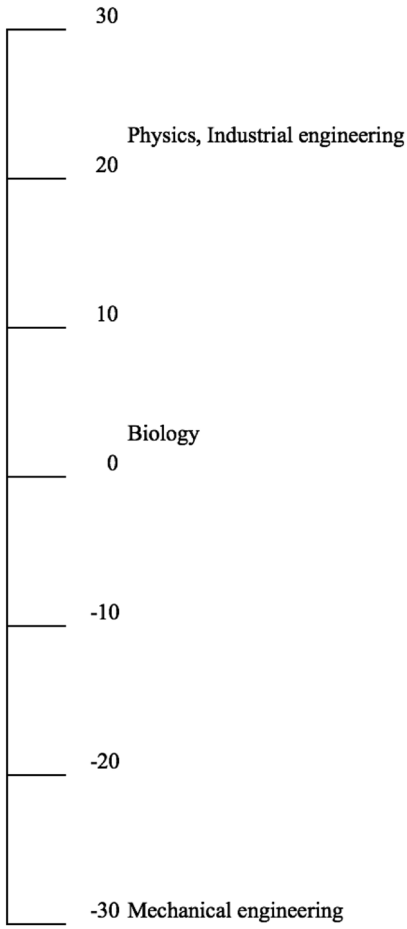

Figure 9.3 Dimension 3 loadings: Content word frequency.

A post hoc DFA categorizing mechanical engineering against all other texts correctly allocated 110 of the 162 essays in the total set, $\chi 2$ (df = 1, n = 162) = 12.629, $p < .001$, for an accuracy of 67.9 percent (the chance level for this analysis is 50 percent). The measure of agreement between the actual text type and the text type assigned by the model produced a weighted Cohen's Kappa of .252, demonstrating fair agreement. The post hoc analysis indicated that the content word frequency component was a strong discriminator of text type at the macro level (i.e., science vs. engineering).

Dimension 4: Word overlap

The fourth component comprised four TAACO indices related to lexical word overlap (both content, function, and all words). The factor explained around 6 percent of the total variance. When the factor scores were computed for this dimension, the most appropriate interpretation for the dimension was in distinguishing gender differences because the dimension separated male and female writers. The dimension is presented

Table 9.12 Classification results for Dimension 2: Text simplicity

Predicted Group Membership		Biology	Industrial engineering	Mechanical engineering	Physics
Count	Biology	39	13	8	7
	Industrial engineering	9	24	5	4
	Mechanical engineering	12	8	8	4
	Physics	3	7	6	5

Table 9.13 Descriptive statistics for weighted component 3

Parameter	Mean	Standard deviation
Biology	.159	3.116
Industrial engineering	.710	2.989
Mechanical engineering	−1.908	2.845
Physics	.981	2.394

Table 9.14 Classification results for Dimension 3: Content word frequency

Predicted Group Membership		Biology	Industrial engineering	Mechanical engineering	Physics
Count	Biology	15	3	23	26
	Industrial engineering	10	2	10	20
	Mechanical engineering	7	3	19	3
	Physics	5	2	4	10

in Figure 9.4. The linguistic indices that separated male and female were related to local cohesion. The indices included sentence overlap for all words, content words, function words, and adjectives.

An ANOVA and a confirmatory DFA were conducted on the weighted component score developed from the fourth factor. The ANOVA results ($F(1,160) = 9.994$, $p < .010$) demonstrated a significant difference between males ($M = -.455$, $SD = 2.169$) and females ($M = .613$, $SD = 1.068$). The stepwise DFA retained the weighted component score as significant predictors of gender. The results demonstrated that the DFA correctly allocated 101 of the 162 essays in the total set, $\chi 2$ (df = 1, $n = 162$) = 9.146, $p < .010$, for an accuracy of 62.3 percent (the chance level for this analysis is 50 percent, see the confusion matrix reported in Table 9.15) for gender assigned by the model produced a weighted Cohen's Kappa of .237, demonstrating fair agreement. The results indicate that females used a greater number of cohesion features between sentences in terms of word overlap than males.

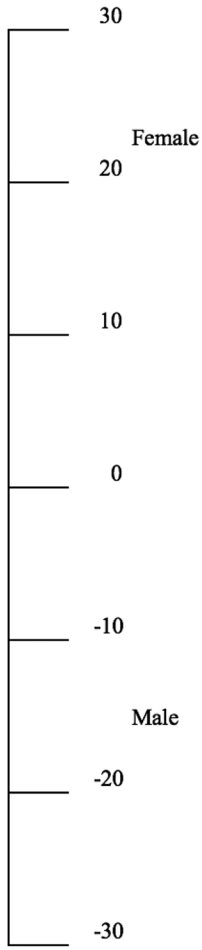

Figure 9.4 Dimension 4 loadings: Word overlap.

Dimension 5: Function word repetition

The fourth factor comprised four TAACO indices related to the repetition of function words. The factor explained 6 percent of the total variance. When the factor scores were computed for this dimension, the most appropriate interpretation for the dimension was in distinguishing macro-disciplines (i.e., engineering from science texts). The dimension is presented in Figure 9.5. The linguistic indices that separated the two macro-disciplines were based on word repetition as found in type-token indices for function words and bigrams and repeated pronouns. In addition, the number function word types loaded in this component. Texts with lower function word repetition loaded positively on the dimension. Hence, the dimension is reverse scaled.

An ANOVA and a confirmatory DFA were conducted on the weighted component score developed from the factor. The ANOVA results ($F(2,160) = 10.092$, $p < .001$) demonstrated a significant difference between science ($M = .588$, $SD = 2.418$)

Table 9.15 Classification results for Dimension 4: Word overlap

Predicted Group Membership		Male	Female
Count	Male	60	33
	Female	28	41

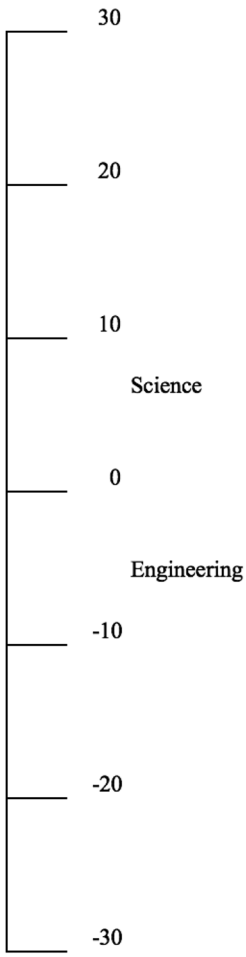

Figure 9.5 Dimension 5 loadings: Function word repetition.

and engineering (M = −.699, SD = 2.745) texts. The stepwise DFA retained the weighted component score as a significant predictor of macro-disciplines. The results demonstrated that the DFA correctly allocated 102 of the 162 essays in the total set, $\chi 2$ (df = 1, n = 162) = 13.966, $p < .001$, for an accuracy of 64.2 percent (the chance level for this analysis is 50 percent, see the confusion matrix reported in Table 9.16 for results). The measure of agreement between the actual text type and the text type

Table 9.16 Classification results for Dimension 5: Function word repetition

Predicted Group Membership		Science	Engineering
Count	Science	52	36
	Engineering	22	52

assigned by the model produced a weighted Cohen's Kappa of .289, demonstrating a fair agreement. The results indicated that science texts, as compared to engineering texts, involved less repetition of function words.

Discussion

The purpose of this study was to examine language differences in student writing using five parameters: macro-disciplines, micro-disciplines, gender, paper type, and year of study. The goal was to examine the potential for lexical and cohesion features to distinguish between these parameters based on the development of component scores derived from an MD Analysis. Our primary interest was in differences between macro- (science and engineering) and micro-disciplines (biology, physics, industrial engineering, and mechanical engineering) because previous research has demonstrated differences at the micro-level for professional writers (Hu and Cao 2015; Ozturk 2007; Ward 2007) and macro-discipline differences at the student level (Durrant 2015; Hardy and Römer 2013), but, to our knowledge, research has not shown differences for both the macro- and micro-level with student writers. Additionally, we wanted to include discourse level features of writing (i.e., text cohesion features) that had not previously been examined in combination with lexical features. Overall, the findings provide evidence that macro-disciplinary and micro-disciplinary differences exist in student writing along with gender differences. However, we find no evidence that differences exist based on paper type or year of study. These findings have important implications for understanding discipline differences.

Our MD Analysis indicated that there were five components that were strongly linked to student writing in MICUSP. These five components explained over 50 percent of the variance in the factor analysis and comprised function word frequency, text simplicity, content word frequency, word overlap, and word repetition. Each of these components was used in an MD Analysis to explore the potential for the components to explain the parameters of interest in the corpus (see Figures 9.1 through 9.5). Further, ANOVA and DFA were conducted to investigate whether the interpretations found in the MD Analysis were significant (i.e., statistically meaningful). The MD Analysis along with the ANOVAs and DFAs indicated that the majority of the factors derived from the factor analysis were related to differences in micro-disciplines. The analyses also supported the notion that gender and macro-disciplines were relevant factors. These analyses are discussed below.

Factor 1

Factor 1 positively loaded a number of linguistic features related to function word and n-gram frequency. In addition, features related to age of acquisition and word meaningfulness for function words loaded negatively. We labeled the factor *Ease of function words* and, when structuring the different parameters based on the MD Analysis, the best interpretation of the resulting dimension was for distinguishing micro-disciplines. The dimension loading indicated that biology and industrial engineering texts contained more sophisticated function words than physics and mechanical engineering texts. As an example, we include excerpts from a physics text and a biology text that reported high component scores for this factor below. In each case, function words are italicized and those function words that are infrequent are also bolded.

1. High Ease of Function Words score
   A viable system *for the* implementation *of a* quantum computer is being actively pursued **by** *a* number *of* research groups *in* varying shapes *and* forms [PHY. G2.03.1].

2. Low Ease of Function Words score

However, unlike many *of the* present models, *it* does allow *for* speciation *in* sympatry **or** parapatry, *as* well *as*, reinforcement *of* allopatrically derived species—**which** have come together *in* secondary contact [BIO.G0.02.4].

Using a weighted component score, we examined if a component comprised of the function word indices demonstrated significant differences between the different micro-disciplines. We found that the component score did demonstrate significant differences and that the component score alone could significantly classify the individual texts according to micro-domain (with an accuracy of 43 percent). These confirmatory analyses provide additional evidence that the MD Analysis interpretation was accurate (i.e., that biology and industrial engineering texts contain more sophisticated function words than physics and mechanical engineering texts).

Factor 2

Factor 2 positively loaded a number of linguistic features related to text simplicity including word frequency, proportion and frequency of trigrams, function word overlap, and the use of causal connectives. In addition, the component loaded an age of acquisition for all words index negatively. We labeled the factor *Text simplicity* because a number of the indices are indicative of less difficult texts such as greater word frequency (McNamara et al. 2010) greater n-gram frequency and proportion scores (Crossley et al. 2012), and greater cohesion (Crossley et al. 2007). When organizing the different parameters based on the MD Analysis, micro-disciplines again gave the clearest interpretation of the resulting dimension. Like in dimension 1, physics loaded highly and biology loaded low. However, in contrast to dimension 1, mechanical engineering texts loaded low and industrial engineering texts loaded high. This indicates that industrial engineering and physics texts included more features related to text simplicity than mechanical engineering and biology texts. As an example, we include

excerpts below from an industrial engineering text, that loaded high on the dimension, and a biology text, which loaded low on the dimension. We focus on trigrams (in bold) as found in the BNC written section.

3. Lower proportion
   Though numerous pharmacologic agents exist to reduce risk factors, **the identification of** novel risk factors and their stimulatory conditions **could lead to** more effective treatments [BIO.G0.16.1].

4. Higher proportion
   Throughout this class we made **the assumption that** the time required **to move the** substrate beneath the insertion point was so small **in comparison to** the time needed to retrieve the element from the correct sleeve **that we were able** to **ignore it** [IOE.G2.03.1].

We next investigated if a weighted component score comprised of the text simplicity indices demonstrated significant differences between the different micro-disciplines. We found that the component score did demonstrate significant differences and that the component score was able to classify the individual texts according to micro-domain with an accuracy of 47 percent. This accuracy was significantly above baseline. These confirmatory analyses provide evidence additional evidence that the MDA interpretation was accurate (i.e., that industrial engineering and physics texts contain more features related to text simplicity than mechanical engineering and biology texts).

Factor 3

Factor 3 positively loaded a number of linguistic features related to content word frequency including word frequency and familiarity indices. Thus, we labeled the factor *Content word frequency*. When examining the MD Analysis loadings, we found similar patterns to Factor 2, supporting an interpretation that favored a micro-discipline separation. Like Factor 2, the dimension loading indicated that physics and industrial engineering texts contained greater scores for word frequency indices, especially when compared to mechanical engineering texts. Biology texts loaded positively, but not strongly on the dimension. The MD Analysis indicates that more frequent words are found in physics and industrial engineering texts as compared to mechanical engineering texts. As an example, we include excerpts from a mechanical engineering text, which scored low in this dimension, and an industrial engineering text that reported high component scores for this dimension. The frequency counts from SUBTLEXus and bolded words are infrequent.

5. Low content word frequency
   Because a **larger percentage** of **Americans use private vehicles** than **walk** or **use public transportation compared** to people in many **countries** in the **European**

Union and around the **world**, **less attention** is **given** in the United States to **pedestrian safety** than in these more **pedestrian-friendly countries** [MEC.G1.03.1].

6. High content word frequency
   We can **immediately** see that if σ is very **large** in **comparison** to the **overall production** time, then it will be in our best **interest** to put all the **boards together** in one **setup** and **manufacture** them **according** to the **optimal setup** for all **boards** [IOE.G2.03.1].

Using a weighted component score, we examined if the indices that loaded into the factor demonstrated significant differences between the different micro-disciplines. We found that the component score did demonstrate significant differences and that the component score significantly classified the individual texts according to micro-domain (with an accuracy of 28 percent). The confusion matrix (see Table 9.14) indicates that the low accuracy resulted from misclassifications of biology texts, which did not show strong trends in the MD Analysis, and from misclassification of industrial engineering texts. These confirmatory analyses provide additional evidence that the MD Analysis interpretation was accurate (i.e., that physics and industrial engineering text contain more frequent words than mechanical engineering texts).

Factor 4

Factor 4 positively loaded a number of linguistic features related to word overlap (i.e., all words, content words, adjectives, and function word). All the indices that loaded into this factor were related to local cohesion and assessed overlap between sentence and not larger text segments like paragraphs. We labeled the factor *Word overlap* and the MD Analysis of the different parameters indicated that the best interpretation of the resulting dimension was in distinguishing genders. The dimension loadings demonstrated that female writers tended to use greater word overlap between sentences than male that writers. As an example, we include excerpts from a text written by a female and a male that reported both high and low component scores respectively for this factor. Words and phrases that are repeated are in bold.

7. High word overlap across sentences
   Intracellular electric fields are of great interest because of the important **role** they **play in** the biological **processes** of a **cell**. It is known that **intracellular electric fields play** a major **role in** sub-**cell processes** like **cell** division [BIO.G3.03.1; female student paper].

8. Low word overlap across sentences
   DNA in query **will** hybridize to **the** matching probe on microarray, if there is, otherwise it **will be** washed away. After hybridization and washing, signals from **the** label of hybridized **DNA will be** detected and recorded by reader instruments such as photometer [BIO.G2.05.1; male student paper].

Using a weighted component score, we examined differences between male and female writers in the disciplines selected from MICUSP. We found that the component score did demonstrate significant differences and that the component score alone could significantly classify the individual texts according to gender (with an accuracy of 62 percent). These confirmatory analyses provide evidence that the MD Analysis interpretation was accurate (i.e., that females use more word overlap at the sentence level than males).

Factor 5

Factor 5 positively loaded two linguistic features related to type-token ratios: function word TTR and bigram TTR. The factor also negatively loaded indices related to the number of function types and the number of pronouns. We labeled the factor *Function word repetition* because function word repetition was high. When structuring the different parameters based on the MD Analysis, the best interpretation of the resulting dimension was in distinguishing macro-disciplines. The dimension loading indicated that science texts contained less function word repetition, while engineering texts included greater function word repetition. As an example, we include excerpts from a science text (biology) that loaded high on the dimension (and thus had lower function word repetition) and an engineering text (industrial engineering) which loaded low on the dimension (and thus had greater function word overlap).

9. Low function word repetition
   However, since **the** casual relationship is often unclear, **the** authors argue prudence when exploring speciation events based **on these** traits [BIO.G0.02.3].
10. High function word repetition
    The components **are** stored **in** sleeves **in a** pipe-organ setup above **the** area where **the** substrates **are** held, with **a** retrieval arm **that** moves between **the** sleeves **to** pick **the** correct component [IOE.G2.03.1].

Using a weighted component score, we examined if the function word repetition indices demonstrated significant differences between the different macro-disciplines. We found that the component score did demonstrate significant differences and that the component scores could significantly classify the individual texts according to macro-domain (with an accuracy of 64 percent). These confirmatory analyses provide evidence that the MD Analysis interpretation is accurate (i.e., that science texts contain fewer function word repetitions than engineering texts).

Implications

Perhaps the most important message to take away from this analysis is that the presumed homogeneity within structured disciplines such as science and engineering is not as clear-cut as thought. The first three factors, which explained almost 40 percent of the variance in the factor analysis, were best explained in terms of difference in

micro-disciplines and not macro-discipline. For instance, physics and mechanical engineering texts grouped together in Factor 1 because both showed less sophistication in function word properties than biology and industrial engineering texts. However, physics and industrial engineering texts grouped together in Factor 2 because both showed more features related to text simplicity as compared to mechanical engineering and biology texts. In terms of content word frequency, physics and industrial engineering texts grouped together because they contained more frequent words in contrast to mechanical engineering texts, which contained less frequent words.

Thus, we see unique linguistic profiles that arise from texts taken from each micro-discipline. We also see some similarities in the profiles across disciplines such as in the case of physics and industrial engineering texts in the second and third factors. The component scores derived from the factor analyses indicate that these interpretations are realized statistically suggesting that there are differences within disciplines, which may be perceived, from the outside, to be similar in nature. The findings from the micro-discipline analyses support anecdotal and qualitative evidence which indicates that both faculty and students within micro-disciplines perceive differences between their micro-disciplines and closely related macro-disciplines (Bazerman and Paradis 1991).

Additionally, the large differences reported between macro-disciplines in previous studies (Durrant 2015; Hardy and Römer 2013) are not fully supported in this analysis. We see in at least one factor (the repetition of function words) that macro-disciplines can be classified based on linguistic features related to lexical sophistication and cohesion, but the factor is relatively weak in terms of the amount of variance it explains (around 6 percent). This does not call into question previous analyses that used different linguistic features, but does indicate that previous findings suggesting that micro-disciplines are similar to the macro-disciplines with which they are associated bears reassessment, especially in terms of varying linguistic constructs.

Future research

The findings on micro-discipline differences seem to indicate that students are aware of disciplinary conventions and expectations at a more granular level. However, more research is needed to support this claim. We envision a number of additional NLP analyses of upper-level student writing from a greater number of disciplines (e.g., humanities, math, linguistics) as well as similar analyses of professional writing in the disciplines. A first candidate for additional student writing analyses would be the British Academic Written English (BAWE) corpus (Nesi and Gardner 2012). BAWE is larger than MICUSP with a greater number of student texts and disciplines. The disciplines overlap with MICUSP, but include different families.

Conclusion

This study provides evidence for both macro- and micro-discipline difference within science and engineering texts written by advanced students. In addition, the study

indicates that gender differences may exist within science and engineering disciplines. Future studies, as discussed above, are warranted in order to validate these findings on larger corpora and a greater range of disciplines. Larger corpora would allow for greater generalizations to be made and provide confidence that the findings reported in this study can be extended to a larger sample of writing.

Acknowledgments

The first author is indebted to the English Department at Iowa State University for hosting the *Symposium* on *Linguistic Perspectives on Discipline Specific Language* in 2014 from which the ideas in this chapter surfaced. Our thanks also to David Russell at Iowa State for talking through some early theoretical issues with us.

References

Ädel, A. and U. Römer (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.

Alsop, S. and H. Nesi (2009), "Issues in the Development of the British Academic Written English (BAWE) Corpus," *Corpora*, 4 (1): 71–83.

Bazerman, C. and J. G. Paradis, eds (1991), *Textual Dynamics of the Professions: Historical and Contemporary Studies of Writing in Professional Communities*, Madison: University of Wisconsin Press.

Biber, D. (1988), *Variation across Speech and Writing*, New York: Cambridge University Press.

Biber, D., S. Conrad, R. Reppen, P. Byrd, M. Helt, V. Clark, V. Cortes, E. Csomay and A. Urzua (2004), *Representing Language Use in the University: Analysis of the TOEFFL 2000 Spoken and Written Academic Language Corpus*, TOEFL Monograph Series, Princeton: Educational Testing Services.

Bird, S., E. Klein and E. Loper (2009), *Natural Language Processing with Python*, Sebastopol: O'Reilly.

Brown, G. D. (1984), "A Frequency Count of 190,000 Words in the London-Lund Corpus of English Conversation," *Behavior Research Methods, Instruments, & Computers*, 16 (6): 502–32.

Brysbaert, M. and B. New (2009), "Moving Beyond Kučera 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): 977–90

Brysbaert, M., A. B. Warriner and V. Kuperman (2014), "Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas," *Behavior Research Methods*, 46 (3): 904–11.

Burstein, J., J. Tetreault and N. Madnani (2013), "The E-Rater Automated Essay Scoring System," in M. D. Shermis and J. Burstein (eds), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, 55–67, New York: Routledge.

Coltheart, M. (1981), "The MRC Psycholinguistic Database," *The Quarterly Journal of Experimental Psychology*, 33 (4): 497–505.

Cortes, V. (2004), "Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology," *English for Specific Purposes*, 23 (4): 397–423.
Coxhead, A. (2000), "A New Academic Word List," *TESOL Quarterly*, 34 (2): 213–38.
Crossley, S. A., Z. Cai and D. S. McNamara (2012), "Syntagmatic, Paradigmatic and Automatic N-Gram," in P. M. McCarthy and G. M. Youngblood (eds), *Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference*, 214–19, Menlo Park: AAAI Press.
Crossley, S. A., D. F. Dufty, P. M. McCarthy and D. S. McNamara (2007), "Toward a New Readability: A Mixed Model Approach," *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, 197–202.
Crossley, S. A., K. Kyle and D. S. McNamara (2016a), "The Tool for the Automatic Analysis of Text Cohesion (TAACO): Automatic Assessment of Local, Global and Text Cohesion," *Behavior Research Methods*, 48 (4): 1227–37.
Crossley, S. A., K. Kyle and T. Salsbury (2016b), "A Usage-Based Investigation of L2 Lexical Acquisition: The Role of Input and Output," *The Modern Language Journal*, 100 (3): 702–15.
Crossley, S. A., K. Muldner and D. S. McNamara (2016c), "Idea Generation in Student Writing Computational Assessments and Links to Successful Writing," *Written Communication*, 33 (3): 328–54.
Crossley, S. A., L. K. Varner and D. S. McNamara (2014), "A Multi-Dimensional Analysis of Essay Writing: What Linguistic Features Tell us about Situational Parameters and the Effects of Language Functions on Judgments of Quality," in T. Berber Sardinha and M. Veirano Pinto (eds), *Multi-Dimensional Analysis, 25 Years On: A Tribute to Douglas Biber*, 199–240, Amsterdam/Philadelphia, PA: John Benjamins.
De Chazal, E. (2013), *Oxford Handbooks for Language Teachers: English for Academic Purposes*, Oxford: Oxford University Press.
Durrant, P. (2014), "Discipline and Level Specificity in University Students' Written Vocabulary," *Applied Linguistics*, 35 (3): 328–56.
Durrant, P. (2015), "Lexical Bundles and Disciplinary Variation in University Students' Writing: Mapping the Territories," *Applied Linguistics*, 38 (2): 165–93.
Enright, M. K. and T. Quinlan (2010), "Complementing Human Judgment of Essays Written by English Language Learners with E-Rater® Scoring," *Language Testing*, 27 (3): 317–34.
Ewer, J. R. and G. Latorre (1969), *A Course in Basic Scientific English*, London: Longman.
Flowerdew, L. (2014), "Learner Corpus Research in EAP: Some Key Issues and Future Pathways," *English Language and Linguistics*, 20 (2): 43–60.
Flowerdew, L. (2015), "Learner Corpora and Language for Academic and Specific Purposes," in S. Granger, G. Gilquin and F. Meunier (eds), *The Cambridge Handbook of Learner Corpus Research*, 465–84, Cambridge: Cambridge University Press.
Guadagnoli, E. and W. F. Velicer (1988), "Relation to Sample Size to the Stability of Component Patterns," *Psychological Bulletin*, 103 (2): 265–75.
Guo, L., S. A. Crossley and D. S. McNamara (2013), "Predicting Human Judgments of Essay Quality in both Integrated and Independent Second Language Writing Samples: A Comparison Study," *Assessing Writing*, 18 (3): 218–38.
Hardy, J. A. and U. Römer (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.
Hu, G. and F. Cao (2015), "Disciplinary and Paradigmatic Influences on Interactional Metadiscourse in Research Articles," *English for Specific Purposes*, 39 (1): 12–25.

Hyland, K. (2002), "Directives: Argument and Engagement in Academic Writing," *Applied Linguistics*, 23 (2): 215–39.

Hyland, K. and P. Tse (2009), "'The Leading Journal in Its Field': Evaluation in Journal Descriptions," *Discourse Studies*, 11 (6): 703–20.

Jung, Y., S. A. Crossley and D. S. McNamara (2015), "Linguistic Features in MELAB Writing Performances," *Cambridge Michigan Language Assessments*, 5 (1): 1–17.

Kucera, H. and N. Francis (1967), *Computational Analysis of Present-Day American English*, Providence: Brown University Press.

Kuperman, V., H. Stadthagen-Gonzalez and M. Brysbaert (2012), "Age-of-Acquisition Ratings for 30,000 English Words," *Behavior Research Methods*, 44 (4): 978–90.

Kyle, K. and S. A. Crossley (2015), "Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application," *TESOL Quarterly*, 49 (4): 757–86.

Kyle, K. and S. A. Crossley (2016), "The Relationship between Lexical Sophistication and Independent and Source-Based Writing," *Journal of Second Language Writing*, 34 (4): 12–24.

Kyle, K. and S. A. Crossley (2017), "Assessing Syntactic Sophistication in L2 Writing: A Usage-Based Approach," *Language Testing*, 34 (4): 513–35.

Loxterman, J. A., I. L. Beck and M. G. McKeown (1994), "The Effects of Thinking Aloud during Reading on Students' Comprehension of More or Less Coherent Text," *Reading Research Quarterly*, 29 (4): 353–67.

MacCallum, R. C., K. F. Widaman, K. J. Preacher and S. Hong (2001), "Sample Size in Factor Analysis: The Role of Model Error," *Multivariate Behavioral Research*, 36 (4): 611–37.

McNamara, D. S., S. A. Crossley and P. M. Mccarthy (2010), "Linguistic Features of Writing Quality," *Written Communication*, 27 (1): 57–86.

McNamara, D. S., E. Kintsch, N. B. Songer and W. Kintsch (1996), "Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning from Text," *Cognition and Instruction*, 14 (1): 1–43.

Miller, G. A. (1995), "Wordnet: A Lexical Database for English," *Communications of the ACM*, 38(11): 39–41.

Nesi, H. and S. Gardner (2012), *Genres across the Disciplines: Student Writing in Higher Education*, Cambridge: Cambridge University Press.

O'Donnell, M. B. and U. Römer (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.

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.

Pawley, A. and F. H. Syder (1983), "Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency," in J. C. Richards and R. W. Schmidt (eds), *Language and Communication*, 191–225, London/New York: Routledge.

Römer, U. and M. B. O'Donnell (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): 159–77.

Simpson-Vlach, R. and N. C. Ellis (2010), "An Academic Formulas List: New Methods in Phraseology Research," *Applied Linguistics*, 31 (4): 487–512.

Skalicky, S., C. M. Berger, S. A. Crossley and D. S. McNamara (2016), "Linguistic Features of Humor in Academic Writing," *Advances in Language and Literary Studies*, 7 (3): 248–59.

Svartvik, J. and R. Quirk (1980), *A Corpus of English Conversation*, Sweden: Gleerup.
Tabachnick, B. G., L. S. Fidell and S. J. Osterlind (2012), *Using Multivariate Statistics*, New York: Pearson.
Thorndike, E. L. and I. Lorge (1944), *The Teacher's Wordbook of 30,000 Words*, New York: Columbia University, Teachers College Bureau of Publications.
Ward, J. (2007), "Collocation and Technicality in EAP Engineering," *Journal of English for Academic Purposes*, 6 (1): 18–35.