English for Specific Purposes 62 (2021) 15–29

Contents lists available at ScienceDirect

English for Specific Purposes

journal homepage: http://ees.elsevier.com/esp/default.asp

Parameters of variation in the use of words in empirical research writing

Taha Omidian [a,*], Anna Siyanova-Chanturia [a,b]

[a] School of Linguistics and Applied Language Studies, Victoria University of Wellington, New Zealand
[b] Ocean University of China, PR China

ARTICLE INFO

Article history:
Available online 4 December 2020

1. Introduction

The late Victorian era has long been associated with developments in the institutionalization of the human sciences (Anderson & Valente, 2002; Russell, 1991; Shumway & Messer-Davidow, 1991). During this formative period, many educational systems (especially in Europe) saw fundamental changes in their social structures and institutional practices. Anderson and Valente (2002), in an edited collection of historical essays entitled *Disciplinarity at Fin de Siècle,* highlight the important role of the educational reform practices of this era in the emergence of discrete intellectual fields in academe. The majority of these practices were mainly related to the compartmentalization of 'scholarly knowledge', as the main product of formal education and academia at large (see e.g., Hoskin, 1993; Russell, 1991). The philosophy which drove these changes was motivated by the recognition that scholarly communities in academia are so diverse and disparate that providing an overarching, unified explanation for the different forms of knowledge they produce is virtually impossible (see Bucher & Strauss, 1961; also see Mannheim, 2013, p. 67 for a detailed discussion of the sociology of knowledge). Included in this recognition is the view that the creation of knowledge is highly likely to be differentially affected by the epistemological orientations of communities and their routine scholarly practices. And thus, accomplishing the task of structuring and organizing such a multi-dimensional phenomenon as knowledge appears to be symbiotically linked to the ability to govern the delineation of academic communities and characterize their intellectual and knowledge practices.

* Corresponding author. Kelburn Parade, Wellington, 6012, New Zealand.
E-mail address: taha.omidian@vuw.ac.nz (T. Omidian).

https://doi.org/10.1016/j.esp.2020.11.001
0889-4906/

Knowledge in disciplinary communities is negotiated through constructing a form of discourse that not only reflects the collective norms of the community but also adheres to the expectations of its members. Such community-specific meaning is primarily communicated through writing in academia (see, e.g., Elbow, 1991; Russell, 1991). The use of writing, as the primary means of knowledge production and transmission, in academia is deeply rooted in the institutionalization of the human sciences and the formation of academic disciplines during the Victorian era. According to Russell (1991), before the 1870s, scholarly knowledge was primarily communicated through speaking (e.g., oratory, recitation, debate), and writing was merely an aid to memory and a way of preserving thinking for speech. In parallel with the formation of academic disciplines during the late nineteenth century, however, the predominant role of writing in the production of specialized meaning began to emerge. This was due to the recognition of the need for the transmission of discipline-specific meaning to specialized audiences.

Due to its distinctive style and conventions, academic writing has spawned major investigatory efforts in both research and practice across various fields of inquiry, such as sociology of knowledge, rhetoric, language education, and applied linguistics (see e.g., Bazerman, 1988; Hyland, 2018). Research in the applied areas of this interdisciplinary inquiry has primarily focused on creating pathways to achieving proficiency in academic writing. One promising avenue that has long been a topic of interest in applied linguistics and educational circles in this area is academic vocabulary knowledge. Research has shown that ample knowledge of such vocabulary can be conducive to gaining academic literacy, in that it provides students with varied lexical choices that are particularly useful for communication in academic discourse (e.g., Corson, 1997; Coxhead & Nation, 2001; Durrant, 2016; also see; Omidian, Beliaeva, Todd, & Siyanova-Chanturia, 2017).

Academic vocabulary is typically classified into two sub-categories: sub-technical and technical vocabulary (e.g., Nation, 2001). Sub-technical (or general academic) vocabulary refers to those items that are not tied to a specific subject area and are commonly used in a wide range of academic disciplines (Coxhead, 2000). This vocabulary, which includes lexes such as *insight*, *exhibit*, *collapse, adequate,* falls somewhere between non-academic and technical vocabulary (see Coxhead, 2020). Technical (or discipline-specific) vocabulary is a set of subject-related items (e.g., *estrogen, periodontal, oxidation*) that are used to create specialized knowledge in a given domain of inquiry (see Liu & Lei, 2020). Such domain-specific lexes are employed by members of a particular academic community to construct a specialized form of discourse which is exclusive to their target audience and may not be readily understood by members of other disciplinary domains.

Knowledge of academic vocabulary (both technical and sub-technical) plays an important role in the construction of meaning in written academic genres. A number of studies have explored the specific use of academic vocabulary in disciplinary fields (e.g., Durrant, 2014, 2016; Hyland & Tse, 2007; Martínez, Beck, & Panza, 2009). The rationale for investigating disciplinary specificity in the use of academic vocabulary is two-fold. First, some words are highly likely to be more important for academic writing in certain disciplines, compared to others (e.g., Durrant, 2016; Hyland & Tse, 2007). Second, it is highly unlikely that these words are evenly dispersed across disciplines and, thus, may not be equally useful for writing across branches of academic study (e.g., Durrant, 2014). In line with this argument, several studies have investigated disciplinary variation in the use of items from academic vocabulary lists (e.g., the AWL, Coxhead, 2000; and the AVL, Gardner & Davies, 2014) in a range of academic disciplines (e.g., Durrant, 2016; Hyland & Tse, 2007; Martínez, et al., 2009). The evidence from this line of research appears to endorse the notion of specificity in academic vocabulary use. However, since these studies mainly focus on the evaluation of such specificity on the basis of previously compiled academic wordlists, it can be reasonably argued that their conclusions are largely restricted to the items included in those lists. The approach adopted in these studies is to use corpus evidence to analyze different patterns of use for a set of pre-defined items. In other words, corpus analysis in such studies is viewed as an evidential tool to validate theoretical presumptions about a collection of linguistic items. However, more generalizable conclusions can be reached by adopting an inductive approach in which patterns of disciplinary specificity in vocabulary use emerge from the analysis of a corpus with minimal a priori assumptions guiding their identification. Such an approach has the potential to inform the steps that should be taken to explore various aspects of the patterns arising from the data (see Baker & Egbert, 2016; Partington, Duguid, & Taylor, 2013). In this approach, it is the emerged linguistic patterns that lead the researcher to a linguistic theory, and not vice versa.

Very few studies to date have adopted such an approach to investigating disciplinarity in the use of single-word items (Durrant, 2014). Fewer still have adopted this approach to examine variation in the use of such vocabulary in empirical research writing, and across different parts of research articles in various disciplinary fields. Empirical research writing can be thought of as a specific sub-register of academic writing, which focuses on written communication of empirical knowledge created as a result of scholarly research activities. Empirical research writing is one of the primary means of knowledge dissemination in academic contexts. This is a highly conventionalized genre in which effective communication of knowledge is contingent on writers' familiarity with the linguistic expectations of their target readership (see e.g., Bhatia, 1993; Hyland, 2001; Swales, 1990). As Hyland (2011) argues, in writing high-stakes genres such as empirical research articles, in which effective communication of knowledge involves the anticipation of alternative interpretations and readers' possible objections, authors often tend to rely on linguistic resources that can reflect the disciplinary values and expectations of their readers. Such linguistic expectations are mostly related to the use of words that are conventionally expected by readers in a given academic community.

Such considerations, together with the fact that writing research papers is now an integral part of academic life across disciplines (see Curry & Lillis, 2017; Hyland, 2016), stress the importance of providing a comprehensive picture of the similarities and differences in the language of research writing across disciplines. As was mentioned earlier, the unique and specific ways in which knowledge in disciplinary communities is negotiated and transmitted to its target audience can be

used as demarcation criteria for a characterization of disciplinary boundaries. However, as was noted above, focusing on a set of pre-defined items (such as those included in vocabulary lists) for highlighting disciplinary commonalties and variations can curtail the generalizability of the patterns that emerge from such an investigation. For this purpose, the present study investigated parameters of variation in empirical research writing across disciplines by adopting an inductive approach in which patterns of specificity in vocabulary use emerged from corpus data. The emerged patterns were then used to shed further light on conventional discourse practices in research writing across various disciplines.

2. Method

2.1. Corpus

The present study is based on a corpus of empirical research articles (c. 4.5 million words) published in high-ranking, accredited journals from ten different disciplines (i.e., biology, chemistry, dentistry, mechanical engineering, physics, applied linguistics, business, management, politics, and sociology). In order to arrive at an optimal design for the corpus, it was necessary to specify the characteristics of the target texts that were to be included in the corpus. Such characteristics are often referred to as the *situational characteristics* of texts that provide important information about their non-linguistic aspects and help characterize the register they represent (see Biber & Conrad, 2009, Chapter 2). These characteristics can, of course, vary depending on the text category under consideration. For example, academic research articles, as a text category most representative of research writing (e.g., Hyland, 2009, Chapter 4), are often characterized by situational parameters such as their type (theoretical or empirical) and their internal structure. Such factors are non-linguistic characteristics of academic journal articles which can potentially affect the ways in which language is used in this genre of writing. Therefore, identification of such non-linguistic factors is an important consideration for designing a corpus that is intended to represent linguistic variation in a particular context of language use.

To this end, prior to sampling, an inductive survey of high-ranking journals nominated by expert informants (based on their 5-year impact factor published by Thomson Reuter's *Web of Knowledge ISI*) was conducted to identify the type of articles commonly published in each discipline.[1] Full-length articles in at least ten volumes of each journal were carefully examined based on the following characteristics to determine their type:

a) The inclusion of quantitative/qualitative data analyses;
b) Presenting a new model or formula that is analyzed and tested through various analytical procedures;
c) Focusing on theoretical aspects of the field without the inclusion of data.

Articles belonging to (a) and (b) were classified as empirical articles and those identified as belonging to the (c) category were considered theoretical articles. In line with the aims of the present research, theoretical articles were not included in the corpus. It should also be noted that, in cases where we were in doubt regarding the type of an article, the article in question was discarded. For instance, in certain disciplines (e.g., physics, chemistry), writers often make methodological and theoretical arguments regarding the validity or accuracy of an established formula (or model) and present various data analyses to prove their point. Since such articles appear to have characteristics of both theoretical and empirical studies, they do not clearly fit into either category and, therefore, were not considered for inclusion in the corpus.

Following this, empirical papers in each discipline were carefully examined to determine their organization structures. This analysis revealed two main organizational patterns. First, papers which included the content of the standard four-part organization (Introduction, Methods, Results, Discussion; IMRD). Second, articles which contained a different and rather distinct sectioning format. For consistency and comparison purposes, empirical articles that belonged to the latter category were not included in the corpus. In addition, the analysis also revealed that the Results and Discussion (R&D) sections in empirical papers were either fully integrated or separated. Every effort was made to ensure that all the articles included in the corpus contained fully separated R&D sections (also see Stoller & Robinson, 2013). The rationale for this decision was that R&D sections can perform distinct discourse functions and fulfill different communicative purposes in research articles (e.g., Lin & Evans, 2012; Swales, 2004). The difference in the communicative aims of R&D sections can in turn give rise to systematic patterns of language use that can vary as a result of the sections' communicative purposes. Therefore, it is important to maintain the distinction between these two sections and recognize the significance of their rather distinct communicative aims in discourse.

The extraction of articles was then carried out based on a stratified random sampling procedure through which five articles were sampled from different issues of a journal in each year. Attention was paid to extract a roughly equal number of articles from different journals in each discipline so as to avoid possible journal influences on writing style. Each extracted article was then thoroughly examined to ensure that it followed the operational definitions and criteria for article selection described above. In addition, care was also taken to arrive at comparable word counts for the ten disciplines. Table 1 provides a description of the corpus analyzed in this study.

[1] It is important to note that, in keeping with the aims of the present research, the focus of this survey was on full-length articles and, therefore, short pieces such as book reviews, commentaries, notes and responses were excluded from the study.

Table 1
Composition of the corpus.

Discipline	Texts	Words
Biology	60	487,386
Chemistry	60	445,786
Dentistry	60	424,359
Mechanical engineering	60	414,455
Physics	60	404,233
Applied linguistics	60	414,097
Business	60	495,235
Management	60	473,461
Politics	60	425,584
Sociology	60	484,655
Total number of texts and words	**600**	**4,469,251**

2.2. Quantifying variation in empirical research writing

To quantify linguistic variation in research writing in terms of word use, two complementary measures were used: lexical dispersion and keyness. The former measure was employed to determine the degree to which frequently used vocabulary in research writing is evenly dispersed across academic disciplines and different sections of research articles. This measure allowed us to classify high-frequency words in research writing based on their level of technicality without relying on any pre-defined categorizations (e.g., technical and sub-technical vocabulary; see Nation, 2001). Variation in research writing across disciplines and IMRD sections was then characterized based on this classification. Following this, the measure of keyness was employed to verify and further explore the patterns identified by the lexical dispersion measure. For this purpose, commonality and variation between disciplinary fields and sections of research articles were determined based on the degree of overlap in key vocabulary use (more on this below). This analysis afforded a mapping of specificity in research writing on the basis of the vocabulary that plays a key role in the communication of knowledge across academic disciplines and IMRD sections. In what follows below, we provide a more detailed explanation of the steps taken to operationalize these two measures.

As a first step, separate listings of high-frequency words were generated for the corpus.[2] Words were considered frequent if they (a) occurred more than 100 times per million words and (b) appeared in at least 10% of texts in the corpus (also see Coxhead, 2000; Gardner & Davies, 2014; Lei & Liu, 2016). It should be noted that the second criterion (i.e., text dispersion) was used to ensure that the identified words were typical of the entire corpus and not restricted to a few texts or certain writing styles.

It is also important to note that 'vocabulary use' in this study was operationalized as the use of individual *word forms* (as opposed to lemmas and word families) in research writing. The rationale for focusing on word forms as the target unit of analysis was to account for their instability in use across different discourse types. Corpus-based vocabulary studies have repeatedly shown that different forms of a word can exhibit disparity in the semantic, syntactic, and thematic ties that they establish with their surrounding context (e.g., Durrant, 2014, 2016; Hyland & Tse, 2007; Sinclair, 1991; also see Omidian & Siyanova-Chanturia, 2020). This level of disparity in use is often influenced by the situational characteristics of the register in which word forms are used. In the case of the present study, these situational characteristics pertain to disciplinary conventions and specific communicative purposes of different sections of research articles. For example, Hyland and Tse (2007) showed that the derivationally-related forms in the word family *analyse* on the AWL were used to varying degrees across disciplines, with the sciences writing making greater use of the adjective form *analytical* and the Social Sciences relying more on the noun form *analysis*. Previous research has also shown that certain inflectional forms are used more or less frequently across sections of research articles. For example, the past tense and past participle forms of verbs are typically more common in Methods as compared with Introductions and Discussions (e.g., see Biber & Conrad, 2009, p. 130; Martínez, et al., 2009; also see Gray, 2015, p. 104). Ignoring such differences by collapsing different forms of a word into more abstract categories (e.g., word families and lemmas) would, therefore, run the risk of obscuring the linguistic clues that can highlight possible variation in vocabulary use in research writing (also see Durrant, 2014).

Having created the lists of high-frequency words, we then used Gries's (2008) *deviation of proportions* statistic to assess the extent and nature of variability in vocabulary use in the corpus and its subsections (also see Biber, Reppen, Schnur, & Ghanem, 2016b). Deviation of proportions (henceforth DP) is a statistical measure that determines the extent to which a given word form is evenly dispersed across different parts[3] of a corpus. To calculate DP, the following steps were taken in the present study (as per Gries, 2008, p. 415):

[2] This, and all the following textual analyses, were performed using computer programs developed in Perl (ver. 5.010).

[3] Corpus parts in the present study were operationalized on the basis of the situational parameters characterizing texts in the corpus (i.e., academic disciplines and IMRD sections, see Egbert, Burch, & Biber, 2020 for a detailed discussion of measuring lexical dispersion across linguistically meaningful corpus parts).

- Determine the size of each sub-section of the corpus and normalize it against the overall size of the corpus to arrive at expected percentages of word forms while taking into account the size of each section;
- Determine the frequency with which word forms occur in a given section and normalize it against the overall number of their occurrences in the entire corpus, creating the observed percentage of use for each word form in each section;
- Compute all pairwise 'absolute' differences between the expected and observed occurrences of word forms across sections, summing these differences and dividing them by two.

Following this, to further examine the patterns of vocabulary use highlighted by DP, key words in each section of the corpus were identified. This was done using Scott's (1996) concept of *keywords*. Keywords are words which have a special status in a given text, or register (Scott, 1996). A single word is considered a *keyword* if it occurs more frequently in a corpus than would be predicted by chance when compared to a reference corpus. The degree to which a word type is 'overused' in a target corpus compared to a reference corpus determines its 'keyness' value, which is typically measured by chi-squared or log-likelihood ratio tests (for further discussion see, Scott & Tribble, 2006). Since log-likelihood tests are thought to provide more accurate results than chi-squared tests when comparing the rates of occurrence of rare events, such as keywords (see Dunning, 1993), log-likelihood tests were used to measure the keyness values of words in the present study (also see Brezina, 2018, Chapter 3 for further discussion).

Separate listings of keywords were generated for each discipline in the corpus, with the written part of the British National Corpus (BNC)[4] as the reference corpus. Words were considered keywords if they (a) were used in the target discipline significantly more frequently than in the reference corpus, with the threshold for significance set at $p < .1E\text{-}7$, (b) appeared in at least 10% of texts in each discipline, (c) occurred at least 20 times per million words in each discipline, and (d) contained at least two alphabetical characters (excluding acronyms) (also see Egbert & Biber, 2019 for detailed discussions of different methods for deriving keywords). The percentage of overlap between the generated keywords lists was then calculated in a pairwise manner to assess the degree of commonality in the use of important words between different subsections of the corpus. This analysis resulted in a matrix of overlap which was then used as a basis of a hierarchical cluster analysis[5] (see Durrant, 2009 for applications of the same method). Cluster analysis is an *unsupervised* statistical method whose goal is to ascertain whether observations fall into relatively distinct groups without taking any a priori assumptions regarding their relations (see James, Witten, Hastie, & Tibshirani, 2013, p. 389). Hierarchical clustering can be classified into two main categories: hierarchical agglomerative clustering and divisive clustering (see James, et al., 2013, p. 393). Hierarchical agglomerative clustering uses a bottom-up approach to build a hierarchy of clusters based on the homogeneity of data points in each group. In contrast, the less commonly-used divisive clustering is based on a top-down approach in which all data points are grouped under a single cluster and then heterogeneous clusters are separated in each iteration. Since the implications of the agglomerative clustering method proved more practical for describing our data in terms of their quantitatively defined commonalities, this technique was selected for the cluster analysis in this study.

3. Analysis and results

3.1. Lexical dispersion in empirical research writing across disciplines

Once the lists of high-frequency words were retrieved from the corpus, the first step in the analysis was to assess the extent to which frequent words in the corpus were evenly distributed across different disciplines. As was mentioned in the previous section, Gries (2008) proposed dispersion measure (DP) was used for this purpose. DP scores for all frequent words in the corpus were calculated and then normalized based on the DP normalization method proposed by Lijffijt and Gries (2012). The normalization procedure was carried out to preclude any chance of DP not reaching its maximal value. DP is maximal when all occurrences of the target word are found in the smallest part of the corpus under analysis (see Lijffijt & Gries, 2012). Factors such as the number of subsections of the corpus or variation in their sizes have the potential to prevent DP from reaching its maximal value. Therefore, to guard against such potential disruptions, all DP values in this study were normalized using the following formula (adopted from Lijffijt & Gries, 2012):

$$DP_{norm} = DP/1-\min(s)$$

where *min(s)* is the size of the smallest part of the corpus normalized against the overall corpus size.

The obtained DP_{norm} values for high-frequency words in the corpus were found to vary substantially, ranging from 0.031 to 0.973 with values close to 0 reflecting an even distribution of the words across disciplines and values close to 1 representing their uneven distribution (see Gries, 2008 for a complete discussion). To aid interpretation, all high-frequency words in the corpus were classified into four major categories based on their DPs: (1) extremely evenly distributed, with a DP_{norm} below

[4] The BNC is a 100-million-word collection of over 4000 samples of English (see http://www.natcorp.ox.ac.uk/corpus/index.xml). The corpus has been widely used in linguistic research in the past two decades. The BNC has the advantage of being one of the largest English corpora freely available to researchers (see Hawtin, 2018 for further discussion).

[5] All statistical analyses in this study were conducted using R (R Core Team 2014. http://www.R-project.org/).

0.25, (2) moderately evenly distributed, with a DP_{norm} ranging from 0.25 to 0.499, (3) moderately unevenly distributed, with a DP_{norm} ranging from 0.5 to 0.749, and (4) extremely unevenly distributed, with a DP_{norm} value of 0.75 or above. Table 2 presents the percentage of words falling within each dispersion range, along with a random selection of words from each.

As can be seen from Table 2, the extreme point at the far left of the continuum represents instances of 'general' vocabulary (*for, of, and, however*), which account for 12.1% of all the words in the corpus. These words clearly distinguish themselves from those residing at the opposite end of the continuum (i.e., extremely unevenly dispersed), which appear to be more technical in nature. These are specialized words (e.g., *entrepreneurial, specimen, precipitation, solar*) which seem to be key in the construction and transmission of field-specific meanings in research writing. These words were found to make up more than one-fifth of high-frequency words in the corpus. Apart from these two extremities of the continuum, the level of specificity appears to slightly vary between the words grouped under the second and third categories (i.e., moderately evenly dispersed and moderately unevenly dispersed, respectively). Understandably, those words that were found to be moderately unevenly distributed seem to be relatively more specific than those with moderately even distributions. As shown in Table 2, the former category (i.e., moderately unevenly dispersed) accommodates about two-fifth of words in the corpus. Classifying the four categories into two broad groups of 'more widely dispersed' (i.e., extremely and moderately evenly dispersed) and 'more narrowly dispersed' (i.e., extremely and moderately unevenly dispersed), we find that the latter group contains 1.6 times (61.5%) more items than the former (38.2%).

Table 2
Examples of high frequency words grouped under four different dispersion classifications.

Cline of dispersion			
More widely dispersed		More narrowly dispersed	
Extremely evenly dispersed	Moderately evenly dispersed	Moderately unevenly dispersed	Extremely unevenly dispersed
0–0.249	0.25–0.499	0.5–0.749	0.75–1
828 (12.1%)	1791 (26.2%)	2698 (39.4%)	1513 (22.1%)
for; of; the; between; also; than; study; can; may; they; has; other; been; however; in; that; with; we; it; two; such; our; when; one; all; results; research; but; time; used; both; each; using; different; together; paper; taken; despite; yet high; level; effect; studies; only; based; social; there; first; effects; into; because; table; while; about;	events; obtained; complex; power; showed; age; features; network; method; function; strength; activities; respondents; items; initial; risk; states; experimental; interactions; focus; samples; status; characteristics; experience; context; activity; respectively; period; measure; quality; year; measures; relationship; values; observed; value; shown; positive; findings; support; participants; rate; range; hypothesis; growth; variable; structure; state; resources; behavior; theory; literature	disproportionate; constructions; innovation; undertake; stakeholders; pathway; synthesis; healthcare; governance; crises; modulation; monitored; ended; metallic; transactions; challenged; elicit; developers; controversy; negotiations; transverse; reputational; crossed; charges; casting; comparability; globalization; compensatory; comprehend; force; experiments; concentrations; containing; velocity; matrix; wave; sector; properties; organizational; conflict; complexity; maximum; ties; yield; curve; maintenance; compliance; commitment; zone; boundary	distilled; districts; longitude; propagating; subsidiary; republican; transnational; party; language; cells; teachers; strain; writing; learners; species; fracture; protein; news; binding; crack; entrepreneurs; entrepreneurial; load; precipitation; periodontal; specimen; solar; thickness; grain; electoral; deformation; acid; voting; earnings; politicians; toughness; immigration; gene; instruction; flux; campaign; war; democratic; radical; winter; neurons; military; resin; violence; ideological

This indicates the importance of this group of vocabulary in the production of scientific knowledge in empirical research articles across academic disciplines.

Table 3
Proportion of dispersion categories across disciplines.

	% Widely dispersed	% Narrowly dispersed
Biology	61.1	38.9
Chemistry	58.1	41.9
Dentistry	57.8	42.2
Mechanical Engineering	63.2	36.8
Physics	58.9	41.1
Mean	59.8	40.1
Minimum	57.8	36.8
Maximum	63.2	42.2
Applied linguistics	73.9	24.1
Business	73.3	24.7
Management	77.3	22.7
Politics	73.5	26.5
Sociology	75.7	24.3
Mean	74.4	24.4
Minimum	73.3	22.7
Maximum	77.3	26.5

To further examine these results, words grouped under these two broad dispersion categories (i.e., 'more widely dispersed' and 'more narrowly dispersed') were analyzed in terms of their degree of concentration in each discipline in the corpus. Table 3 presents the percentages of each category across disciplines.

Table 3 shows that there are differences in the extent to which writers in different disciplinary fields draw on words from the two categories. On average, widely-dispersed words appear to outnumber narrowly-dispersed items across all disciplines. However, words grouped under the former category appear to account for only 60% of high-frequency word types in the hard sciences. As shown in the table, two-fifth of individual words in research papers written in the hard disciplines were found to have a narrower and less widely distributional range. This means that, for every five high-frequency individual words in these disciplines, two had a rather narrow distribution characteristic. In contrast, widely-dispersed words were found to have a greater concentration in the word lists retrieved from the soft sub-corpora, accounting for 74.4% of word types in these disciplines. Words with a narrower range in these disciplines, however, were found to have a mean percentage of 24.4%, which is about 1.65 times less than that in the hard sciences. This variation would suggest that research writing in the hard fields demands the knowledge of a type of vocabulary that has a relatively narrow range of use and applicability.

3.2. Overlap in key vocabulary between disciplines

To further explore the observed disciplinary variation, a list of words that were found to be 'key' in each discipline was generated. Following this, the percentage overlaps between the keyword lists of each discipline were computed. This procedure yielded a symmetric matrix representing the degree of overlap in the use of key vocabulary between all disciplines in the corpus. Table 4 provides more information about this matrix.

Table 4
Overlaps in key vocabulary between disciplines in the corpus.

	BI	CH	DE	ME	PH	AL	BU	MA	PO	SO	Mean	SD	Max	Min
BI	---	51.1	40.6	36.8	36.1	26.5	29.6	31.1	27.9	30.6	34.5	7.7	51.1	26.5
CH	51.1	---	44.3	41.8	40.8	21.8	22.5	23.5	21.8	22.4	32.2	12	51.1	21.8
DE	40.6	44.3	---	36.6	31.2	23.6	25.4	25.9	24.5	25.5	30.8	7.8	44.3	23.6
ME	36.8	41.8	36.6	---	44.9	24.9	24.6	26.1	24.3	24.7	31.6	8.4	44.9	24.3
PH	36.1	40.8	31.2	44.9	---	21.2	24.6	24.4	25.7	27.5	30.7	8.2	44.9	21.2
AL	26.5	21.8	23.6	24.9	21.2	---	44.0	44.8	38.0	38.7	31.5	9.7	44.8	21.2
BU	29.6	22.5	25.4	24.6	24.6	44.0	---	66.7	49.0	50.6	37.4	15.7	66.7	22.5
MA	31.1	23.5	25.9	26.1	24.4	44.8	66.7	---	45.0	48.9	37.4	14.9	66.7	23.5
PO	27.9	21.8	24.5	24.3	25.7	38.0	49.0	45.0	---	50.9	34.1	11.7	50.9	21.8
SO	30.6	22.4	25.5	24.7	27.5	38.7	50.6	48.9	50.9	---	35.5	11.9	50.9	22.4

BI: Biology CH: Chemistry DE: Dentistry ME: Mechanical Engineering PH: Physics AL: Applied Linguistics
BU: Business MA: Management PO: Politics SO: Sociology

A number of points stand out from the matrix. First, the overlap values between disciplines appear to correspond to the *hard* and *soft* distinction proposed by Biglan (1973), Hyland (2004), and Durrant (2017). As shown in Table 4, there is a low degree of overlap between hard and soft disciplines in terms of the keywords they commonly use.

Secondly, the degree of homogeneity in the use of key vocabulary appears to vary even within the hard and soft disciplinary domains, with the confluence of certain disciplines being stronger than others. For instance, there is a great deal of overlap between the key vocabulary used in biology and chemistry research papers. In fact, if we calculate the mean (40.42) and standard deviation (5.44) of overlap percentages for the cells belonging to the hard disciplines, we find that the percentage overlap between these two disciplines (i.e., chemistry and biology) is about two standard deviations above the average (z-score = 1.94), indicating their high degree of overlap compared to the other hard disciplines in the corpus. In addition, as can be seen from the matrix, mechanical engineering and physics also showed a greater degree of overlap with each other than with Biology, chemistry, and dentistry. Among the soft disciplines, business and management were found to show a stronger overlap with each other (z-score = 2.42) than they did with the other soft disciplines. And finally, applied linguistics was found to be an outlier among the soft disciplines in terms of key vocabulary use, as it showed the lowest average degree of overlap with these disciplines (M = 41.3).

In order to extend our evaluation of these commonalities to a more quantitatively-defined level, a hierarchical cluster analysis was carried out based on the above matrix of overlaps between disciplines (see Durrant, 2008, 2009 for similar applications of this approach). In doing so, Euclidean distance was used as the dissimilarity measure to determine the distances between disciplines based on their degree of overlap (see James, et al., 2013, p. 402). Disciplines were then grouped according to the *average linkage* method which computes all pairwise distance dissimilarities between clusters and calculates the *average* of these dissimilarities to produce a hierarchy of the target variables (for a discussion of linkage methods in hierarchical clustering see James et al., 2013, pp. 399–402). This analysis produced the hierarchical arrangement of the disciplines based on the degree of overlap in their use of key vocabulary, as shown in Figure 1.

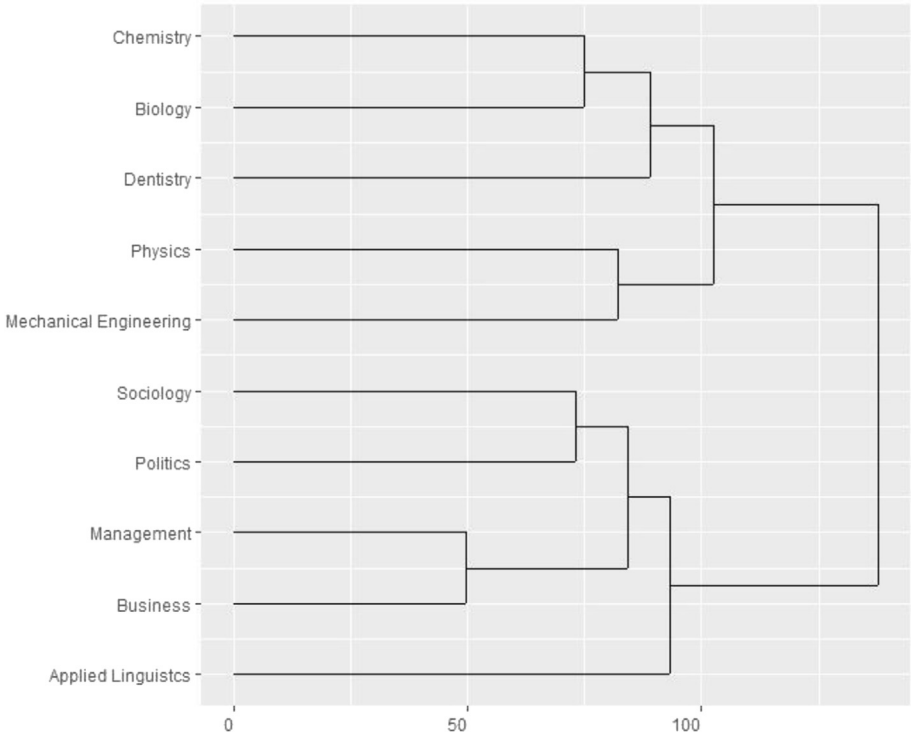

Figure 1. Cluster analysis of overlaps between disciplines.

As illustrated in Figure 1, the cluster analysis distinguished the hard sciences (top half) from the soft fields (bottom half). Reading the top half of the dendrogram from left to right, we can see that biology and chemistry are assembled within the same cluster at the first level of the analysis. At the second level, physics and mechanical engineering form their own cluster. Following this, the biology-chemistry cluster is joined by a single-point cluster formed by dentistry, which reflects a relative degree of homogeneity in the use of keywords between these disciplines. This group of three disciplines then joins up with the mechanical-physics cluster to form a macro group of disciplines, representing the broad category of hard knowledge fields.

Reading the bottom half of the clustering tree by the same logic, we find that management and business were the first to form a cluster due to their strong degree of overlap. In the next step, politics and sociology merge into their own cluster, which then joins up with the management-business group. This group of four disciplines then combines with applied linguistics, creating the broad category of soft knowledge fields. The fact that the single-point cluster created by applied linguistics joins the other four soft disciplines at the final steps of the analysis indicates its low degree of overlap with these disciplines and provides an indication of its outlier status as a soft field in terms of key vocabulary use. Together, these results suggest that the degree of homogeneity in the use of high frequency words that play a key role in the construction of knowledge in academic disciplines not only varies between hard and soft knowledge fields but also within these two broad disciplinary domains.

3.3. Lexical dispersion across sections of empirical research articles

A more nuanced view of these differences can be achieved if we further explore variation in research writing practices in terms of vocabulary use across different sections of research articles. To achieve this, the corpus was divided into four separate sub-corpora, each corresponding to the four main sections (i.e., IMRD) of the research article. High frequency words were then extracted from each sub-corpora based on the inclusion criteria mentioned earlier. This procedure was repeated for each section of research articles in each discipline, resulting in 40 separate listings of high frequency words for the ten disciplines in the corpus. Following this, Gries (2008) DP was again used to assess the extent to which high frequency words were evenly distributed across different sections of research articles in different disciplines. Table 5 presents the degree of dispersion for frequent words in the four main sections of research articles across the disciplines.

Table 5
Degree of dispersion for frequent words in IMRD sections.

	Range	More widely dispersed		More narrowly dispersed	
		0–0.249	0.25–0.499	0.5–0.749	0.75–1.00
Introduction	0.03–1.00	230 (9.7%) the, of, and, in, study, from, between, these, however	477 (20.2%) literature, et, al, attention, little, whether, known, practices	1035 (44.0%) status, phase, theories, supply, induced, modes, industry, intensity,	609 (25.8%) political, cell, academic, dental, deformation
Methodology	0.02–0.99	284 (8.4%) and, the, with, number, data, method, used, was, analysis, following	757 (22.3%) samples, information, performed, calculated, obtained, variable	1642 (48.5%) temperature, respondents, surface, protein, incubated, volume, pressure, gender	697 (20.6%) firm, buffer, corpus, steel, election, soil, vertical, purified, crack, texts
Results	0.02–0.94	297 (10.0%) and, the, significantly, table, different, higher, average, may	803 (27.0%) interaction, respectively, relationship, due, example, range, seen, per, evidence	1290 (43.4%) surface, region, fit, stress, density, items, pressure, wave, perceived, concentrations	580 (19.5%) fracture, protein, precipitation, thickness, race, solar, amplitudes
Discussion	0.02–0.99	363 (11.5%) the, and, from, explanation, consistent, possible, associated, in, line, supported	731 (23.2%) findings, et, al, quality, implications, determined, improve, contributes, enhance	1349 (42.9%) performance, social, risk, complexity, density, article, transfer, ratio, detected,	701 (22.2%) strain, reading, mechanical, angle, ion, writers, bond, affinity

As can be seen from Table 5, DP_{norm} values for high-frequency words in the IMRD sections vary substantially. On average, more widely dispersed words were found to accommodate about 30 percent of high-frequency word types across sections (M = 33.07%), whereas more narrowly dispersed words accounted for more than three-fifth of these words (M = 66.72%). These results indicate a considerable degree of unevenness across the four sections of research articles in terms of vocabulary use.

More differences can be identified by looking at these results in light of disciplinary preferences for the use of widely and narrowly dispersed words across the IMRD sections. For this purpose, the lists of high frequency words from the four sections of research articles across the ten disciplines were analyzed in terms of the extent to which they included items from the two dispersion categories. Tables 6–9 present the percentages of each category in the IMRD sections across disciplines.

Table 6
Proportion of dispersion categories across disciplines in the introduction section.

	% Widely dispersed	% Narrowly dispersed
Biology	64.29	35.32
Chemistry	64.52	35.18
Dentistry	61.66	38.00
Mechanical Engineering	59.79	40.21
Physics	57.94	39.97
Mean	61.64	37.74
Minimum	57.94	35.18
Maximum	64.52	40.21
Applied linguistics	65.50	34.50
Business	66.79	33.21
Management	66.96	33.04
Politics	68.56	31.44
Sociology	67.60	32.40
Mean	67.08	32.92
Minimum	65.50	31.44
Maximum	68.56	34.50

Table 7
Proportion of dispersion categories across disciplines in the methods section.

	% Widely dispersed	% Narrowly dispersed
Biology	48.47	51.53
Chemistry	42.04	57.95
Dentistry	45.45	54.55
Mechanical Engineering	51.06	48.94
Physics	54.24	45.76
Mean	48.25	51.75
Minimum	42.04	45.76
Maximum	54.24	57.95
Applied linguistics	74.48	25.52
Business	74.86	24.92
Management	77.50	22.27
Politics	77.63	22.37
Sociology	76.89	23.11
Mean	76.27	23.64
Minimum	74.48	22.27
Maximum	77.63	25.52

Table 8
Proportion of dispersion categories across disciplines in the results section.

	% Widely dispersed	% Narrowly dispersed
Biology	69.47	30.53
Chemistry	71.37	28.63
Dentistry	75.01	24.99
Mechanical Engineering	66.08	33.92
Physics	62.45	37.55
Mean	68.88	31.12
Minimum	62.45	24.99
Maximum	75.01	37.55
Applied linguistics	67.24	32.76
Business	73.32	26.68
Management	81.13	18.87
Politics	70.75	29.25
Sociology	69.19	30.81
Mean	72.33	27.67
Minimum	67.24	18.87
Maximum	81.13	32.76

Table 9
Proportion of dispersion categories across disciplines in the discussion section.

	% Widely dispersed	% Narrowly dispersed
Biology	70.15	29.85
Chemistry	68.54	31.46
Dentistry	66.95	33.05
Mechanical Engineering	65.30	34.70
Physics	63.70	36.30
Mean	66.93	33.07
Minimum	63.70	29.85
Maximum	70.15	36.30
Applied linguistics	67.27	32.73
Business	66.89	33.11
Management	67.69	32.31
Politics	71.95	28.05
Sociology	69.69	30.31
Mean	68.70	31.30
Minimum	66.89	28.05
Maximum	71.95	33.11

As can be seen from Tables 6, 8 and 9, lists of high frequency words from introduction, results and discussion sections in hard and soft disciplines contained similar proportions of individual words from the two categories. However, these proportions were found to substantially vary in the methods section. As shown in Table 7, on average, more than half of high-frequency word types in the methods sections in hard science research articles fall into the category of narrowly dispersed words (M = 51.75%). This is twice the proportion of only 23.64% found for the methods sections in the soft fields. In contrast, widely dispersed words were found to be more present in soft field methods, accounting for more than 75% of frequent individual words in this section. These results indicate that explaining research methods in hard knowledge fields demands a vast repertoire of a type of vocabulary which has a somewhat narrow range of use in research writing. Table 10 provides random examples of this vocabulary employed in the methods sections of hard science research papers in the corpus.

Table 10
Examples of narrowly dispersed vocabulary in the methods sections of science research papers.

	Narrowly dispersed	
	Moderately unevenly dispersed	Extremely unevenly dispersed
	0.5–0.749	0.75–1.00
Biology	temperature, incubated, protein, surface, site, diameter	purified, acetate, buffer, binding, diluted, centrifuged, resuspended,
Chemistry	energy, cells, protein, acid, reactions, concentration	buffer, species, equation, plasmid, ion, spectra, diluted, residue
Dentistry	acid, incubation, antibody, resin, temperature, protein, diameter	specimen, dental, implant, periodontitis, plaque, tissues, saliva
Mechanical Engineering	pressure, volume, velocity, temperature, displacement, laser	thickness, load, steel, equation, simulations, geometry, deformation
Physics	density, energy, waves, chamber, dynamics, intervals, spatial	flux, pulse, simulations, equation, spectra, solar, zonal, zenith, electrons

As Table 10 shows, there appears to be a certain degree of overlap in the use of narrowly dispersed words in Methods sections in these disciplines. To obtain a quantifiable measure of such commonality, the degree of overlap in the use of keywords was calculated for each of the sections across disciplines. To this end, separate listings of keywords for the IMRD sections across disciplines were generated based on the inclusion criteria for keywords mentioned earlier. Following this, the percentage overlaps between the keyword lists of each section across the ten disciplines were computed. This procedure yielded a symmetric matrix with 780 entries (i.e., $(40 \times 39)/2 = 780$), representing the degree of overlap between 40 keywords lists retrieved from the IMRD sections across ten disciplines. This procedure was then followed by a cluster analysis to identify potential groupings of disciplines based on the degree of their vocabulary overlap across the IMRD sections of the research article (i.e., IMRD).

3.4. Overlap in key vocabulary between disciplines across sections

Figure 2 plots the hierarchical clustering of disciplines based on the degree of overlap in key vocabulary use across IMRD sections.

Looking at Figure 2, the first point that stands out is that the clustering tree created by the analysis is divided into two main clusters: one belonging to the IMRD sections in the hard disciplines (top half) and the other representing the four sections in the soft fields (bottom half). Reading the top half of the dendrogram, we can see that there are two main groups comprising the hard sciences cluster: one created by the IMRD sections in mechanical engineering and physics and the other formed by the sections in biology, chemistry and dentistry. Figure 2 shows that each of these two groups is also composed of certain (sub)clusters. If we look at clusters in the former group, we find that results and discussions in mechanical engineering papers were among the first sections to form a merged pair. Interestingly, introductions and methods in this discipline created their own cluster, indicating a high degree of overlap between these two sections in mechanical engineering. At the next level of grouping for mechanical engineering, the results–discussions pair joins up with the introductions-methods cluster to represent key vocabulary use in this discipline. For physics, results and discussions were also the first to create their own cluster, which then combined with the methods sections in this discipline. This three-point cluster is then joined by a single-point cluster created by introductions in this discipline. The clusters created by the IMRD sections in mechanical engineering and physics then combined to form a stand-alone cluster for these two disciplines, representing a degree of similarity between these two fields in terms of key vocabulary use across the four sections.

Looking at the formation of (sub)clusters in the latter group (biology, chemistry, and dentistry), we can see that, at the first step of clustering in this group, results and discussions in chemistry formed a cluster. This pair is then joined by a cluster created by the same sections in biology, indicating a confluence of these two fields in terms of vocabulary overlap in results and discussions. This four-point cluster is then combined with a merged pair formed by introductions in biology and chemistry. As can be seen from Figure 2, for dentistry, these three sections (introduction, results and discussions) were found

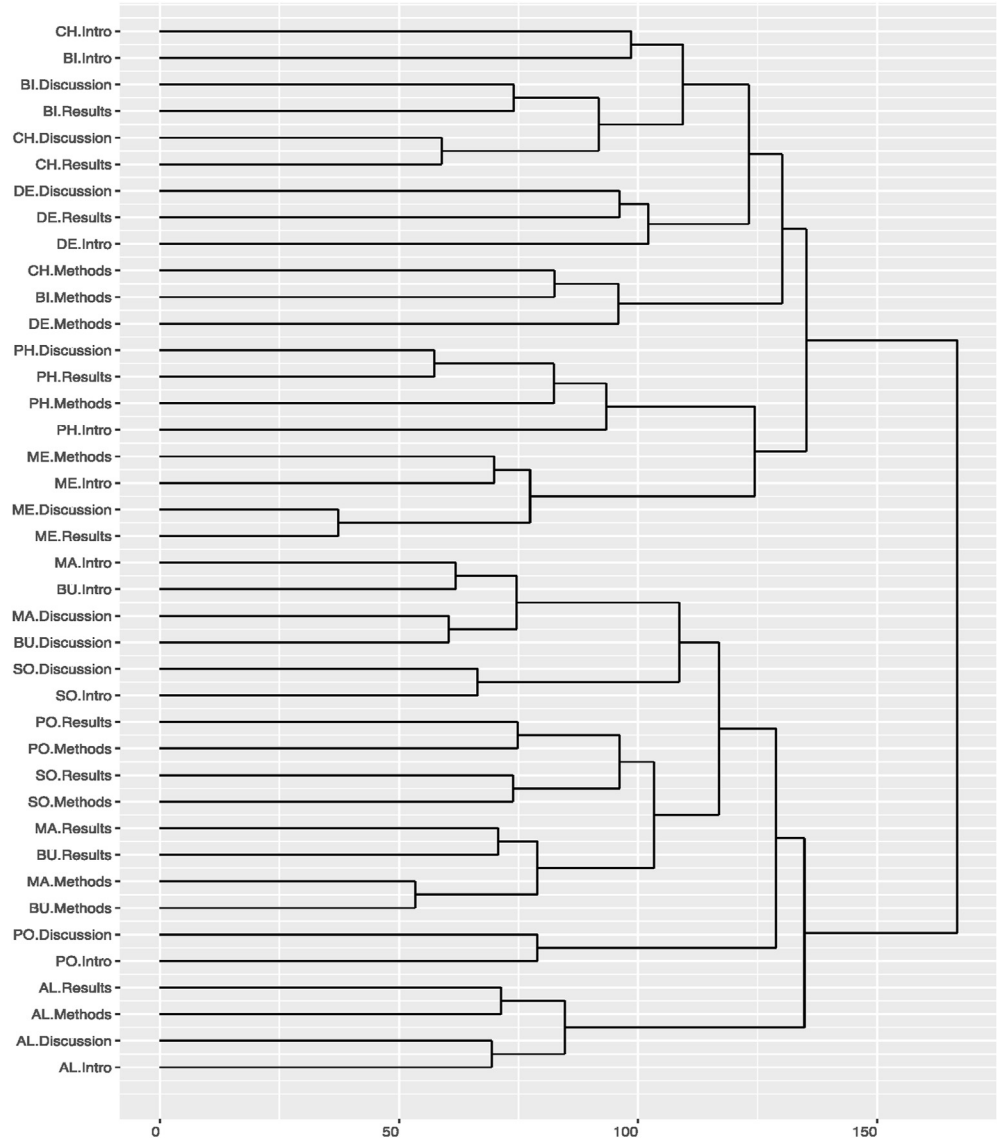

Figure 2. Cluster analysis of overlaps between disciplines across IMRD sections. BI: Biology CH: Chemistry DE: Dentistry ME: Mechanical Engineering PH: Physics AL: Applied Linguistics BU: Business MA: Management PO: Politics SO: Sociology.

to form their own cluster prior to merging with the group assembled by the same sections in biology and chemistry. Further, as shown in the dendrogram, there is a separate cluster for methods in biology, chemistry, and dentistry. This cluster is formed by methods in biology and chemistry creating a pair in the first stage of clustering and then joining up with methods in dentistry. This group of three then combines with the larger cluster formed by introductions, results and discussions across the three disciplines. This group in turn merges with that forged by the sections in mechanical engineering and physics, representing the broad category of the hard sciences.

Reading the bottom half of the dendrogram, we can see that the grouping of sections in the soft disciplines is rather different from that in the hard sciences. The first point that can be noted is that, unlike the hard sciences, results and discussions were not found to create a pair in any of the soft fields. Instead, it is introductions and discussions that appear to have a high degree of vocabulary overlap in research articles written in the soft disciplines. Interestingly, these two sections were found to make a pair in all the soft fields. A similar trend can also be observed for methods and results. These two sections were also found to show a high degree of commonality in terms of key vocabulary use across all the soft disciplines. Another interesting point is that applied linguistics was, once again, found to be an outlier among the soft fields. From Figure 2, we can see that the IMRD sections in applied linguistics separate themselves from those in the other soft disciplines by forging their own group first and then joining the other disciplines in the final round of clustering. This means that the four main sections of research papers in applied linguistics had the lowest average degree of *dissimilarity* with those in the other soft disciplines

only when they were grouped together, and not separately. Taken together, these findings lend support to the idea that there are various parameters of variation in vocabulary use across academic disciplines and that the hard-soft distinction can only represent one axis of this rather systematic variation.

4. General discussion

The present study aimed at investigating the commonalities and differences in the language of empirical research writing across disciplines. For this purpose, we examined specificity in the use of words in the main sections of empirical research articles written in ten different disciplines. In doing so, an inductive approach was adopted, through which specific patterns of vocabulary use were systematically identified and analyzed. The key strength of this approach is that it allows disciplinary-specific patterns to emerge from the analysis with minimal a priori assumptions guiding their identification. It also allows the researcher to triangulate and verify the emerged patterns using different methods of data analysis. It can be argued that such a methodological triangulation can yield results that can lay greater claim to generalizability, allowing the researcher to anchor findings in more robust theoretical interpretations (see, e.g., Baker & Egbert, 2016; Layder, 1993). Below, we discuss the findings that emerged from our analysis on the basis of the analytical steps taken at different stages of the study.

The first step in the analysis was to assess the degree to which frequently used words in research writing were evenly dispersed across the disciplines. The results revealed substantial variability (ranging from 0.031 to 0.973) in the dispersion values (DP_{norm}) for high-frequency words used in the corpus, indicating that the language of empirical research articles is characterized by highly specialized, field-specific discourses that are far from being homogeneous. These results stand in stark contrast to the theoretical position that views the language of academe as comprising an array of relatively homogeneous discourses, whose communicative functions are often realized by a set of shared linguistic items (Blue, 1988; Jordan, 1997; also see; Hutchinson & Waters, 1987). Our analysis showed that about 60 percent of the vocabulary used in the corpus were narrowly dispersed words (i.e., words that were used more frequently in one discipline than the other). These results not only highlight the importance of this group of vocabulary in the production of scholarly knowledge in empirical research articles, but also show that the majority of the high-frequency vocabulary used in research writing is not widely shared among academic disciplines, and so may well not be of equal value to writing in different branches of academia. It was further found that about two-fifth of individual words in research papers written in the hard disciplines were narrowly dispersed, that is, for every five high-frequency individual words in these disciplines, two had a rather narrow distribution characteristic. Further, the mean percentage of narrowly dispersed words in the sciences was found to be about 1.65 times more than that in the soft disciplines, implying that research writing in the sciences demands the knowledge of a type of vocabulary that has a narrow range of use and applicability. These results corroborate those of Durrant (2014), who found that university students' writing in Science and Technology fields makes greater use of specialized vocabulary, compared to Social Sciences and Humanities. The patterns observed in the present study extend Durrant's (2014) findings by indicating that disciplinary differences in the degree of reliance on specialized vocabulary (with narrow range of applicability) is even more pronounced in writing for research publication purposes. This indicates that conventional practices of disciplinary communities can have a profound impact on the lexical choices that members make in this particular genre of writing. Moreover, this finding also provides evidence in support of the argument that regards research writing as a highly conventionalized genre in which disciplinary knowledge plays an important role in adhering to the linguistic expectations of the target readership (see Bhatia, 1993; Hyland, 2001; Swales, 1990).

To further explore the observed disciplinary patterns, the next step in the analysis focused on the degree of overlap in the use of key vocabulary between all disciplines in the corpus. A number of key differences and similarities were observed. First, the overlap values between disciplines corresponded to the hard-soft classification established in previous research (Biglan, 1973; Durrant, 2017; Hyland, 2008). It is important to note that the disciplinary classifications proposed in these studies are based on parameters of variation other than single word vocabulary, with Biglan (1973) focusing on faculty members' judgements and perceptions of knowledge fields, and Hyland (2008) and Durrant (2017) basing their categorizations on the use of multi-word expressions, such as lexical bundles. Secondly, the degree of homogeneity in the use of key vocabulary was found to substantially vary even within the hard and soft broad divisions. For instance, mechanical engineering and physics showed a greater degree of overlap with each other than they did with biology, chemistry, and dentistry. Within the soft category, applied linguistics was found to be an outlier in terms of key vocabulary use, as it showed the lowest average degree of overlap with other soft fields (management, business, sociology, and politics). This offers support to Durrant (2017), who also found a general tendency for Humanities to be an outlier in multi-word vocabulary use in university students' writing. The observed patterns of overlap were further verified by the results of a cluster analysis. The cluster analysis provided a systematic mapping of the levels of commonality and variations between disciplines, illustrating that the degree of homogeneity in the use of high frequency words that play a key role in the construction of knowledge in academic fields not only varies between hard and soft disciplines but also within these two broad categories.

Our further investigation focused on exploring variation in research writing practices in terms of vocabulary use across different sections of research articles. The analysis showed that the dispersion of high frequency words substantially varied across the IMRD sections of research articles. On average, more narrowly dispersed words accounted for more than three-fifth of high-frequency word types across the sections, whereas more widely dispersed words were found to accommodate about 30 percent of these words. This finding suggests that there is a considerable degree of unevenness in single-word vocabulary use across the main sections of research articles. Further differences were identified by looking at these results in light of

disciplinary preferences for the use of widely and narrowly dispersed words across the sections. It was found that, compared to writing in Social Sciences and Humanities, science writing made greater use of narrowly dispersed words across all four sections (IMRD). These patterns of results parallel those observed in the first part of the analysis, which revealed science writers' reliance on less-widely used vocabulary, and extend these results by showing that this pattern of use is consistent across all the main sections of research papers. It was also observed that more than half of high-frequency word types in methods sections in hard science research articles were narrowly dispersed across the IMRD sections. This was found to be twice the proportion of only 23.64% found for methods sections in Social Sciences and Humanities. This particular observation suggests that explaining research methods in hard knowledge fields demands a vast repertoire of a type of vocabulary which can have a very narrow range of use in research writing. These findings also serve to illustrate that disciplinary variation can exist at various levels of specificity (e.g., in the overall text, across the article sections, across different parts of a single section of research articles, also see Cortes, 2013; Omidian, Shahriari, & Siyanova-Chanturia, 2018). Indeed, there is no one correct level at which such variation should be identified; rather, it depends on the goals of the researcher (see Biber & Conrad, 2009). However, it can be argued that investigating patterns of disciplinary use at different levels of granularity can result in more generalizable conclusions (see, e.g., Biber, Egbert, Gray, Oppliger, & Szmrecsanyi, 2016a).

Further, to explore possible commonalties between the sections of research articles in terms of vocabulary use across disciplines, the degree of overlap in the use of high-frequency keywords was calculated for each of the four sections across the ten disciplines. It was found that the degree of overlap between sections is also governed by disciplinary conventions. More specifically, we observed that introductions and discussions had a high degree of vocabulary overlap in research articles written in Social Sciences and Humanities. However, in the sciences, it was the results and discussion sections that showed a high degree of commonality in terms of keywords. This finding indicates that there are fundamental differences in how justification of findings and making claims about their significance are handled across disciplinary fields. This provides support for Hyland, 2004, 2008) assertion that disciplines can characteristically vary in how they convince the reader to assent to a particular interpretation or a knowledge claim in their research papers. The patterns of results revealed in this part of our analysis suggest that the communicative purposes of discussion sections written in the soft fields are often realized through the lens provided by the arguments already constructed in the Introduction section. In comparison, the discussion of findings and the claims about their importance in hard science research articles appears to be primarily based on the observations described in the Results section. In other words, while justification of results in Social Sciences and Humanities is realized through situating them within a theoretical basis often established in the Introduction section, discussion of findings and their validity in the sciences is often carried out by reiterating their grounded, experimental basis and anchoring them in empirical observations, rather than theoretical interpretations. It can, therefore, be argued that creating a convincing discourse for research findings in science writing, in comparison to writing in the soft fields, is more empiricist and less interpretative in nature.

Finally, applied linguistics was, once again, found to be an outlier among the soft fields, with the IMRD sections in this discipline forging their own group first and then joining the other disciplines in the final round of clustering. This suggests that, for writing the main sections of their research papers, writers in the field of applied linguistics appear to need an inventory of lexis which is not often shared among other soft knowledge fields.

5. Conclusion

Empirical research writing is one of the central activities of academic institutions. Scholars in various academic fields share and disseminate the outcome of their scientific endeavors through writing and publishing academic journal articles. It is through this particular form of knowledge dissemination that a rich body of scientific knowledge about a given phenomenon is accumulated. The results presented in this study show that the lexical choices academics make in the process of composing their research papers is differentially affected by the standards and conventions of scholarly activities in their field. It was found that such conventions have the potential to govern the delineation of authors' linguistic decisions at the most basic levels, such as the lexis.

The present research also attempted to show the strength of inductive methods and methodological triangulation in highlighting and verifying data-driven patterns of disciplinary writing. Future research could adopt such methodological approaches to shed further light on various aspects of research writing across disciplines.

Acknowledgements

We wish to thank Phil Durrant for his helpful comments on an earlier draft of this paper.

References

Anderson, A., & Valente, J. (2002). *Disciplinarity at the Fin de Siècle*. Princeton: Princeton University Press.
Baker, P., & Egbert, J. (2016). *Triangulating methodological approaches in corpus linguistic research*. New York: Routledge.
Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science*. London: University of Wisconsin Press.
Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. London: Longman.
Biber, D., & Conrad, S. (2009). *Register, genre, and style*. New York: Cambridge University Press.
Biber, D., Egbert, J., Gray, B., Oppliger, R., & Szmrecsanyi, B. (2016a). Variation versus textlinguistic approaches to grammatical change in English: Nominal modifiers of head nouns. In M. Kytö, & P. Pahta (Eds.), *Handbook of English historical linguistics* (pp. 351-375). New York: Cambridge University Press.

Biber, D., Reppen, R., Schnur, E., & Ghanem, R. (2016b). On the (non) utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics, 21*(4), 439-464.
Biglan, A. (1973). The characteristics of subject matter in different academic areas. *Journal of Applied Psycholinguistics, 57*(3), 195-203.
Blue, G. (1988). Individualising academic writing tuition. In P. Robinson (Ed.), *Academic writing: Process and product* (pp. 95-99). Hong Kong: Modern English Publications & The British Council.
Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press.
Bucher, R., & Strauss, A. L. (1961). Professions in process. *American Journal of Sociology, 66*, 325-334.
Corson, D. (1997). The learning and use of academic English words. *Language Learning, 47*(4), 671-718.
Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes, 12*(1), 33-43.
Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.
Coxhead, A. (2020). Academic vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 97-110). New York: Routledge.
Coxhead, A., & Nation, I. S. P. (2001). The specialized vocabulary of English for academic purposes. In J. Flowerdew, & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 252-267). London: Cambridge University Press.
Curry, M. J., & Lillis, T. (2017). Problematizing English as the privileged language of global academic publishing. In M. J. Curry, & T. Lillis (Eds.), *Global academic publishing: Policies, perspectives and pedagogies* (pp. 1-20). Bristol: Multilingual Matters.
Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics, 19*(1), 61-74.
Durrant, P. (2008). *High-frequency collocations and second language learning*. England: The University of Nottingham (Unpublished doctoral dissertation).
Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes, 28*(3), 157-169.
Durrant, P. (2014). Discipline and level specificity in university students' written vocabulary. *Applied Linguistics, 35*(3), 328-356.
Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? *English for Specific Purposes, 43*, 49-61.
Durrant, P. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics, 38*(2), 165-193.
Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora, 14*(1), 77-104.
Egbert, J., Burch, B., & Biber, D. (2020). Lexical dispersion and corpus design. *International Journal of Corpus Linguistics, 25*(1), 89-115.
Elbow, P. (1991). Reflections on academic discourse: How it relates to freshmen and colleagues. *College English, 53*(2), 135-155.
Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305-327.
Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics, 13*(4), 403-437.
Hawtin, A. (2018). The British national corpus revisited: Developing parameters for written BNC2014. Available at: http://www.birmingham.ac.uk/Documents/collegeartslaw/corpus/conference-archives/2017/general/paper39.pdf.
Hoskin, K. (1993). Education and the genesis of disciplinarity: The unexpected reversal. In E. Messer-davidow, D. R. Shumway, & D. J. Sylvan (Eds.), *Knowledges: Historical and critical studies in disciplinarity* (pp. 271-304). Charlottesville & London: University of Virginia Press.
Hutchinson, T., & Waters, A. (1987). *English for specific purposes*. Cambridge: Cambridge University Press.
Hyland, K. (2001). Bringing in the reader: Addressee features in academic articles. *Written Communication, 18*(4), 549-574.
Hyland, K. (2004). *Disciplinary discourses*. Ann Arbor: University of Michigan Press.
Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21.
Hyland, K. (2009). *Academic discourse*. London & New York: Continuum.
Hyland, K. (2011). Disciplines and discourses: Social interactions in the construction of knowledge. In D. Starke-Meyerring, A. Pare, N. Artemeva, M. Horne, & L. Yousoubova (Eds.), *Writing in knowledge societies* (pp. 193-214). Colorado & South Carolina: The WAC Clearinghouse and Parlor Press.
Hyland, K. (2016). Academic publishing and the myth of linguistic injustice. *Journal of Second Language Writing, 31*, 58-69.
Hyland, K. (2018). *The essential Hyland: Studies in applied linguistics*. London & New York: Bloomsbury Publishing.
Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL quarterly, 41*(2), 235-253.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.
Jordan, R. (1997). *English for academic purposes*. Cambridge: Cambridge University Press.
Layder, D. (1993). *New Strategies in social research*. Cambridge: Polity Press.
Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes, 22*, 42-53.
Lijffijt, J., & Gries, S. T. (2012). Correction to Stefan Th. Gries' "Dispersions and adjusted frequencies in corpora". *International Journal of Corpus Linguistics, 17*(1), 147-149.
Lin, L., & Evans, S. (2012). Structural patterns in empirical research articles: A cross-disciplinary study. *English for Specific Purposes, 31*(3), 150-160.
Liu, D., & Lei, L. (2020). Technical vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 111-124). New York: Routledge.
Mannheim, K. (2013). *Ideology and utopia*. New York: Routledge.
Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes, 28*(3), 183-198.
Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
Omidian, T., Beliaeva, N., Todd, L., & Siyanova-Chanturia, A. (2017). The use of academic words and formulae in L1 and L2 secondary school writing. *New Zealand Studies in Applied Linguistics, 23*(2), 39.
Omidian, T., Shahriari, H., & Siyanova-Chanturia, A. (2018). A cross-disciplinary investigation of multi-word expressions in the moves of research article abstracts. *Journal of English for Academic Purposes, 36*, 1-14.
Omidian, T., & Siyanova-Chanturia, A. (2020). Semantic prosody revisited: Implications for language learning. *TESOL Quarterly, 54*(2), 512-524.
Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies*. Amsterdam/Philadelphia: John Benjamins Publishing.
Russell, D. (1991). *Writing in the academic disciplines: A curricular history*. Carbondale & Edwardsville: Southern Illinois University Press.
Scott, M. (1996). *WordSmith tools*. Oxford: Oxford University Press.
Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins Publishing.
Shumway, D. R., & Messer-Davidow, E. (1991). Disciplinarity: An introduction. *Poetics Today, 12*(2), 201-225.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Stoller, F. L., & Robinson, M. S. (2013). Chemistry journal articles: An interdisciplinary approach to move analysis with pedagogical aims. *English for Specific Purposes, 32*(1), 45-57.
Swales, J. (1990). *Genre analysis: English in academic and research settings*. New York: Cambridge University Press.
Swales, J. (2004). *Research genres: Exploration and applications*. New York: Cambridge University Press.

Taha Omidian is a PhD student at Victoria University of Wellington, New Zealand. He specializes in the use of corpus linguistic and computational methods to explore systematic patterns in language data. His research interests include corpus linguistics, computational linguistics, quantitative linguistic research methods, English grammar, register variation, vocabulary, phraseology, language learning, language for specific purposes, multilingualism, and academic writing.

Anna Siyanova-Chanturia is Senior Lecturer in Applied Linguistics at Victoria University of Wellington, New Zealand. Anna's research interests include psychological aspects of second language acquisition, bilingualism, usage-based approaches to language acquisition, processing and use, vocabulary and multi-word expressions, and quantitative research methods (corpora, eye movements, EEG/ERPs).