A multi-dimensional view of collocations in academic writing

Maria Carolina Zuppardi and Tony Berber Sardinha
São Paulo Catholic University

This chapter discusses the identification of the major sets of interrelated collocations in academic writing across different disciplines, or dimensions of collocation. A corpus of textbooks and research articles from Social, Behavioral and Economic Sciences, containing 10.6 million words across 230 texts was analyzed using a collocational multi-dimensional analysis framework (Berber Sardinha 2017). Results show four dimensions of collocation for Social, Behavioral and Economic Sciences, namely: (1) Human Nature, Culture and Research Methods vs. Economics; (2) Human Evolution and Society; (3) Business and Finance; and (4) Statistics. The chapter highlights a methodology for identifying collocational patterns through multi-dimensional analysis, and contributes to a better understanding of academic writing in Social, Behavioral and Economic Sciences.

1. Introduction

Instructional approaches in English for Academic Purposes (EAP) assume that the linguistic needs of learners vary according to different language tasks and academic disciplines (Biber 2006). Vocabulary plays a central role in EAP, and there has been increasing interest in corpus-based research that identifies the major vocabulary items in EAP contexts. Indeed, when defining measures of lexical proficiency, Crossley et al. (2015) ranked the ability to use collocations ("the co-occurrence of two items in a text within a specified environment" (Sinclair, Jones & Daley 1970/2014: 10); the "comings-together-of-words", (Palmer 1933: 1 in Barnbrook, Mason & Krishnamurthy 2013: 24) as the most important index of vocabulary development. The role of collocation in developing vocabulary competence in EAP has been recognized and previous research on collocations in academic writing has focused on producing collocation lists deemed important for students of English for Academic Purposes to be successful in their studies (Durrant 2009; Ackermann & Chen 2013; Lei & Liu 2018). However, research has shown that collocations can

https://doi.org/10.1075/scl.95.14zup

form patterns that guide topical coherence (cf. Berber Sardinha 2017), highlighting the need to move beyond lists of single collocations to groups of interrelated collocations that are shared across different topics and registers.

Durrant, in an effort to analyze "positionally-variable academic collocations […] across a wide range of disciplines" in academic writing (2009: 157), created a 2-word cross-disciplinary collocation list containing both fixed and variable items as well as both lexical and functional words. The corpus was comprised of research articles published in peer-reviewed journals and the disciplines were defined through clustering of the schools and faculties of the researcher's own university, resulting in five major discipline areas: Arts and Humanities; Life Sciences; Science and Engineering; Social-Administrative; and Social-Psychological. Academic collocations were defined as "word pairs which co-occur with at least moderate frequency across a wide range of academic disciplines, but which are not often found in non-academic language" (2009: 161–2). The final list contained the 1,000 "most distinctively academic collocations" (2009: 163). Durrant (2009) attempted to address two issues that are commonly raised in academic collocations research: (1) to what extent collocations are distributed across disciplines; and (2) multi-word collocational patterns within a specific window span. However, limitations become apparent in his study, for example the fact that the inclusion of both lexical and function words resulted in the majority of the collocations in the list being incomplete, grammatical collocations (e.g., *and respectively* and *can be*), whose pedagogical value may be questioned.

Ackermann and Chen (2013) also took a corpus-driven approach to derive a pedagogically-relevant Academic Collocation List (ACL); however, there were two major differences from Durrant's approach: (1) only lexical collocations were extracted so there were no structurally incomplete collocations, and (2) there was an added step of having the list reviewed, reduced and vetted as pedagogically relevant by a panel of expert judges. The corpus used was the Pearson International Corpus of Academic English (PICAE), which contains a range of academic registers, both written and spoken, from five English-speaking countries, and it is divided into four major discipline areas: Applied Sciences and Professions; Humanities; Social Sciences; and Natural/Formal Sciences. The authors defined collocation as "a single word that tends to co-occur in the span of ± 3 words from the reference word, co-occurring at least five times in total across at least five different texts with a [MI] score of at least 3 and a t-score of at least 2" (2013: 237). The collocation extraction and final list creation involved the following steps: (1) a list of node words that appeared at least five times in at least five different texts was extracted from the list of all content words in the corpus, excluding those in common with the General Service List (West 1953); (2) using the node list, potential collocations were extracted resulting in over 130,000 items; (3) the list was reduced based on

the strength of association cut-offs established, leaving 16,174 items as potential entries in the final list; (4) a further reduction was achieved by selecting lexical collocations of only these types: verb + noun, adjective + noun, adverb + adjective, and adverb + verb, and keeping only free and restricted word pair combinations; (5) the resulting 6,808 items were further assessed by the two researchers and reduced based on combination criteria related to particular types of meanings (e.g., concrete geographical references, common transparent adjectives, among others), degree of fixedness, and completeness; (6) the final 4,558 items were then manually reviewed and vetted by a panel of six experts (four professors, a dictionary consultant, and a lexicographer/publisher), who systematized and reduced the final list to 2,468 collocations. It is undisputed that the ACL is regarded as beneficial and relevant for learners of ESL and EAP and it contains more than double the items in Durrant's (2009) list. Nevertheless, a major limitation may well be what to the authors is a distinguishing factor in their methodology: using subjective human judgement (researchers and a panel of experts) to screen and reduce the list by more than 60% may have left out collocations they deemed idiomatic and irrelevant for students, but whose prevalence in corpora may still be high and therefore of relevance to English learners. There was also moderate inter-rater reliability (Intraclass Correlation Coefficient of 0.524), which further weakens the argument in favor of human judgement (Ackermann & Chen 2013: 240).

Following a different approach, Lei and Liu (2018) developed an academic English collocation list (AECL) aimed at identifying lexical collocations the authors considered "most useful to ESL/EFL learners" (2018: 216). The corpus used was comprised of texts from three existing corpora (BNC Academic, BAWE, and JDEST) and from three corpora compiled by the researchers (containing published journal articles, doctoral dissertations in applied linguistics, and book reviews) (2018: 222). The final corpus covered a wide array of registers, ranging from published and unpublished academic and research texts (such as textbooks and journal articles) to graduate and undergraduate student writing, as well as five discipline areas: natural sciences, applied sciences/engineering, social sciences, applied social sciences, and humanities. In spite of their corpus having 43.1 million words across a rich variety of registers and disciplines, the authors chose a top-down approach to select the nodes for collocation extraction, which were the 3,015 content words (nouns, verbs, adjectives, and adverbs) from Gardner and Davies's Academic Vocabulary List (2014). The corpus was queried for each node word and word pairs were extracted following collocation patterns of interest (for example, adjective + noun, verb + noun, or noun + noun). Rather than using window span, Lei and Liu (2018) used syntactic dependency relations to "make sure that the extracted combinations are true collocations of the types" (2018: 226) of interest. According to the authors, in addition to ensuring the collocations belong to the types of structures

they desired, dependency relations may help identify additional word pairs that could be missed by window span restrictions. Collocations were then extracted based on the dependency relations, strength of association cut-offs (MI score of at least 3; t-score of at least 2), as well as frequency and dispersion measures (frequency: 10 occurrences pmw; dispersion: 0.2 occurrence in each discipline). The resulting list, containing 9,208 collocations, was then manually reviewed for exclusion of proper nouns or duplicates and harmonizing of conflicting spellings, leaving a total of 9,049 collocations in the final list.

The collocation lists described above have been instrumental in supporting EAP teaching and learning; nevertheless, such lists lack information that is arguably important to language learners such as how the collocations might co-occur with one another and form networks, what general topics they are used to talk about and what registers they are associated with.

The present study takes a statistical view on collocation, whereby the goal is to identify what Sinclair, Jones and Daley (1970/2014: 10) called "significant collocation", that is, "regular collocation between two items, such that they co-occur more often than their respective frequencies, and the length of text in which they appear, would predict". This is achieved by calculating an association measure that gauges the attraction between a node word (the word under scrutiny) and its neighboring collocates (the words surrounding the node). This Sinclairean perspective is arguably the prevailing view of collocation in corpus-based studies, but there are exceptions, such as by taking into consideration syntactic dependence relation analysis to extract word pairs of the dependency relations of interest (e.g., Lei & Liu 2018).

The common practice in the collocations literature has been to analyze individual collocations in a corpus without regard for how collocations vary across registers, understood here as "a cover term for any language variety defined by its situational characteristics, including the speaker's purpose, the relationship between speaker and hearer, and the production circumstances" (Biber 2009: 823). In addition, the norm has been to see collocations as independent of each other, as if they did not often inter-collocate, that is, collocate with each other (cf. Phillips 1989; Berber Sardinha 2017; Berber Sardinha, Mayer & Ferreira in press). By acknowledging the principle of inter-collocation, manifested by collocates of a particular collocation overlapping to some degree with the collocates of a different collocation, it becomes possible to identify sets of collocations, that is, groups of nodes that share to some degree a number of different collocates. In addition, this enables us to see to what extent these sets of collocations are distributed across registers, in order to determine the degree of the relationship between collocation and register. So far, there is some evidence of the relationship between registers and collocations in the literature, as in Berber Sardinha (2017), which showed register could discriminate among the collocations in the COCA corpus, but there is no

previous comprehensive research on the relationship between collocation use and academic registers such as research articles and college textbooks.

Since collocations occur in differing situational contexts, such as disciplines and registers, in addition to general collocation lists, it becomes imperative to identify the collocational patterns of variation underlying those disciplines and registers. None of the studies into academic collocations carried out to date has accounted for differences in the registers that comprised their corpus of study. Instead, their focus has been on identifying collocations that are characteristic of a given register or genre and describing their occurrence, thus "little empirical evidence is available about the degree to which genres, registers, and modes of communication affect the strength of association between words" (Gablasova et al. 2017: 167). Gablasova et al. (2017), focusing on the use of collocations in language learning research, provide a comprehensive critical review of corpus-based formulaic research, including the application of strength of association measures, comparison of collocational use across corpora with varying social and situational characteristics, as well as interpretation of collocation patterns produced by L1 and L2 speakers. Gablasova et al. (2017) claim that, just as frequency and co-occurrence of specific words will vary according to different registers or linguistic settings (cf. Biber et al. 1999), so should collocational strength in different (sub)corpora representing different registers and modality. The results of an analysis of three types of collocational patterns across the BNC (comparing the full corpus and a subset of its subcorpora) using t-score, MI, and logDice seem to support the assumption that formulaicity, i.e., strength and frequency of collocational patterns, will vary across genres, registers, and modes of communication (Gablasova et al. 2017: 169–70). The authors also indicate that beyond identification of collocations through association measures, "a subsequent structural and semantic analysis of collocations identified by high or low [association measure] scores can be crucial for getting a clearer picture of collocational patterns associated with certain groups of speakers or a certain type of word co-selection" (Gablasova et al. 2017: 173). In addition to groups of speakers or word co-selections, we argue that further semantic analysis of collocational patterns is crucial to understanding use of collocations in varying registers and disciplines as well as how sets of interrelated collocations guide topical coherence in texts (cf. Berber Sardinha 2017).

This chapter describes the results of a study aimed at identifying the major sets of interrelated collocations, or dimensions of collocation, in two registers of academic writing across different disciplines. The principle underlying the dimensions of collocations is inter-collocation, i.e., the fact that collocations collocate among themselves. This study attempts to identify dimensions of collocation through a collocational multi-dimensional (MD) analysis (Berber Sardinha 2017; Berber Sardinha, Ferreira & Mayer in press; Berber Sardinha, Mayer & Ferreira in press).

It can be argued that a collocational MD analysis provides a way to supply not only a data-driven lexical collocation listing respecting register and discipline constraints, but also patterns of collocational co-occurrence across disciplines and registers, through a text-linguistic approach, that could be even more relevant for pedagogical use in EAP. We expect that teachers and learners of EAP can benefit from knowing the sets of correlated collocations (dimensions) that reflect the major topics and subject-matters across disciplines and registers in academic English.

Unlike most common methods where collocations are usually extracted and analyzed individually from corpora with no text boundaries, a collocational MD analysis allows for retrieval of a large number of collocations respecting register and text boundaries. Although it has been established that "patterns of linguistic variation and use are dramatically different across registers" (Biber 2012: 39), most previous research on collocations have disregarded register differences. Studies that looked at register from a collocational perspective are rare. An early example is Sinclair, Jones and Daley (1970), who compared collocations found in a science magazine (New Scientist) and in conversation and found that the collocations could discriminate between the two registers: "from a linguistic point of view it is interesting to find that 'strength of collocation' provides a useful discriminant between different types of English and it would be interesting to see if the results were so encouraging for two texts which differ very little" (1970: 133). In this chapter, we follow their suggestion and explore the degree to which two closely related registers (research articles and university textbooks) can be distinguished on the basis of collocation. A comprehensive study of the relationship between register and collocations in English is Berber Sardinha (2017), which introduced the notion of dimensions of collocation to capture the sets of correlated collocations in English. This study identified nine different dimensions based on the Corpus of Contemporary American English (COCA), namely: 1. Literate discourse; 2. Oral discourse; 3. Objects, people, and actions; 4. Colloquial and informal language use; 5. Organizations and the government; 6. Politics and current affairs; 7. Feelings and emotions; 8. Cooking; and 9. Education.

As aforementioned, we know that collocations co-occur in texts and this repeated co-occurrence leads to collocation networks, such that the occurrence of a collocation in text selects other collocations nearby. The notion that collocations form networks has been recognized in the literature in various forms – for instance, Williams (1998: 157) used the term collocational networks to mean clusters of words sharing statistical collocates that embody the "multitudinous linkage potential of lexical items." Brezina, McEnery and Wattam (2015: 141–2) acknowledge that "collocation networks have the potential to provide us with an insight into important lexical connections in discourse." The idea of collocation networks relates to

Phillips's notion of 'lexical networks' (1989), where networks of collocating words reflect lexical patterns of discourse in a given text.

More recently, specific computer software has been developed with the aim of visualizing collocational networks graphically, such as the one developed by Brezina, McEnery and Wattam (2015). Their software can build collocation networks based on user-defined corpora and show graphically – through symbols and connecting lines – the relationships among nodes and collocates. However, it does not identify the collocations present in the uploaded corpus, rather it shows the collocation networks for a node word chosen by the user as a search word. The fact that the software displays the networks formed by a single word at a time provides a restricted view of collocational networks. In our understanding, collocational networks are not centered on any one single word, rather they are formed by the linkages among multiple words that share similar collocational environments.

Collocation networks are patterned for topical coherence, that is, the major discourses and topics discussed in the text guide collocational patterns. This is suggested by Biber (1993), who used factor analysis to detect word senses from word association patterns. According to Biber, "if we assume that texts are topically coherent, it follows that groupings of collocations that co-occur frequently in texts will often reflect different underlying word senses" (1993: 532). The study showed that each factor "represent[ed] a separate grouping of collocations that tend to co-occur frequently in texts" (Biber 1993: 536), that is, a different word sense. Biber (1993) identified the major collocational patterns for two target words: *certain* and *right*. The factorial structures for both words show a "highly systematic patterning among collocations" (Biber 1993: 536), where each factor reflected a grouping of collocations that shared the same word sense, for example, factor 2 for the word *right* contains collocations that refer to immediacy, i.e., *right* as "immediately", "directly", or "exactly", as in *right there* or *right now*.

In addition, collocation networks can be shaped by register constraints if particular collocations are found more often in particular registers (Berber Sardinha 2017). A collocational MD analysis is able to show the extent to which collocations can predict register categories or vice-versa, which previous studies have not been able to demonstrate. For example, Berber Sardinha (2017) showed that while collocation was generally a poor predictor of register differences in general English, registers in turn were good predictors of collocation. In the next sections, the collocational MD analysis framework employed in this study is presented, followed by a discussion of the dimensions of collocation in academic writing across Social, Behavioral and Economic Sciences disciplines.

2. Collocational multi-dimensional analysis

The corpus in this study was analyzed using a novel variant of the multi-dimensional (MD) framework introduced by Biber (1988), known as collocational MD analysis (Berber Sardinha 2017). The main goal of MD studies is to identify the underlying parameters of variation in the data. In a traditional functional MD analysis, the goal is to identify patterns of variation in texts across different registers based on co-occurrence of lexicogrammatical characteristics. Texts are the units of observation and linguistic features that may or may not be present in those texts are the variables. The texts being analyzed are tagged for those lexicogrammatical features and normed counts of the features present in the texts are submitted to a statistical procedure called factor analysis. In a factor analysis, the variables – in this case, the linguistic features present in the corpus – are reduced to a limited number of factors comprised of the set of variables that co-occur in the texts. Each factor is then analyzed qualitatively to identify a functional dimension that reflects one or more shared communicative functions underlying the patterns of co-occurrence of the linguistic features that comprise that factor. The final result is a set of dimensions that reflect the underlying patterns of variation in linguistic features across the registers being analyzed.

Similarly, the goal of a collocational MD analysis is also to identify patterns of variation; however, rather than analyzing the co-occurrence of linguistic features, it aims to determine the patterns of co-occurrence and the underlying use of collocations in texts across different registers. In this type of analysis, variables are the nodes and units of observation are the collocates. The texts being analyzed are tagged for part of speech and lemma and frequency counts are used to calculate a strength of association measure, which is then submitted to a factor analysis. A factor analysis is carried out to determine the statistical co-occurrence of nodes across all the collocates. Each factor is then analyzed qualitatively, and dimensions of collocation are identified and named based on analysis of underlying topics or subject-matter expressed, semantic fields, and semantic preference. Table 1 summarizes the main characteristics of the two types of MD analysis, functional and collocational.

The interpretation of collocational dimensions is carried out taking into consideration lexical and/or semantic field, semantic preference, and shared topics. At the level of the text or register, lexical or semantic field is defined as "a segment of reality symbolized by a set of related words [which] share a common semantic property. Most often, fields are defined by subject matter" (Brinton & Brinton 2010: 112). In this sense, a lexical or semantic field may be defined by the set of collocations that co-occur in a given text or register, which in turn contribute to

Table 1. Types of MD analysis

	Traditional MD (Biber 1988)	**Collocational MD (Berber Sardinha 2017)**
Unit of Analysis	Text	Node + Collocate
Measurement	Normed frequency	Strength of association measure
Variables	Lexico-grammatical features	Nodes
Rows of Observation	Texts	Collocates
Factor scores	Calculated for texts in the corpus	Calculated for collocates
Primary Interpretation	Functional and communicative	Semantic and topical preference

build topical coherence in that text. How a lexical or semantic field is defined highlights the importance of taking into account disciplinary differences in patterns of collocational variation. At the level of the word pairs, "semantic preference refers to what has traditionally been known as a lexical field: a class of words which share some semantic feature" (Stubbs 2007: 178). Semantic preference, "the restriction of regular co-occurrence of items which share a semantic feature" (Sinclair 1998: 16), is related to the topic of the surrounding co-text. In other words, the co-occurrence of lexical sets of semantically related words, may reflect and typify topics that are characterized by the relationships underlying the variation in collocational use across disciplines and registers.

3. Collocational multi-dimensional analysis of academic writing

The corpus used in this study is comprised of textbooks and research articles from Social, Behavioral and Economic Sciences, containing 10.6 million words across 230 texts. Texts were selected based on the subareas of Social, Behavioral and Economic Sciences as defined by the National Science Foundation (NSF),[1] namely: Behavioral and Cognitive Sciences (BCS) and Social and Economic Sciences (SES). For each of those two subareas, 15 textbooks and 100 articles from 10 different journals were collected. The Behavioral and Cognitive Sciences subarea encompassed the following subfields: Anthropology, Cultural Anthropology, Biological Anthropology, Archeology, Evolution, Social Psychology, Cognitive Psychology, Learning and Behavior, Cognitive Sciences. The following subfields comprised the Social and Economic Sciences subarea: Sociology, Political Economy, Applied Economics, Financial Studies, Political Analysis, Micro and Macroeconomics, Administrative

1. The list of areas and subareas can be found at: <https://www.nsf.gov/about/research_areas.jsp>

Sciences. Every effort was made to ensure that each subfield was equally represented and had the same number of texts represented in the corpus. Textbooks were collected based on the bestsellers list for textbooks on Amazon.com, for each subarea. Research articles were collected from the top-ranking journals in each area/subarea according to the SJR – Scimago Journal & Country Rank – Rankings.[2] These two registers were chosen because they are the most often cited registers in academic syllabi and hence the most likely to be encountered by students in higher education. Table 2 shows the composition of the corpus.

Table 2. Composition of the corpus

Register	Subareas	# of texts	# of tokens
Articles	Behavioral and Cognitive Sciences (BCS)	100	1,543,065
	Social and Economic Sciences (SES)	100	1,506,232
Textbooks	Anthropology; Political Science; Psychology; Sociology; Economics	30	7,575,507
	Total	**230**	**10,624,804**

For every textbook, the content of the introduction and of every chapter was kept, including text from tables, figures, key terms, exercises, cases studies, and endnotes; the following were excluded: boilerplates, acknowledgements, preface, references, appendices, and supplemental materials. For every article, boilerplates, acknowledgements, references, and supplemental materials were excluded and the following material was kept: abstract, keywords, full article content, end-of-article methods section (when applicable), footnotes and/or end-of-article notes (when present), and glossaries (when present). All texts were converted and cleaned up using computer scripts, and subsequently tagged for part of speech and lemma with the Tree Tagger (Schmid 1994). Manual verification of automatic tagging was carried out on a sample and corrections were performed in the whole corpus as needed. Only lemmatized nouns, adjectives, and verbs were considered for analysis. This means that only lexical collocations were analyzed since shared topics, topical coherence and semantic fields are expressed by lexical words rather than grammatical items. The study presented in this chapter follows the Sinclair tradition and defines collocation as a word pair that tends to occur in the span of up to 4 words on either side of the node (Sinclair, Jones & Daley 1970/2014: 13).

First, all of the word pairs occurring in a span of up to four words from each other in each text were identified using a computer script. Next, the strength of

2. <http://www.scimagojr.com/journalrank.php>

association for each of these word pairs was calculated using the logDice coefficient (Rychly 2008). The logDice statistic was chosen because it serves as a middle ground between the Mutual Information (MI) and t-score measures of association, capturing collocations for a wide range of word frequency (Berber Sardinha 2017). According to Hunston (2002: 71), "the MI-score is a measure of the strength of collocation", capturing rare word pairs with low frequencies in the corpus, and the "t-score is a measure of certainty of collocation", capturing the most frequent word combinations. In addition, unlike MI and t-score, the logDice formula does not include expected frequency, being a standardized measure on a delimited scale; therefore, it highlights exclusivity while avoiding the low-frequency bias found in the MI-score (Gablasova et al. 2017).

Word pairs with frequency equal to 1 were removed. Then, the 2,000 most frequent nodes that were verbs, nouns or adjectives in each register were extracted along with their collocates. Any overlapping lemmas and any acronyms were removed, resulting in a pool of 2,538 nodes and 45,311 collocates. An initial factor analysis was performed using SPSS v.23, with Principal Axis as the extraction method. As mentioned above, the nodes were the variables entered in the factor analysis. KMO measure was .996, indicating the sample is 'marvelous' for a factor analysis to be carried out (see Friginal & Hardy 2014).

Next, nodes with communalities of at least .3 were retained and entered into a rotated factor analysis, with Promax extraction, based on the optimal number of factors identified. The optimal number of factors was determined through analysis of the scree plot, a plot of the eigenvalues associated with each factor, and a four-factor solution was deemed the most interpretable after rotation. A total of 2,110 variables (nodes) remained in the analysis at this stage. Factor scores were then calculated for each collocate, by summing up the normalized logDice values of the nodes loading on each factor, and mean factor (or dimension) scores were computed for each register (article and textbook). Finally, each factor was interpreted, and the dimensions of collocation identified were named based on an analysis of the topics, subject-matters, semantic fields, and semantic preferences, through an examination of concordance lines and by extensively reading text excerpts. It is important to highlight that it is expected that nodes will share some of their collocates thereby fulfilling the concept of inter-collocation, that is, how they inter-collocate contributes to build the underlying topical coherence and semantic field that characterize a dimension, as exemplified in Table 3.

Table 3. Example of nodes sharing collocates

Nodes*	Collocate*	Collocation
human (j)	evolution (n)	human evolution
human (n)		evolution (of) human
biological (j)		biological evolution
cultural (j)		cultural evolution
language (n)		language evolution
evidence (n)		evidence (for) evolution
adaptation (n)		adaptation (and) evolution
model (n)		evolution models
human (j)	behavior (n)	human behavior
human (n)		behavior (of) humans
social (j)		social behavior
cultural (j)		cultural behavior
evolution (n)		evolution (of / of the) behavior
pattern (n)		pattern (of) behavior
adaptation (n)		behavior (and) adaptation
model (n)		model(s) (of) behavior

* j: adjective; n: noun.

4. Dimensions of collocation in academic writing

The four dimensions of collocation are discussed below, each comprising sets of interrelated nodes and collocates:

1. Human Nature, Culture and Research Methods vs. Economics
2. Human Evolution and Society
3. Business and Finance
4. Statistics

The positive pole of Dimension 1 (Human Nature, Culture and Research Methods) has 1,405 nodes. A highly significant statistical difference exists for the mean dimension scores ($F = 1680.06$; $p = .000$), with articles having a positive score (161.09), and textbooks a negative score (−93.03), which suggests that the collocations associated with the positive pole are typical of journal articles, whereas those associated with the negative pole are common in the textbooks. Table 4 shows the nodes with the highest loadings in this pole, divided into nouns, adjectives, and verbs, as well as the most frequent collocations for each node.

As reflected in the interpretive label for the dimension, the major topics for the positive pole are: (1) Human nature and Culture; and (2) Research Methods. Topics

Table 4. Nodes with the highest loadings in the positive pole of Dimension 1 and examples of collocation found in the corpus

Node category	Nodes with highest loadings	Collocation examples
Nouns	literature (.78), culture (.70), language (.70), domain (.70), human (.69), intervention (.68), motivation (.67), actor (.67), behavior (.65), brain (.64), childhood (.63), evaluation (.62), paradigm (.62), group (.62), people (.62)	*literature review; common culture; human population; government intervention; human behavior; childhood disadvantage; processual paradigm; ethnic group; ask people*
Adjectives	human (.72), cultural (.71), neural (.69), biological (.69), clinical (.66), prior (.65), social (.64), causal (.64), developmental (.63), genetic (.63), perceptual (.61), different (.61), behavioral (.60), early (.60), contemporary (.60)	*human tendency; biological basis; clinical psychology; social life; developmental disorders; genetic variation; behavioral adaptation; early input*
Verbs	highlight (.65), review (.65), study (.62), learn (.61), understand (.60), encode (.60), overlap (.59), suggest (.58), shape (.58), identify (.57), see (.57), think (.57), know (.57), recognize (.57), live (.56)	*highlight importance; review evidence; learned behaviors; evidence suggest; shape behavior; think ~ act; process known; people recognize; live alone*

related to Human nature are reflected by collocations related to human development, cognition, and literacy. Collocations that address cultural issues relate to cultural practices, both unique to a given social/regional ground and common by different communities; shared beliefs and customs; community identity; and, particularly, human beings as a social entity, especially concerning active participation in different aspects of societal life, social attributes, and societal issues such as gender and race. The following are text excerpts with collocations that exemplify these concepts (Examples (1) to (5)):

(1) *It is impossible to separate **human nature** from **human culture**. As popular as it may be to think that nature has driven our **development** as **humans**, even our long evolutionary process has been deeply **influenced** by **culture**.*

(5_7_1.txt – BCS)

(2) *IQ tests, in fact, **measure** particular **knowledge** and **abilities** that are largely learned through one's culture. They are valuable. Relatively low scores may sometimes point out a **learning disability** that has a **biological basis**. Because they measure the kinds of skills **required** by **education** in our culture as well as by many occupations, they do have predictive value as to one's **success** within the **culture**.*

(5_7_3.txt – BCS)

(3) *specialization theory is more consistent with white families and may not be an appropriate perspective for understanding the wage impact of marriage for black women. Failure to account for the intersection of* ***gender*** *and* ***race*** *in the mechanisms leading to the wage effect of marriage will oversimplify the complex, heterogeneous* ***nature*** *of* ***marriage****.* (5_6_183.txt – SES)

(4) *These different depictions of men and women reflect* ***gender roles, societal beliefs*** *about how men and women are* ***expected*** *to* ***behave****. For example, in many cultures, women are expected to* ***assume*** *the* ***role*** *of wife and mother and have limited opportunities to pursue other careers.* (5_7_14.txt – BCS)

(5) *Despite Abdullah and Brown's (2011) review, the relationship between* ***culture*** *and* ***stigma*** *is complex (Yang et al., 2007). In the Abdullah and Brown (2011) review, for example, disparate* ***cultural groups*** *were reduced to continent-level constructs (Asian or African). As Abdullah and Brown noted, Indians and Chinese in Asia may be as different from each other as Western Europeans are from East Asians.* (5_6_64.txt – BCS)

The collocations related to the last main topic, research methods, capture issues that are typically found in the research methods sections of the journal articles. The use of collocations in this case is not motivated by topical coherence, but rather by academic writing conventions regardless of area or subarea. The following excerpts (Examples (6) and (7)) show the most frequent collocations in this topic.

(6) *Our extensive* ***literature review*** *shows that researchers have only rarely examined these types of outcomes. For example, very little* ***evidence exists*** *regarding the relationship between age and the basis of reemployment, and no meta-analyzable correlations were available.* (5_6_19.txt – BCS)

(7) *From a* ***methodological perspective****, this finding* ***highlights*** *the* ***importance*** *of examining a disorder-specific bias at the level of the individual.* (5_6_23.txt – BCS)

Unlike the positive pole, the number of nodes loading on the negative pole is restricted, with only 26 words, all centering around Economics. As mentioned, this pole is statistically associated with the textbooks in the corpus. It is important to highlight that the topics that came up in this pole are all related to Social and Economic Sciences, even though both subareas were equally represented in the textbook subcorpus. Table 5 shows the nodes that loaded in this pole, broken down by part of speech, as well as the most frequent collocations.

Table 5. Nodes loading in the negative pole of Dimension 1 and examples of collocation found in the corpus

Node category	Nodes (loadings)	Collocation examples
Nouns	saving (–.41), currency (–.39), corporation (–.37), liability (–.37), budget (–.36), shareholder (–.36), stockholder (–.36), seller (–.34), expense (–.34), financing (–.34), export (–.34), owner (–.33), transaction (–.32), bill (–.32), depreciation (–.32)	*national saving; foreign currency; large corporation; tax liability; budget deficit; shareholder value; common stockholders; buyers ~ sellers; operating expenses; debt financing; net exports; business owners; market transaction; tax bill; depreciation rate*
Adjectives	fiscal (–.33), extra (–.32), nominal (–.31)	*fiscal policy; extra revenue; nominal rate*
Verbs	finance (–.38), purchase (–.38), issue (–.36), borrow (–.35), supply (–.35), charge (–.32), cost (–.30)	*finance investment; purchase bond; issue currency; borrowing constraints; quantity supplied; charge price; cost more*

All collocations identified on this pole reflect such themes as economic theory, economic policy, and fiscal policy. These are illustrated in the text excerpts numbered (8) through (11).

(8) *One such comparison involves government-issued securities. These are free of much of the variability we see in, for example, the stock market. The government* ***borrows money*** *by* ***issuing bonds*** *in different forms.* (5_7_17.txt – SES)

(9) *In contrast to the assumption of the neoclassical model, firms cannot always raise funds to* ***finance investment. Financing constraints*** *make investment sensitive to firms' current cash flow.* (5_7_18.txt – SES)

(10) *When receipts exceed spending, the government is said to run a* ***budget surplus****. The government finances a* ***budget deficit*** *by borrowing from the public.* (5_7_19.txt – SES)

(11) *Because NX is the distance between the* ***saving schedule*** *and the investment schedule at the world interest rate, this shift reduces NX. Hence, starting from balanced trade, a change in* ***fiscal policy*** *that reduces* ***national saving*** *leads to a trade deficit.* (5_7_18.txt – SES)

Dimension 2, Human Evolution and Society, had a single pole, with 417 nodes. Collocations found in this dimension had the highest loadings in textbooks. The difference between registers is statistically significant ($F = 1720.65$; $p = .000$), with a positive mean dimension score for textbooks (22.52) and a negative mean score for articles (–38.85), which suggests that these collocations are usually found in

textbooks rather than articles. Table 6 presents the nodes with the highest loadings and typical collocations.

Table 6. Nodes with the highest loadings in Dimension 2 and examples of collocation found in the corpus

Node category	Nodes with highest loadings	Collocation examples
Nouns	specie (.78), primate (.76), fossil (.74), species (.73), hominins (.69), ape (.69), ancestor (.68), anthropologist (.68), bone (.68), chimpanzee (.65), monkey (.64), archaeologist (.63), tree (.62), village (.61), hominin (.61)	*animal fossil; separate species; ape behavior; cultural anthropologist; bone tools; chimpanzee population; family tree; village life;*
Adjectives	ancient (.67), African (.67), fossil (.65), archaeological (.64), evolutionary (.63), European (.57), sapiens (.56), native (.55), ancestral (.54), animal (.52), reproductive (.51), religious (.49), southern (.49), Neandertal (.49), archaic (.48)	*ancient remains; African populations; evolutionary change; European descent; Native Americans; ancestral land; religious beliefs; southern Africa; archaic states*
Verbs	discover (.62), date (.61), eat (.58), survive (.57), remember (.56), forage (.53), die (.51), preserve (.50), possess (.49), reconstruct (.49), recover (.48), spread (.48), walk (.48), gather (.48), belong (.48)	*anthropologist discover; date fossils; remember discussion; preserve ~ protect; individuals possess; remains recovered; walk upright; gather data*

The collocations on this dimension reflect such themes as the environment, anthropology, human evolution, culture, and social issues. Excerpts (12) to (14) illustrate collocations from this dimension as they appear in the texts.

(12) *Because it is a characteristic of all vertebrates including fish, reptiles, birds, and mammals, bilateral symmetry does not contribute to the reconstruction of* ***evolutionary relationships*** *among* ***fossil primates****. Instead, paleoanthropologists pay particular attention to recently evolved derived features in order to construct* ***evolutionary relationships*** *among* ***fossil groups****.* (5_7_13.txt – BCS)

(13) *A* ***state religion*** *was usually practiced, even in areas of linguistic and ethnic diversity. The Classic Maya, Aztec, Inka, and* ***ancient*** *Egyptian* ***societies*** *are examples of* ***archaic states****.* (5_7_10.txt – BCS)

(14) *There are some significant contrasts between industrial and nonindustrial economies. When* ***factory workers*** *produce for sale and for their employer's profit, rather than for their own use, they may be alienated from the items they make.* (5_7_6.txt – BCS)

Dimension 3 (Business and Finance) has a total of 237 nodes, and examples of their collocations appear in Table 7. As with the other dimensions, there is a statistically significant difference between registers ($F = 926.47$; $p = .000$), with articles having a positive mean dimension score (25.54), and textbooks, a negative mean dimension score (−14.75), which suggests that these collocations appear frequently in articles but not in textbooks.

Table 7. Nodes with the highest loadings in Dimension 3 and examples of collocation found in the corpus

Node category	Nodes with highest loadings	Collocation examples
Nouns	dollar (.69), sale (.66), revenue (.65), debt (.65), bank (.64), asset (.64), tax (.64), payment (.63), investment (.63), price (.63), cash (.63), profit (.62), purchase (.62), stock (.61), fund (.60)	*dollar cost; revenue ~ sale; bank account; asset price; interest payments; investment bank; higher price; cash flow; government purchases; mutual funds*
Adjectives	net (.67), total (.63), annual (.57), domestic (.55), marginal (.54), foreign (.54), average (.52), private (.51), equal (.50), financial (.49), aggregate (.47), low (.46), risky (.44), retail (.42), additional (.42)	*net worth; annual income; marginal cost; foreign competition; financial crisis; aggregate demand; additional revenue; retail stores*
Verbs	buy (.65), sell (.65), pay (.64), earn (.59), rise (.59), exceed (.58), invest (.56), lower (.55), decline (.50), raise (.48), fall (.48), fix (.48), trade (.47), save (.47)	*buy ~ sell; sell products; pay dividend; price exceeds; invest capital; lower price raise funds; stock traded*

Collocations that comprise this dimension generally refer to themes such as investment banking, business operations, business management, and the stock market. More specifically, collocations are related to making investments and the operations of running a business or company, unlike the negative pole of Dimension 1, which focuses on economic theory and how the economy is shaped by factors such as government policy. Examples appear in text excerpts (15) and (16).

(15) *[T]he POC also provides a hedge to entrepreneurs who prefer greater* ***net worth*** *when the return to* ***internal funds*** *is high.* (5_6_141.txt – SES)

(16) *They argue that the market is dominated by large* ***bond traders*** *who* ***buy and sell securities*** *of different maturities each day, that these traders focus only on* ***short-term returns****, and that they are not concerned with* ***maturity risk****. According to this view, a* ***bond trader*** *is just as willing to* ***buy*** *a 20-year* ***bond*** *to pick up a* ***short-term profit*** *as he or she is to* ***buy*** *a 3-month* ***security****.* (5_7_21.txt – SES)

Finally, Dimension 4 (Statistics), which includes 25 nodes (examples in Table 8), also distinguishes between registers ($F = 339.78$; $p = .000$), albeit less sharply than with the previous dimensions, with articles having a positive dimension score (1.76), and textbooks, a negative score (−1.02), again indicating that these collocations are slightly more typical of articles than textbooks.

Table 8. Nodes loading in Dimension 4 and examples of collocation found in the corpus

Node category	Nodes (loadings)	Collocation examples
Nouns	variance (.44), coefficient (.42), regression (.41), parameter (.41), estimator (.40), estimate (.39), error (.39), equation (.39), deviation (.38), slope (.38), intercept (.37), residual (.36), correlation (.36), estimation (.35), interval (.33), elasticity (.32), panel (.32), column (.31)	*error variance; correlation coefficient; multiple regression; population parameter; measurement error; steep slope; estimate intercept; residual variance; effects estimation; confidence interval; panel data; shown ~ column*
Adjectives	linear (.37), standard (.36), estimated (.35), explanatory (.33), conditional (.32)	*linear model; standard deviation; estimated effect; explanatory variable; conditional assumption*
Verbs	compute (.40), estimate (.34)	*compute average; estimate model*

As with the positive pole of Dimension 1, the node and collocate pairs loading in Dimension 4 capture register-specific features, reflecting collocations used to talk about empirical methods and statistical terminology. Examples of collocations in this dimension appear in excerpts (17), (18), and (19).

(17) *Another method is to correct the conventional* ***standard errors*** *using what is known as the "sandwich* ***variance estimator****" attributed to Huber (1967) and White (1982), which is widely available in software for* ***estimating*** *multilevel* ***models****, including SAS PROC MIXED (see Sterba, 2009).* (5_6_55.txt – BCS)

(18) *The correlation across respondents in the rating of each industry is high, with the* ***correlation coefficient*** *between the responses of any two reviewers varying between 0.31 and 0.83.* (5_6_130.txt – SES)

(19) *This measure is designed to* ***capture deviations*** *from the long-run sustainable size in the market. This is a market-specific measure; market tightness in a given market is measured relative to that market's maximum size. The investment policy functions are* ***estimated*** *using* ***linear regression****.* (5_6_131.txt – SES)

5. Conclusion

This study applied a novel and powerful methodology for identifying sets of collocations through a lexical version of the MD analysis framework (Berber Sardinha 2017). The analysis identified the interrelationships among more than 2,000 different nodes and 45,000 different collocates. The results showed four distinct dimensions of collocation in English academic writing, each reflecting the major collocations found in articles and textbooks in Social, Behavioral and Economic Sciences. Unlike most previous studies on collocation, which tend to focus on individual collocations selected ahead of time, in this study all of the collocations in the corpus were extracted. Through the application of collocational MD analysis, we were able to see how the collocations clump together around common topics, themes, issues, and register rhetoric. Following on from Sinclair, Jones and Dailey (1970/2014), we demonstrated that collocations are sensitive to register differences, in that all of the dimensions distinguished between articles and textbooks. At the same time, a limitation of the current study is that the collocation dimensions are sensitive to topic and subject matter differences (Berber Sardinha, Mayer Acunzo & São Bento Ferreira 2016; Berber Sardinha 2017), and therefore slight differences in the content of the articles and textbooks in our corpus may have contributed to the statistical differences associated with register found in this study. Although great care was taken in designing the corpus to minimize the influence of topic and subject matter, even slight differences could result in one register having different collocations from another. More research in collocation-based MD analysis is needed before we can determine the extent to which topic and subject matter influence the register distinctions picked up collocation.

The collocational dimensions can be of relevance to EAP, not only because they comprise the most typical collocations found in particular fields, but also because each dimension provides a coherent set of collocations that can potentially be taught as topical units. For teachers and students, it is always challenging to determine which collocations to learn, given the sheer number of them in use, and therefore it would make sense to give priority to collocations that are used to talk about similar issues in particular registers, which is what the dimensions of collocation can offer. The dimensions can be seen as an alternative to teaching collocations from arbitrary lists, as the nodes are selected statistically from all of the words in the corpus. The dimensions comprise over 2,000 nodes, each pairing up with different collocates to form thousands of collocations, which provide a comprehensive picture of the formulaic repertoire of research articles and university textbooks. Because awareness of collocation and proficiency in using collocation both play a pivotal role in academic writing, it is not hard to imagine a place in the

EAP classroom for the collocations found through collocational MD analysis. We hope that more studies will consider taking this approach to the investigation of collocation in corpora.

Acknowledgements

The research presented here was supported by CAPES (Brazil), the Fulbright Commission (Doctoral Dissertation Research Award – DDRA 2018/2019), and CNPq (Brasília, Brazil, grant #306994/2017–8).

References

Ackermann, Kirsten & Chen, Yu-Hua H. 2013. Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes* 12(4): 235–247. https://doi.org/10.1016/j.jeap.2013.08.002

Barnbrook, Geoff, Mason, Oliver & Krishnamurthy, Ramesh. 2013. *Collocation: Applications and Implications*. Houndmills: Palgrave Macmillan. https://doi.org/10.1057/9781137297242

Berber Sardinha, Tony. 2017. Lexical priming and register variation. In *Lexical Priming: Applications and Advances* [Studies in Corpus Linguistics 79], Michael Pace-Sigge & Katie J. Patterson (eds), 189–229. Amsterdam: John Benjamins. https://doi.org/10.1075/SCL.79.08BER

Berber Sardinha, Tony, Mayer, Cristina & Ferreira, Telma. In press. Dimensions of collocation in Brazilian Portuguese: Exploring the Brazilian Corpus on SketchEngine. In *Essays in Lexical Semantics in Honor of Adam Kilgarriff*, Mona Diab & Aline Villavicencio (eds). Berlin: Springer.

Berber Sardinha, Tony, Ferreira, Telma & Mayer, Cristina. In press. *A Dictionary of Portuguese Collocations*. London: Routledge.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511621024

Biber, Douglas. 1993. Co-occurrence patterns among collocations: A tool for corpus-based lexical knowledge acquisition. *Computational Linguistics* 19: 549–556.

Biber, Douglas. 2006. *University Language: A Corpus-Based Study of Spoken and Written Registers* [Studies in Corpus Linguistics 23]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.23

Biber, Douglas. 2009. Multi-dimensional approaches. In *Corpus Linguistics: An International Handbook*, Vol. 2, Anke Lüdeling & Merja Kytö (eds), 822–855. Berlin: Walter de Gruyter.

Biber, Douglas. 2012. Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory* 8(1): 9–37. https://doi.org/10.1515/cllt-2012-0002

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Brezina, Vaclav, McEnery, Tony & Wattam, Stephen. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2): 139–173. https://doi.org/10.1075/ijcl.20.2.01bre

Brinton, Laurel & Brinton, Donna. 2010. *The Linguistic Structure of Modern English*, rev. edn. Amsterdam: John Benjamins. https://doi.org/10.1075/z.156

Crossley, Scott, Salsbury, Tom & McNamara, Danielle. 2015. Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics* 36(5): 570–590.

Durrant, Philip. 2009. Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes* 28(3): 157–169. https://doi.org/10.1016/j.esp.2009.02.002

Friginal, Eric & Hardy, Jack A. 2014. Conducting multi-dimensional analysis using SPSS. In *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber* [Studies in Corpus Linguistics 60], Tony Berber Sardinha & Márcia Veirano Pinto (eds), 298–316. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.60.10fri

Gablasova, Dana, Brezina, Vaclav & McEnery, Tony. 2017. Collocations in corpus-based language learning research: Identifying, comparing and interpreting the evidence. *Language Learning* 67(1): 155–179. https://doi.org/10.1111/lang.12225

Gardner, Dee & Davies, Mark. 2014. A new academic vocabulary list. *Applied Linguistics* 35: 305–327. https://doi.org/10.1093/applin/amt015

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: CUP. https://doi.org/10.1017/CBO9781139524773

Lei, Lei & Liu, Dilin. 2018. The academic English collocation list: A corpus-driven study. *International Journal of Corpus Linguistics* 23(2): 216–243.

Palmer, Harold E. 1933. *The Second Interim Report on English Collocations*. Tokyo: Kaitakusha.

Phillips, Martin. 1989. *Lexical Structure of Text* [Discourse Analysis Monograph 12]. Birmingham: University of Birmingham.

Rychly, Pavel. 2008. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing*, RASLAN 2008, Petr Sojka & Aleš. Horák (eds), 6–9. Brno: Masaryk University.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Sinclair, John M. 1998. The lexical item. In *Contrastive Lexical Semantics* [Current Issues in Linguistic Theory 171], Edda Weigand (ed.), 1–24. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.171.02sin

Sinclair, John, Jones, Susan & Daley, Robert. 1970/2004. *English Collocation Studies: The OSTI Report*. London: Continuum.

Stubbs, Michael. 2007. On texts, corpora and models of language. In *Text, Discourse and Corpora*, Michael Hoey, Michaela Mahlberg, Michael Stubbs & Wolfgang Teubert (eds), 127–161. London: Continuum.

West, Michael. 1953. *A General Service List of English Words*. London: Longman.

Williams, Geoffrey. 1998. Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics* 3(1): 151–171. https://doi.org/10.1075/ijcl.3.1.07wil