Publication type and discipline variation in published academic writing

Investigating statistical interaction in corpus data

Jesse Egbert
Brigham Young University

This study uses Multi-Dimensional analysis to describe linguistic variation in a corpus of published academic writing across three publication types in two disciplines. The resulting five dimensions were labeled: “Affective synthesis versus specialized information density”, “Definition and evaluation of new concepts”, “Author-centered stance”, “Reader-friendly narrative”, and “Abstract observation and description”. Factorial ANOVAs were used to test for significant interactions between publication type and discipline on each of the five linguistic dimensions. Statistical interactions were discovered for four of the five dimensions. The appropriate tests for statistical differences, either for main effects or simple effects, were performed, and publication type and discipline patterns were interpreted for all five dimensions. This paper highlights the importance of accounting for all of the independent factors in a corpus, using factorial ANOVAs where appropriate, in order to appropriately analyze and interpret patterns of linguistic variability.

Keywords: register, discipline, academic writing, statistical interaction, Multi-Dimensional analysis

1. Introduction

The primary purpose of this study is to describe the linguistic characteristics of published academic writing across six registers, or three publication types (journal articles, university textbooks, and popular academic books) in two disciplines (biology and history). This description is both quantitative, including the results of a Multi-Dimensional analysis and a series of ANOVAs and corresponding post hoc tests, and qualitative, including interpretations that are based on situational differences among three publication types and between two disciplines.
My secondary goal is to explore two models of analyzing linguistic variability in published academic writing. In the 'Register' model, linguistic variation in the corpus is analyzed along a single variable — register — that accounts for all of the situational characteristics of published academic writing. In the 'Publication Type x Discipline' model, I attempt to account for the same linguistic variation by measuring the separate effects of two situational variables: publication type and discipline. This second model allows us to tease apart the differential effects of publication type and discipline, two situational variables that have emerged as important sources of variability in studies of academic language.

Although the 'Register' and 'Publication Type x Discipline' models each reveal important and complementary insights about the nature of published academic writing, this study suggests that factorial models, such as the 'Publication Type x Discipline' model, are more appropriate in cases where multiple independent and crossed variables exist in a corpus sample. Factorial designs are seldom used in corpus linguistic research despite the relatively common presence of crossed independent variables in corpus designs. Overlooking important interactions can result in incomplete interpretations, at best, and inaccurate conclusions, at worst. In this article, I will discuss these issues and demonstrate appropriate methodological procedures for analyzing corpus data with multiple factors.

Section 1.2 includes a broad overview of the state of the art in corpus-based research on register variation. Section 1.3 focuses specifically on linguistic variation within published academic writing.

1.2 Register variation

Corpus linguistic research has established register as a key predictor of linguistic variation (see e.g. the survey in Atkinson & Biber 1994). However, Biber (2012) recently has noted that register differences are often disregarded by authors of reference works and linguistic research on grammatical and lexico-grammatical patterns in English. These authors attempt to describe “general English”, ignoring wide variation in patterns of English language use across situational contexts. Descriptions of “general English” are commonly based on a corpus designed to represent the full range of speech and writing in English. Patterns of language use are then identified within the entire corpus without reference to variation within the corpus.

An analysis of nouns and verbs in Davies’ (2008-) Corpus of Contemporary American English (COCA) will suffice to illustrate some of the problems that can result from ignoring register variation. Taken as a whole, COCA seems to show that nouns are used more than verbs in “general English”. However, when we compare the rates of occurrence for nouns and verbs in speech to those in academic prose, we discover that nouns are much more common than verbs in academic
texts, but the opposite is true in speech (see Table 1). An important point to notice here is that ignoring register differences not only causes us to miss important information about register variation in the use of nouns and verbs. It also produces a misleading representation of the use of these two features in English. The “general English” findings do not accurately represent any variety of English. These numbers are central tendencies calculated based on the full range of variability in the corpus, and they fall either well above or well below the rates of occurrence for each of the five major register categories in COCA.

It is clear that register differences should, at the very least, be considered as potential predictors of linguistic variation. However, this raises questions about the most appropriate level of granularity, or specificity, in defining register categories. Is it enough to distinguish academic writing from other registers or is it necessary to classify texts at a more specific level (e.g. journal articles vs. textbooks)? Gries (2006) addresses questions such as this in a study on linguistic variability within and between corpora. He demonstrates a number of robust statistical techniques that can be used to measure the degree of variability within a corpus, and concludes that the degree of granularity (i.e. register, sub-register, etc.) is closely related to the variability within a sample of texts. As with any sample of data, we expect the variance to be larger within a less homogenous sample (e.g. register) and smaller within a more homogenous sample (e.g. sub-register). One important finding from Gries’ (2006) study is that registers (or sub-registers) differ not only in their rates of occurrence for a given feature but also in the degree of variability in the use of that feature across the texts in the sample.

This supports the results of Biber’s (1988) Multi-Dimensional analysis which revealed a narrow range of variation within certain registers (e.g. official documents) and a much wider amount of variability within other registers (e.g. academic prose) (Biber 1988: 176). These findings, particularly with regard to the large range of variation in academic writing, were characterized by Biber (1988: 178) as “surprising and contrary to popular expectation”. However, subsequent empirical research during the past two and a half decades has attributed much of this variability to the many publication types and disciplines within the general register of academic writing.

Table 1. Normed counts of nouns and verbs in “General English”, speech, and academic writing

<table>
<thead>
<tr>
<th></th>
<th>“General”</th>
<th>Speech</th>
<th>Academic writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>215.57</td>
<td>163.93</td>
<td>258.90</td>
</tr>
<tr>
<td>Verbs</td>
<td>181.20</td>
<td>207.88</td>
<td>148.24</td>
</tr>
</tbody>
</table>
1.3 Variability within published academic writing

There is a growing body of research focused on linguistic variation within academic writing. Many of these studies have analysed linguistic variation across publication types. For example, Biber (2006b) shows large differences between university textbooks, course management writing (e.g. syllabi), and institutional writing (e.g. university catalogs) in their use of many linguistic features (e.g. lexical bundles, vocabulary, stance features). Hyland (2006) finds important differences between research articles and popular science articles in the authors’ use of proximity (e.g. stance, engagement). Finally, Biber & Gray (2010) have also shown that specialist science writing uses more nouns, in general, and pre-modifying nouns, specifically, than multi-disciplinary science writing.

Linguistic variation across disciplines has also received a great deal of attention in recent years. Each of the studies introduced above included a description of linguistic variation across disciplines in their corpus. Biber (2006b) shows variation in the use of lexical bundles in university textbooks, both in terms of frequency and function. Hyland (2010) shows a large amount of variability in the most frequent reporting verbs used in journal articles across eight disciplines. Finally, the corpus data in Biber & Gray (2013) reveals that pre-modifying nouns are more common in science research articles than in research articles in the social sciences and humanities.

Studies of individual linguistic features, such as those mentioned in the two preceding paragraphs, have revealed meaningful differences between registers, publication types, and disciplines of academic writing. This body of research is complemented by investigations of multivariate patterns of linguistic variation within the macro-register of academic writing, such as Biber’s (1988) Multi-Dimensional (MD) analysis. In MD analysis, factor analysis is applied to the normed rates of occurrence of linguistic variables in a corpus in an effort to reduce these variables down to a much smaller set of functionally interpretable dimensions of linguistic variation (see Introduction to Conrad & Biber 2001). MD analyses have revealed sizeable differences between academic publication types (Biber 2006b, Grabe 1984, Conrad 1996b), across sections of medical research articles (Biber & Finegan 2001), as well as across disciplines in journal articles (Gray 2011), textbooks (Carkin 2001), and popular science writing (Grabe 1984).

Despite this body of research, there is a great deal we do not yet understand about variability within the macro-register of academic writing. On this note, Biber (2006b: 227) calls for additional studies that investigate “particular university registers at a much more specified level,” and Hyland (2008: 20) suggests that more cross-discipline comparisons are needed to create “a fuller picture of community-specific practices”.
In addition to the need for further investigations of the independent effects of situational variables such as publication type and discipline, additional research is needed to investigate statistical interactions between these and other factors. In language research, an interaction exists between two factors when the linguistic patterns in a particular level of one factor are determined by the level of another factor. In other words, the two factors are not independent of each other or they are non-additive.

There is a fair amount of research that has shown the existence of interacting factors in corpus-based studies of academic writing. For example, Biber et al. (in press) identify a strong register (news, personal letters, science articles) by time (1650–1999) interaction in the use of genitives and pre-modifying nouns in the ARCHER corpus. Biber et al. (2014) find that, for many linguistic variables associated with grammatical complexity, mode (spoken, written) interacted significantly with task (independent, integrated) in TOEFL iBT responses. Csomay (2007) identifies statistical interactions between discipline (business, education, engineering, humanities, natural sciences, social sciences) and speaker (teacher, student) and discipline and level of instruction (lower division, upper division, graduate). Finally, using a research approach similar to the one adopted in the present study, Conrad (1996a) applies Biber’s (1988) MD analysis framework and reports significant interaction effects between discipline (ecology, history) and publication type (research articles, textbooks) on two of the five dimensions. However, interaction between factors is not the focus of Conrad’s (1996a) study. Therefore, her interpretations are mostly focused on main effect differences.

These studies show that factors such as time, mode, publication type, and discipline can interact in meaningful ways. However, the possibility of interaction is often overlooked in analyses of corpora that contain multiple factors. In studies where statistical tests are used to measure interaction, inappropriate subsequent analyses are sometimes carried out (e.g. testing for main effects instead of simple effects in the presence of an interaction), thus affecting the interpretations of the interaction effects. Finally, in studies where these analyses are appropriately carried out, the qualitative interpretation of interaction effects and corresponding main effects or simple effects often lacks the detail and depth necessary for a complete understanding of the relationships between the interacting factors.

2. Corpus compilation

The corpus in this study is a balanced collection of texts from three academic publication types: journal articles, university textbooks, and popular academic books (see Table 2). For the purposes of this study, journal articles are defined
as intraspecialist research reports published in peer-reviewed academic journals. University textbooks are defined as pedagogical books written for students at the undergraduate or graduate university levels. Popular academic books are defined broadly as non-fiction books written for a general, non-specialist audience on scientific or academic topics.

One reason for selecting these particular publication types is that they span at least two separate situational clines. The first is a cline of intended audiences that ranges from non-expert to expert. The level of expected reader expertise and specialization of knowledge will also affect the size and homogeneity of the intended audience. For example, we would expect readers of peer-reviewed micro-biology articles to be relatively small and homogeneous compared with the readers of an introductory micro-biology textbook. Likewise, we would expect an even larger and more heterogeneous intended audience for a popular academic book on topics within the same discipline.

Another situational cline that can be used to describe publication types of academic writing is Cloitre & Shinn’s (1985) four-stage continuum, which consists of (i) intraspecialist, (ii) interspecialist, (iii) pedagogical, and (iv) popular (Cloitre & Shinn 1985). The main distinction made between the first two stages is that intraspecialist articles are written to a more narrow audience of experts who specialize in a particular area of research, whereas interspecialist articles are written to a broader, more interdisciplinary audience. While there are subtle differences between these two groups, they share many things in common. Most importantly, they are both written by specialists for other specialists. In this study, intraspecialist and interspecialist writing will be grouped together within the publication type of journal articles, and the pedagogical and popular stages will be represented by university textbooks and popular academic books, respectively.

Within each of these three publication types I have sampled an equal number of texts from two disciplines: biology and history. These two disciplines were chosen to represent some of the variability present in academic writing. Whereas biology is a natural science, history is traditionally classified with the humanities. Moreover, Conrad (1996a: 59) points out that history and biology “have different research and methodological traditions”. Together, the situational variables of

<table>
<thead>
<tr>
<th></th>
<th>Biology</th>
<th>History</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Popular academic</td>
<td>25 (14,306)</td>
<td>25 (14,349)</td>
<td>50 (28,655)</td>
</tr>
<tr>
<td>Textbooks</td>
<td>25 (13,875)</td>
<td>25 (14,042)</td>
<td>50 (27,917)</td>
</tr>
<tr>
<td>Journal articles</td>
<td>25 (14,067)</td>
<td>25 (14,269)</td>
<td>50 (28,336)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>75 (42,248)</td>
<td>75 (42,660)</td>
<td>150 (84,908)</td>
</tr>
</tbody>
</table>

* Word counts are in parentheses. See Egbert (2014) for a complete list of source texts
publication type and discipline in this study can be combined to form a single situational variable — register — with 6 levels.

The journal article samples were selected from recent (2005–2012), peer-reviewed academic journals. Previous research has identified important linguistic differences across the four traditional journal article sections (introduction, methods, results, discussion), especially in quantitative research reports, such as those published in biology journals (see e.g. Biber & Finegan 2001). Therefore the journal article excerpts were selected in a stratified manner from across the various sections of the journal articles. Textbook samples were selected from recent (1999–2012) university textbooks ebrary.com or myilibrary.com, two electronic book databases. Textbook excerpts were selected from a variety of chapters and locations within chapters in order to avoid a sampling bias. The popular academic text samples were selected from the New York Times Books website which contains free sample passages of books reviewed by the New York Times. A notable limitation of this study is that all popular academic passages were taken from the first or second chapter of the book.

Each text sample in the corpus is relatively short, being only 500–600 words long, with an overall mean of 566 words. The results presented in this paper are part of a larger study that included an analysis of readers’ perceptions of writing quality and style. The perceptual aspect of that study necessitated the use of relatively short texts. While the short text samples can be seen as a limitation of this study, Biber (1993) finds that frequent lexico-grammatical features were quite stable in texts even shorter than this. In an effort to investigate the “optimal text length,” Biber (1993:248) uses texts from the LOB and the London-Lund corpus to measure the stability of a variety of linguistic features across 200-word text segments. He summarizes his findings by stating: “Common linear linguistic features are distributed in a quite stable fashion within texts and can thus be reliably represented in relatively short text segments” (Biber 1993:252). Furthermore, Biber (1993) reports that most grammatical features “occur in the first 200 words, with relatively few grammatical categories being added after 600 words” (Biber 1993: 251). These findings offer strong support for the length of the texts used in this study. However, the representativeness of the specific texts in this corpus is an empirical question.

In order to determine the extent to which the 500–600 word texts in this corpus represent the registers they were sampled from, I compared the linguistic patterns in this corpus with those in two sub-corpora from Conrad’s (1996a) dissertation study that were designed to represent the exact same registers (university textbooks in history and journal articles in biology). Conrad’s (1996a) corpus contained 3,200 word samples of journal articles and 5,000 word samples from university textbooks, both from two disciplines: biology, specifically ecology, and
history, specifically American history. Conrad (1996a) presents the results of an
MD analysis, using the dimension structures of Biber’s (1988) first five dimen-
sions. I performed the same analysis on the texts in my corpus, using the same
set of linguistic features to calculate dimension scores for each text and mean di-
mension scores for each register. The results from my analyses can be compared
with Conrad’s (1996a) results for university textbooks in history in Figure 1 and
journal articles in biology in Figure 2.

![Figure 1](image1.png)
**Figure 1.** Mean Biber (1988) dimension scores for the history university textbook sub-
corpora in Conrad (1996a) and the present study

![Figure 2](image2.png)
**Figure 2.** Mean Biber (1988) dimension scores for the biology journal article sub-corpo-
ra in Conrad (1996a) and the present study
The results presented in Figures 1 and 2 reveal that the patterns of linguistic variation across the various dimensions are almost identical between the corpus in this study and Conrad’s (1996a) corpus, which is composed of texts that are much longer than those in my corpus. This lack of deviation between the two corpora offers strong support for the representativeness of the corpus used in this study. The results of these case studies combined with Biber’s (1993) evidence for the representativeness of short texts are sufficient to support the use of 500–600 word texts samples in this corpus.

3. Corpus analysis

Sixty-two linguistic features were included in this analysis. Fifty-six of these features were selected based on previous MD analyses of academic writing, such as Conrad (1996a), Biber (2006a), and Gray (2011). Six additional features were included in this study because of their hypothesized prevalence in published academic writing. These six features include nouns as pre-nominal modifiers, and the percent of the text comprised of Davies & Gardner’s (2013) Academic Vocabulary List and Core Vocabulary List (1–500 and 501–3000 words). This study also included phrasal verbs and lexical bundles, or commonly occurring four word sequences. The phrasal verbs were those that appeared in the lists of frequent phrasal verbs reported in Gardner & Davies (2007) and Biber et al.’s (1999) Longman Grammar of Spoken and Written English. The academic lexical bundles were those included on Simpson-Vlach & Ellis’s (2010) Academic Formulas List and Biber et al.’s (2004) list of frequent bundles in university textbooks.

The Biber Tagger (see Biber 1988:211–245) was used to grammatically annotate each of the texts for the majority of the linguistic features. Perl programs were written to count noun-noun sequences, phrasal verbs, and lexical bundles. Wordandphrase.info was used to calculate the percent of academic and core vocabulary. The final 62 features selected for inclusion in this study were counted for each text and, with the exception of the vocabulary percentages, their frequencies were normed to per 1,000 words.

Using the normed rates of occurrence for this set of 62 features in each of the texts, factor analysis was performed in an effort to reduce this large set of linguistic variables down to a smaller set of interpretable underlying dimensions. The statistical software R (R Development Core Team 2012) was used to perform the factor analysis procedure. This was done using the R function ‘fa’ (factor analysis) within the ‘psych’ library, with a principal axis factoring method and a Promax rotation (Revelle 2012). A Promax rotation was used because it allows for minor correlations among the factors, which is to be expected with linguistic data (Biber
The scree plot of eigenvalues showed a clear break between factors 6 and 7; thus, a six factor solution was used. However, it was determined that only the first five dimensions were interpretable. The cumulative percentage of shared variance accounted for by these five factors was 31%. Variables were only included in the analysis if they achieved a minimum factor loading threshold of +/−.30, and each variable was included in the factor where it loaded the strongest (see Biber 1988: 93, Gorsuch 1983: 268). After determining which variables belonged to each of the factors, the positive features were separated from those that loaded negatively. The reader is referred to Egbert (2014) for the complete results of the factor analysis performed here.

The next step was to calculate dimension scores for each text in the corpus by first standardizing the rates of occurrence for each linguistic feature to a mean of 0 and a standard deviation of 1 using the z-score formula: \( z = \frac{x - \mu}{\sigma} \). This was done in order to ensure that all features have an equal influence on a text’s dimension score. The standardized counts for the negatively loading features were then summed and subtracted from the sum of the counts for the positively loading features for each dimension to create five dimension scores for each text.

In the Register model, a MANOVA and subsequent one-way ANOVAs were used to test for significant difference across the six register groups on each of the five dimensions, and Tukey HSD post hoc tests were performed for each of the dimensions, using register as the independent variable.

In the Publication Type x Discipline model, a factorial MANOVA and a series of 2x3 factorial ANOVAs was also performed in order to (i) test for significant interaction effects between publication type and discipline, and (ii) measure the separate effects of these two variables. An interaction term was included in the model to test whether the effect of discipline is determined by register (and vice versa) or whether the effects of discipline and register operate independently of one another. In each case, the significance of the interaction term was tested and reported first because the outcome of this test determines whether it is most appropriate to perform tests for significant main effects or simple effects. In cases where no statistical interaction was found the main effects of the two factors were investigated and interpreted. Following Kuehl (2000: 179), when a statistical interaction was discovered, main effects were considered uninterpretable and simple effects ANOVAs were used to test for the effect of one factor within one level of the other factor. In the cases where a significant simple effect or main effect was found, post hoc Tukey HSD pairwise mean comparisons were performed to determine which pairs differed significantly.
4. Results

This section contains a description of the five linguistic dimensions, including the quantitative and qualitative results for register, discipline, and publication type. Each of the five dimensions was interpreted using an iterative three stage process that included: (i) studying the co-occurrence patterns of linguistic features, (ii) investigating these patterns in the texts, and (iii) reviewing the results of previous MD analyses. Tables 3–7 display the positive and negative features associated with each dimension along with example text excerpts, with positively loading linguistic features bolded and negatively loading features italicized in both excerpts.

In order to interpret Dimension 1, it may be most useful to begin by looking at the features that loaded negatively. Three of the features on this dimension are nominal in nature, including all nouns, pre-modifying nouns (e.g. activation sequence), and technical concrete nouns (e.g. chromosome, diagram). Texts with these features share at least two functional characteristics: specialization of information and density of information. Nominal structures, in general, transmit information to the reader, but the use of technical concrete nouns suggests that the information is specialized or technical, and the use of nouns as nominal pre-modifiers adds additional layers of informational density. Further evidence of the informational focus of the negative side of this dimension is the presence of agentless passives. Passives are common in writing, especially writing that presents abstract, decontextualized information. Agentless passives are especially common in academic writing, a register that commonly deemphasizes the agent, usually the researcher or writer, in order to focus on what was done rather than who did it.

Table 3. Dimension 1: Non-technical synthesis vs. specialized information density

<table>
<thead>
<tr>
<th>Features with positive loadings</th>
<th>PS_HI_10 (D1 score: 15.82)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbs: <strong>HAVE</strong> (0.36)</td>
<td>Although in its most general sense democracy is ancient, the <strong>form of democracy</strong> I shall be mainly discussing in this book is a product of the twentieth century. Today we have come to assume that democracy must guarantee virtually every adult citizen the right to vote.</td>
</tr>
<tr>
<td>Vocabulary: <strong>core 1–500</strong> (0.61)</td>
<td></td>
</tr>
<tr>
<td>Other: <strong>adverbials</strong> (0.59); <strong>conjuncts</strong> (0.51); <strong>amplifiers</strong> (0.43); <strong>phrasal coordinators</strong> (0.39); <strong>factive adverbials</strong> (0.37); <strong>emphatics</strong> (0.36); <strong>that relative clauses</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features with negative loadings</th>
<th>JA_BI_24 (D1 score:−21.67)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns: <strong>pre-nominal modifiers</strong> (−0.73); <strong>nouns</strong> (−0.73); <strong>technical concrete nouns</strong> (−0.31)</td>
<td>An upstream activation sequence (UAS) containing P element, GS1664, inserted upstream of the foxO gene, disrupts the embryonic axonal scaffold when crossed to the panneuronal driver ElavGal4.</td>
</tr>
</tbody>
</table>
The positive end of this dimension, in contrast, contains features such as general adverbials, certainty adverbials, emphatics (e.g. very, highly), and amplifiers (e.g. a lot, just, really) which have interactive and affective functions. Emphatics and amplifiers are both used as adverbial intensifiers to add emotive emphasis to a message. In addition, a high percentage of frequent vocabulary suggests non-technical discourse, and phrasal coordination and adverbial conjuncts (e.g. alternatively, consequently) function to connect ideas in a coherent synthesis.

These features, contrasted with the negative features, suggest a continuum based on two related functional considerations. The first functional consideration is the amount of technical language that is used to transmit the information in the text. The negative end of this dimension is highly technical, and the positive end is non-technical. The second functional consideration is the number and broadness of the sources of information used in the writing. Texts on the negative end tend to present dense information from a small number of narrow sources (e.g. results of a scientific experiment), while texts with high positive scores contain a synthesis of information taken from many broad sources (e.g. historical overview of major advances in a scientific field). Taken together, the positive and negative features on Dimension 1 support a label of “Non-technical synthesis vs. specialized information density”.

Dimension 2 highlights a very different facet of linguistic variation. This dimension has no negative features. The positive co-occurrence of present tense verbs, be verbs, and predicative adjectives have functions in definitions and

<table>
<thead>
<tr>
<th>Features with positive loadings</th>
<th>TB_BI_21 (D2 score: 29.18)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Verbs:</strong> present tense (0.73); be (0.59)</td>
<td></td>
</tr>
<tr>
<td><strong>Nouns:</strong> concrete (0.30)</td>
<td></td>
</tr>
<tr>
<td><strong>Modals:</strong> possibility (0.67); prediction (0.38)</td>
<td></td>
</tr>
<tr>
<td><strong>Other:</strong> non-finite to-clauses controlled by adjectives (0.68); Academic lexical bundles (0.58); that-clauses controlled by stance adjectives (0.57); predicative adjectives (0.53)</td>
<td></td>
</tr>
</tbody>
</table>

It implies that the state variables, which are considered as the result of ecological models also may be spatially distributed. Spatial distribution is characteristic for most ecosystems and it would therefore be important to be able to develop ecological models with spatial distribution. The question that will be discussed in this chapter is: how is it possible to develop spatial models?

<table>
<thead>
<tr>
<th>Features with negative loadings</th>
<th>JA_BI_25 (D2 score: −9.13)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cells</strong> were lysed with NETN buffer (0.5% NP-40, 150 mM NaCl, 50 mM Tris, and 1 mM EDTA) at 4°C. Cell debris was removed by centrifugation, and the supernatant was incubated with 5 µg of the appropriate antibody and protein A beads at 4°C for 4 h.</td>
<td></td>
</tr>
</tbody>
</table>
explanations of terms and concepts. The features of possibility and prediction modals and that clauses controlled by stance adjectives show added evaluation from the author regarding prediction and possibility. Taken together, these features support a Dimension 2 label of “Definition and evaluation of new concepts”.

The third dimension also has no negative features. The presence of communication verbs (e.g. *say*, *assert*), mental verbs (e.g. *think*, *reveal*), suasive verbs (e.g. *agree*, *urge*), cognitive nouns (e.g. *consideration*, *idea*), and several types of that-clauses controlled by various stance features shows a large degree of author stance. This is further supported by the presence of first person pronouns. These features support a Dimension 3 label of “Author-centered stance”.

The positive features on Dimension 4 include past tense verbs, third person pronouns, and aspectual verbs (e.g. *begin*, *finish*). These are strong indicators of narrative prose. The addition of phrasal verbs (e.g. *look up*, *take over*), which are relatively common in informal spoken registers, and the negatively loading academic words suggest that the narrative prose associated with this dimension is also relatively easy to process. These features support the label of “Colloquial narrative” for Dimension 4.

Positive-loading features for the final dimension include nominalizations and moderate frequency vocabulary, suggesting the use of relatively common words with dense informational packaging. The addition of abstract nouns (e.g. *arrangement*, *transition*) and topic adjectives (e.g. *natural*, *physical*) show a high degree of abstraction. These features are contrasted with time adverbials (e.g. *now*,

---

**Table 5.** Dimension 3: Author-centered stance

<table>
<thead>
<tr>
<th>Features with positive loadings</th>
<th>PS_BI_24 (D3 score: 18.99)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbs: communication (0.58); mental (0.58); suasive (0.42)</td>
<td>This book proposes that our minds evolved not just as survival machines, but as courtship machines. Every one of our ancestors managed not just to live for a while, but to convince at least one sexual partner to have enough sex to produce offspring. [...] Following this insight, I shall argue that the most distinctive aspects of our minds evolved largely through the sexual choices our ancestors made.</td>
</tr>
<tr>
<td>Nouns: human (0.41); cognition (0.39)</td>
<td></td>
</tr>
<tr>
<td>Stance: that-clauses controlled by non-factive verbs (0.53); that-clauses controlled by likelihood verbs (0.46); that-relative clauses controlled by stance nouns (0.38); that-clauses controlled by factive verbs (0.30)</td>
<td></td>
</tr>
<tr>
<td>Other: infinitives (0.53); 1st person pronouns (0.36)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features with negative loadings</th>
<th>JA_HI_02 (D3 score: −7.88)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crucial to these discourses was a narrative of progress and modernity versus tradition and stagnation that reflected the dual society paradigm central to the enactment of Zionist ideology and politics.</td>
<td></td>
</tr>
</tbody>
</table>

This book proposes that our minds evolved not just as survival machines, but as courtship machines. Every one of our ancestors managed not just to live for a while, but to convince at least one sexual partner to have enough sex to produce offspring. [...] Following this insight, I shall argue that the most distinctive aspects of our minds evolved largely through the sexual choices our ancestors made.
Table 6. Dimension 4: Colloquial narrative

<table>
<thead>
<tr>
<th>Features with positive loadings</th>
<th>TB_HI_04 (D4 score: 8.93)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verbs: past tense (0.59); activity (0.54); common phrasal (0.47); aspectual (0.44); present progressive (0.37)</td>
<td>Mitrokhin’s most anxious moment came when he arrived at his weekend dacha to find a stranger hiding in the attic. He was instantly reminded of the incident a few years earlier, in August 1971, when a friend of the writer Aleksandr Solzhenitsyn had called unexpectedly at his dacha while Solzhenitsyn was away and surprised two KGB officers in the attic who were probably searching for subversive manuscripts. Other KGB men had quickly arrived on the scene and Solzhenitsyn’s friend had been badly beaten.</td>
</tr>
<tr>
<td>Other: 3rd person pronouns</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features with negative loadings</th>
<th>TB_BI_06 (D4 score: −6.69)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary: academic words (−0.64)</td>
<td>Salmonellae are Enterobacteriaceae that are widely distributed in the environment and include more than 2000 serotypes. The Salmonella numbers in wastewater range from a few to 8000 organisms/100 mL; they are the most predominant pathogenic bacteria in wastewater and cause typhoid and paratyphoid fever, and gastroenteritis.</td>
</tr>
</tbody>
</table>

Table 7. Dimension 5: Abstract observation and description

<table>
<thead>
<tr>
<th>Features with positive loadings</th>
<th>JA_HI_18 (D5 score: 9.76)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns: nominalizations (0.71); process (0.59); abstract (0.47)</td>
<td>Along these same themes, the circuitous routes on which visual images of production traveled reinforce the view that “china” (porcelain) effected a global culture. Specific to the pictorial themes discussed in this article, production as a visual theme was consumed as a product in itself, and the mode of viewing, a historically constructed visuality about porcelain production, was what gained global purchase.</td>
</tr>
<tr>
<td>Adjectives: attributive (0.53); topic (0.47)</td>
<td></td>
</tr>
<tr>
<td>Other: word length (0.65); core vocabulary (501–3000)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Features with negative loadings</th>
<th>PS_BI_12 (D5 score: −8.85)</th>
</tr>
</thead>
<tbody>
<tr>
<td>time adverbials (−0.32)</td>
<td>Cocoa beans are harvested from the cocoa tree, which grows best in warm, moist climates and within 20 [degrees] latitude of the Equator. The world production of cocoa beans is two million tons a year, and they are grown in Brazil and Mexico for the North American market, and in West Africa for the European market.</td>
</tr>
</tbody>
</table>
tomorrow), which are concrete and descriptive in nature. The positive features seem to function as linguistic tools used by authors to characterize and interpret their findings. In addition, the positive features seem to be associated with abstract interpretation rather than the explanation of concrete evidence. Therefore, the label of “Abstract observation and description” is used here.

4.1 Variation across registers

Linguistic variation across register categories is graphically displayed in Figure 3 below. Each of the six cells in this figure contains a radar plot for one of the six register groups. Each of these plots displays two shapes, a dotted shape for the overall corpus means and a solid shape for the register means. Each of these shapes has five points, one for each of the five linguistic dimensions (LD). It is worth noting that for each of the five dimensions the overall corpus mean was very close to zero (< 0.25).

A one-way MANOVA was performed using the five linguistic dimensions as dependent variables and register as the independent variable. Using the Wilks’ criterion, the combination of the dimension scores were significantly affected by register \(F(5, 140) = 11.79, p < .001\). In order to investigate the impact of register on each dependent variable, separate one-way ANOVAs were performed for each of the five dimensions, using a Bonferroni-adjusted alpha criterion of .01 \((\alpha = .05 / 5\) comparisons). The results can be seen in Table 8 below.

The one-way ANOVAs revealed significant register differences across each of the five linguistic dimensions. However, the effect sizes \((R^2)\) show that some of the dimensions are stronger predictors of register variation than others. In particular, LD1 had the largest effect size, showing that register differences account for more than 50% of the variance in the use of linguistic features associated with “Non-technical synthesis vs. specialized information density”. In contrast, register variation only accounted for 17% and 13% of the variance in the linguistic features associated with Dimensions 3 and 5, respectively.

Tukey HSD post hoc analyses were performed for each dimension. These results are reported in Figure 3 next to the register labels in the form of groupings using letters (e.g. A, AB, etc.). Within the Tukey HSD groupings, pairwise differences are statistical on a particular dimension in cases where two registers do not share a single letter in common. These results reveal that a substantial amount of linguistic variability can be attributed to register differences. By viewing Figure 3 as a table with six cells, we can see that the columns contain the results for the two disciplines, and the rows contain the results for the three publication types.

In looking at these results, we might expect to see that registers that are more similar to each other situationally are also the most similar linguistically. For example, we might expect journal articles in history to be more similar to journal
Figure 3. Radar plots of the mean dimension scores (register and full corpus) for the six registers
articles in biology than to textbooks in biology because they share a situational parameter (publication type). However, this is clearly not the case. Journal articles in history are statistically different from journal articles in biology on four of the five dimensions, whereas they only differ from university textbooks in biology on one dimension. One way of interpreting these results would be to simply conclude that each of these registers is an independent text type. The problem with this approach is that it fails to account for key underlying situational variables (e.g., publication type and discipline) that are crossed in the model. If we accept the Register model as the best way of representing the data, then we are essentially concluding that there is as much variability within publication types and disciplines as between them. This would raise serious questions about the validity of any study that compares language use across disciplines or publication types.

Furthermore, the findings of the Register model outlined in this section suggest that discipline and publication type are distinct variables that interact to create complex patterns of linguistic variation in published academic writing. For example, there are no statistical differences between popular academic books in history and biology on any of the five dimensions. On the other hand, as mentioned above, there are statistically significant differences between journal articles in biology and history on four of the five dimensions. These findings suggest that the situational grouping variables of discipline and publication type can be important predictors of linguistic variation. However, as we observed above, this is not the case all of the time. This shows that the Register model may be too simple to account for the complex situational relationships among registers groups. One possible explanation for the confusing register patterns is that the effect of publication type on language use is moderated by discipline, and vice versa. The next section will present the results of a series of factorial ANOVAs that will be used to identify the presence of moderator variables (i.e., an interaction effect).

Table 8. One-way ANOVA results for the five linguistic dimensions.

<table>
<thead>
<tr>
<th>Linguistic Dimension</th>
<th>df1, df2</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD1: Non-technical synthesis vs. specialized information</td>
<td>5, 144</td>
<td>871.99</td>
<td>30.40</td>
<td>&lt;.001</td>
<td>.51</td>
</tr>
<tr>
<td>LD2: Definition and evaluation of new concepts</td>
<td>5, 144</td>
<td>340.10</td>
<td>11.55</td>
<td>&lt;.001</td>
<td>.29</td>
</tr>
<tr>
<td>LD3: Author-centered stance</td>
<td>5, 144</td>
<td>167.47</td>
<td>6.02</td>
<td>&lt;.001</td>
<td>.17</td>
</tr>
<tr>
<td>LD4: Colloquial narrative</td>
<td>5, 144</td>
<td>135.91</td>
<td>10.62</td>
<td>&lt;.001</td>
<td>.27</td>
</tr>
<tr>
<td>LD5: Abstract observation and description</td>
<td>5, 144</td>
<td>84.93</td>
<td>4.33</td>
<td>.001</td>
<td>.13</td>
</tr>
</tbody>
</table>
4.2 Variation across publication types and disciplines

A two-way factorial MANOVA was performed using the five linguistic dimensions as dependent variables and publication type and discipline as the two independent variables. Using the Wilks’ criterion, the results showed a statistical interaction between publication type and discipline, $F(10, 280) = 8.06, p < .001$. The combination of the dimension scores were also significantly affected by publication type, $F(10, 280) = 14.97, p < .001$, and discipline, $F(5, 140) = 22.87, p < .001$.

Separate two-way ANOVAs were also performed for each of the five dimensions, using a Bonferroni-adjusted alpha criterion of .01 ($\alpha = .05 / 5$ comparisons). The results of these analyses are reported below. In cases where a significant interaction effect is discovered, simple effect contrasts are carried out to test for significant differences between levels of one factor on one level of the other. In every case, the alpha levels for these simple effects contrasts are adjusted using the appropriate Bonferroni correction. In cases where it is determined that the factors do not interact, tests for the significance of the main effects of each factor are carried out.

There was a significant interaction effect between publication type and discipline for Dimension 1, $F(2, 144) = 23.46, p < .001$. The $R^2 = .514$ indicating that the interaction effect between publication type and discipline accounts for approximately 51% of the variability in Dimension 1 scores. Figure 4 below displays the marginal means for the Dimension 1 scores of each publication type within the

![Figure 4](image_url)
two disciplines. The plot shows a clear difference in the trends for the registers within biology versus history.

Whereas biology popular academic writing contains more features associated with non-technical synthesis than popular academic books in history, biology journal articles contain much more specialized information density than those in history. Simple effects ANOVAs for the discipline pairs within each register showed no significant differences between biology and history popular academic books and textbook writing. However, the Dimension 1 scores for biology journal articles were significantly lower than those for history, $F(1, 144) = 64.23, p < .001$. Furthermore, while there were no simple effects differences among the history publication types, statistically significant differences were found between the biology publication types, $F(2, 144) = 64.17, p < .001$.

The publication type x discipline interaction effect was also significant for Dimension 2, $F(2, 144) = 11.27, p < .001$, $R^2 = .318$. The most striking difference that can be seen in Figure 5 is between biology and history textbooks. The simple effects ANOVA results showed this difference to be significant, $F(1, 144) = 35.99, p < .001$, whereas there were no statistical discipline differences for the other two publication types. The extremely high Dimension 2 scores for biology textbooks also contributed to significant differences across the three biology publication types, $F(2, 144) = 24.79, p < .001$.

The Dimension 3 data presented in Figure 6 suggest a change in direction in the trends for the biology and history publication types, resulting from the higher amount of author-centered stance in history journal articles relative to those from

Figure 5. Marginal means plot for Dimension 2: “Definition and evaluation of new concepts”
biology. However, the ANOVA results did not suggest a significant interaction effect \( F(2, 144) = 3.15, p = .046, R^2 = .173 \). The main effects ANOVAs revealed a significant main effect of publication type, with popular academic books containing significantly more author-centered stance than the other two publication types. There was no significant main effect of discipline. A qualitative analysis of the results displayed in Figure 6 shows that biology writing contains incrementally less author-centered stance as the expertise and specialization of the target audience increases. In history this is true between popular academic writing and textbooks, but history journal articles actually contain more author-centered stance than university textbooks.

Unlike the first three dimensions, there is no significant interaction effect for Dimension 4. This shows that Dimension 4 scores for publication types do not depend on the discipline and vice versa. This is clearly shown by the nearly parallel lines in Figure 7. The main effects were significant for both discipline, \( F(1, 144) = 39.674, p < .001 \), and publication type, \( F(2, 144) = 16.22, p < .001 \). This reveals that there are significantly more features associated with colloquial narrative in history than in biology. In order to determine where the publication type differences lie, a post hoc analysis was conducted using Tukey HSD pairwise mean comparisons. These analyses showed that biology journal articles contain significantly less colloquial narrative than biology popular academic writing \( (p < .001) \) and biology textbooks \( (p = .001) \).

The interaction effect between discipline and publication type was not statistically significant for Dimension 5, \( F(2, 144) = 4.012, p = .02, R^2 = .131 \). However, I proceeded to perform and interpret simple effects due to the relatively large

![Figure 6. Marginal means plot for Dimension 3: “Author-centered stance”](image-url)
interaction effect and a visual review of the data in Figure 8. Simple effects ANOVAs confirm the results displayed in Figure 8. There are significant differences among the three history publication types, $F (2, 144) = 6.34, p = .002$, but not among the publication types of biology. The only publication type simple effects that were found were between the two disciplines of journal articles, $F (1, 144) = 12.01, p = .001$, showing that history journal articles contain significantly more abstract observation and description than biology journal articles.

Figure 7. Marginal means plot for Dimension 4: “Colloquial narrative”

Figure 8. Marginal means plots for Dimension 5: “Abstract observation and description”
5. Discussion

In this section, I summarize and discuss the results of this study and their implications for research on register and discipline variation in published academic writing and for corpus research, generally. In the first part of this study I performed a MD analysis of 62 linguistic features within the 150 texts of the corpus. The following five interpretable dimensions of linguistic variation in published academic writing were identified:

i. “Non-technical synthesis vs. specialized information density”
ii. “Definition and evaluation of new concepts”
iii. “Author-centered stance”
iv. “Colloquial narrative”
v. “Abstract observation and description”

After calculating a unique dimension score for each text in the corpus, comparisons were made (i) across registers and (ii) between disciplines and across publication types.

5.1 What can we learn from the Register and Publication Type x Discipline models?

Two models were used to quantitatively analyze linguistic variability in the corpus. On the level of pairwise *post hoc* tests, the statistical results of these two models show identical patterns. The key difference is that the Publication Type x Discipline model makes a distinction between the situational variables of publication type and discipline, whereas the Register model collapses these two variables into one situational parameter. The most important benefit of the Register model is that it makes it possible to measure differences between specific register categories in a manner that is fine-grained and efficient. The main benefit of the Publication Type x Discipline model is that it quantifies the effects of the separate situational factors of publication type and discipline, as well as possible interactions between them.

The Register model was a useful first step for the analyses presented here because it (i) allowed us to graphically display the six register groups on a single parameter, (ii) established that the variable of register has a significant effect on linguistic variability in the corpus, and (iii) established the significant pairwise differences and register groupings within the corpus. This model offers a general overview of patterns in the data. However, this method also suggested the need for a more complex model in order to explain some of the statistical differences and groupings that could not be explained by simple register differences. Additionally,
the statistical procedures used in the Register model make it impossible to simultaneously determine the statistical effects of discipline and publication type.

The Publication Type x Discipline model was a logical next step in order to gain a more complete understanding of the variability due to publication type and discipline. As mentioned before, the post hoc analyses for these two approaches ultimately reveal the same patterns of difference between textual categories. However, the results of the previous analyses supported the use of both models in order to describe different types of variability in the data. For example, the results showed that the effect of discipline differs depending on the publication type. The Publication Type x Discipline model revealed significant differences between journal articles in biology and history on four of the five dimensions. In contrast, significant discipline differences emerged on only two of the dimensions for university textbooks and on none of the dimensions for popular academic books. This suggests that discipline, at least between biology and history, is a more important factor in some written publication types (e.g. journal articles) than in others (e.g. university textbooks, popular academic books).

Interestingly, although there are bigger differences between the two disciplines in journal articles than in the other two publication types, the linguistic variation (measured using standard deviation) within the publication type of journal articles is not consistently larger than the other two publication types. This shows that the discipline variation within journal articles cannot be simply attributed to greater overall variation within this publication type. On the contrary, discipline seems to be a more powerful explanatory variable within journal articles than in the other two publication types. This finding suggests that the extent of discipline variation may depend on the publication type. Further investigation of this pattern in a wider range of disciplines and publication types would be an interesting area for future research on academic writing.

Another way of looking at interaction effects is from the perspective of within-discipline variation. Within the discipline of biology, significant publication type differences emerged on four of the five dimensions. In contrast, significant publication type differences within the discipline of history were only found on one dimension. This reveals that the variable of publication type seems to play a more important role within biology writing than within history writing.

The results of this study strongly support the importance of register as a predictor of linguistic variation in academic writing. However, the Publication Type x Discipline model revealed that discipline and publication type interact in complex ways. This is due, at least in part, to the situational relationships between discipline and publication type. For example, the lack of discipline variation within popular academic writing may be due to differences in the situational variables of purpose and nature of the data or evidence. In contrast, there are stark discipline
differences within the publication type of journal articles; whereas history journal articles report an author’s ideas and observations, biology journal articles transmit empirical results of new scientific findings.

In summary, this study supports the findings of previous research on academic writing by demonstrating the importance of the situational variables of publication type and discipline. The situational definition of a register category encompasses the variables of discipline and publication type. However, these two variables have emerged as two of the most important situational parameters within the definition of academic registers. Additionally, unlike many of the other situational considerations in a register definition, it is possible to design a corpus so that the variables of publication type and discipline are completely crossed, as they are in this study. Balancing the design of the corpus in this way makes it possible to measure the effects of publication type and discipline, as well as interactions between them. This design is attractive because it allows for a more complete description of the effects of publication type and discipline on register variation.

5.2 Interpreting variation among publication types

In this section, I briefly summarize and interpret the linguistic results of this study, organized according to publication type. Discipline variation is also discussed throughout. The Dimension 1 scores for popular academic writing suggest that popular academic writers rely on a writing style that is more emotional and less informationally dense. Popular academic writing also contains more features associated with defining and evaluating new concepts than journal articles but not nearly as many as biology textbooks. Additionally, Dimension 5 shows that popular academic books use very little “Abstract observation and description”, or elaborated, abstract language, relative to the textbooks and history articles. However, the Dimension 5 scores for popular academic books and biology journal articles were very similar. This does not necessarily imply, however, that biology journal articles and popular academic writing have the same reasons for not using “Abstract observation and description”. A closer investigation of the texts revealed that the reason abstract elaboration is less common in popular academic writing is that authors focus instead on making concrete scientific findings and evidence relevant and exciting to the reader. In summary, popular academic writing in history and biology can be described as an emotional, author-focused register that is written to be readable and appealing to a wide audience.

The results from the first dimension show that textbooks contain less non-technical synthesis and more informational density than popular academic books in both disciplines. The large difference between the Dimension 2 scores for biology and history textbooks is one of the clearest examples in this study of the
importance of accounting for discipline variation in a corpus-based study. This difference is not altogether surprising once we consider the nature of pedagogy in these two disciplines. Biology is highly conceptual, requiring textbook authors to define and explain new terms and processes. Biology can also be quite difficult for learners to relate to, as much of it is studied at the microscopic level. History, on the other hand, does not rely as heavily on new concepts but on narrating and sequencing the details of past events. In many cases, these events are highly relevant to the reader, thus not requiring extensive explanation and evaluation. To summarize, textbooks tend to fall somewhere between journal articles and popular academic writing in most areas. However, biology textbooks focus on defining and evaluating new concepts, whereas history textbooks contain more colloquial narrative.

Relative to other publication types within their respective disciplines, journal articles in biology and history tend to have more specialized language and density of information. Journal articles in both disciplines also contain less definitions and evaluations of new concepts and colloquial narrative. However, biology textbooks contain less author-centered stance, less colloquial narrative, and less “Abstract observation and description” than journal articles in history. These patterns are unsurprising when we consider the highly objective informational focus of writing in the natural sciences. The supporting evidence used by biology writers, in most cases, results from analyses of empirical data. This eliminates many of the factors that motivate the use of author-centered stance and abstract elaboration. Supporting evidence in history articles, on the other hand, is often heavily based on the interpretation and rhetoric of the author regarding causes for and relationships among events.

5.3 Implications

The findings from this study have implications for research on language variation, in general, and register, publication type, and discipline variation within academic writing, specifically. This paper began with a discussion of register as an important predictor of linguistic variation. In addition to register there are many other predictors of variation in language use, including discipline, time, dialect, speaker status, mode, and task. While a large amount of research has focused on differences across levels of these factors independent of other factors, very little research has investigated potential interactions among sets of two or more of these variables. The results of this study have shown that these interaction effects, if they exist, have a profound impact on the interpretation of our research findings and the conclusions we can draw from our data. This is not to say that the use of factorial ANOVAs to measure interaction effects is appropriate for every study. Rather
it is to say that anytime two or more independent categorical variables or factors exist in our corpus sample, we should at least test whether they interact with each other. This will help us ensure we are appropriately and accurately interpreting the results of our study.

In cases where researchers are not interested in possible interactions between independent factors, they should design the corpus in such a way that it represents the range of variability along a single factor (e.g. academic journal articles), without introducing additional factors (e.g. disciplines, time periods). By controlling for variables that are deemed extraneous to the research questions of interest, the researcher will reduce the noise, or variance, in the sample, allowing for a more accurate description and/or comparison of the levels within the factor of interest. Although this will necessarily limit the scope and generalizability of the study, it will increase the interpretability of the findings. In other words, it is the researcher’s responsibility to create a representative corpus by including the full range of variability for the language variety of interest and limiting additional variability due to factors and factor levels that are beyond the scope of the study.

Finally, representativeness in corpus design has been an area of emphasis in the field of corpus linguistics for at least two decades (see e.g. Biber 1993), but the specific implications of representativeness for corpora in studies of multiple factors have not received adequate attention. This study has shown that issues of representativeness must be considered at each stage of a corpus study. After establishing the research questions for a corpus study, we should design our corpus sample so that it (i) includes the full range of variability within the factor(s) (e.g. publication type, discipline) and factor levels (e.g. journal articles, textbooks) that we are interested in measuring, and (ii) excludes, to the extent possible, extraneous variables from the corpus sample. We should then perform appropriate analyses which account for variability within each of the factors in the corpus. In cases where more than one factor exists in a corpus sample, a statistical technique, such as a factorial ANOVA, should be used to test for interactions. The presence or absence of an interaction effect will then determine how the researcher should proceed to analyze and interpret those patterns. Finally, patterns should be interpreted based on the corpus sample used in the study and generalized only to the factors and factor levels included in that corpus sample.

This study has shown that some of the variability within registers of academic writing can be accounted for by publication type and discipline variation. The large differences in mean dimension scores between the three publication types in this study have shown that descriptions of linguistic variation based on situation of use may be more useful and meaningful if they are carried out at a higher level of granularity. Accounting for publication type variation may be particularly useful in studies such as this one that include many linguistic variables because patterns
across publication types are likely to vary based on the linguistic feature. For example, the linear cline I hypothesized at the beginning of the study (in which journal articles and popular academic writing are at the extreme ends with textbooks in the middle) was only supported by two of the five dimensions. Dimensions 2, 3, and 5 resulted in situations where textbooks had dimension scores more extreme than one or both of the other two publication types in at least one discipline. This shows that there is value in measuring publication type and discipline variation rather than merely describing patterns within a general macro-register, such as academic writing, and making unwarranted assumptions about universal patterns of register variation.

5.4 Conclusion and future research

It is hoped that this corpus-based study, although relatively small-scale and limited in scope, has contributed valuable insights into the nature of publication type and discipline variation in published academic writing. The findings presented here have raised important new questions about the existence of interaction between other linguistic factors and the interpretation of those interaction effects. These empirical questions are worthy of future study. Replication of the design used in this study with larger corpora containing additional publication types and disciplines would reveal whether interaction effects exist between disciplines and publication types other than the ones included here. The factorial design used in this study can also be applied to other language variables, including time, mode, dialect, and L1 background for non-native speakers. Any of these investigations can be carried out using individual linguistic features or large sets of language variables using multivariate methods such as Multi-Dimensional analysis. Future research that incorporates simultaneous investigations of multiple factors will improve our understanding of the complex nature of linguistic variation.

References


*Author’s address*

Jesse Egbert  
Department of Linguistics and English Language  
Brigham Young University  
4050 JFSB  
Provo, UT 84602  
USA  
Jesse_Egbert@byu.edu