High-frequency collocations of nouns in research articles across eight disciplines

Matthew Peacock
City University of Hong Kong (China)
enmatt@cityu.edu.hk

Abstract

This paper describes a corpus-based analysis of the distribution of the high-frequency collocates of abstract nouns in 320 research articles across eight disciplines: Chemistry, Computer Science, Materials Science, Neuroscience, Economics, Language and Linguistics, Management, and Psychology. Disciplinary variation was also examined – very little previous research seems to have investigated this. The corpus was analysed using WordSmith Tools. The 16 highest-frequency nouns across all eight disciplines were identified, followed by the highest-frequency collocates for each noun. Five disciplines showed over 50% variance from the overall results. Conclusions are that the differing patterns revealed are disciplinary norms and represent standard terminology within the disciplines arising from the topics discussed, research methods, and content of discussions. It is also concluded that the collocations are an important part of the meanings and functions of the nouns, and that this evidence of sharp discipline differences underlines the importance of discipline-specific collocation research.

Keywords: collocations, corpus analysis, interdisciplinary research writing, genre analysis.

Resumen

Colocaciones muy frecuentes de sustantivos en artículos de investigación en ocho disciplinas académicas

En este trabajo se analiza la distribución de las colocaciones más frecuentes de sustantivos abstractos en un corpus de 320 artículos de investigación en ocho disciplinas diferentes: Química, Informática, Ciencias de los Materiales, Neurociencia, Economía, Lengua y Lingüística, Administración de Empresas y Psicología. Se examina también la variación según las diferentes disciplinas, un

aspecto poco tratado en investigaciones anteriores. En el análisis del corpus se utilizó WordSmith Tools. Se identificaron los 16 sustantivos más frecuentes en cada una de las ocho disciplinas así como las colocaciones más frecuentes en cada caso. En cinco disciplinas se dio hasta un 50% de variación respecto a los resultados combinados de todas las disciplinas examinadas. Se concluye que las diferencias identificadas guardan relación con las características de cada disciplina y que éstas también están relacionadas con la terminología estándar dentro de cada disciplina en relación con los temas tratados, los métodos de investigación empleados y el contenido de los artículos. Se concluye además que las colocaciones constituyen una parte importante de los significados y de las funciones de los sustantivos, y que la existencia de marcadas diferencias disciplinarias subraya la necesidad de investigar las colocaciones en las distintas disciplinas académicas.

Palabras clave: colocaciones, análisis de corpus, artículos de investigación en diferentes disciplinas, análisis de género.

Introduction

This paper describes a corpus-based analysis of the high-frequency collocations of common nouns in 320 research articles (RAs) across eight disciplines. The following definition of collocation is adopted, "chunks of language-sequences of words-that are used repeatedly by speakers and writers" (Biber, Conrad & Cortes, 2004: 377). Collocation is the co-occurrence of two or more words significantly more often than is found in similar language genres. For the present research high-frequency is defined as a frequency of at least 40 per million words (pmw), following the definition of Biber, Conrad and Cortes (2004: 376): "we take a conservative approach (…) [a] frequency cut-off of 40 times per million words to be included in the analysis".

Williams (2002) asserts that discourse communities develop codes for communication through the use of patterns and that this code, rather than individual words, is one of their defining characteristics. Collocations appear to be among these patterns. However, description and discussion of the distribution of the high-frequency nouns themselves is beyond the scope of this research, which focuses on their collocations.

The RA was chosen for this research because of its significance for the spread of knowledge. RAs have been called the key medium for legitimating findings and disciplines (Hyland, 1996), and the preferred genre for

discourse communities to communicate (Williams, 1998). Their language defines these communities.

One of the earliest mentions of the word collocation is Firth (1957), who also wrote the well-known related phrase "You shall know a word by the company it keeps" (Firth, 1968: 179). Sinclair (1991) broke new ground with his suggestions that word combinations are not random and that they make an important contribution to the organization of language, while Hoey (1991) contends that collocation plays an important role in lexical cohesion. In a later work (Hoey, 2007a), he argues that exposure to collocations primes or prepares us to recall their correct meaning, and use them correctly, whenever we re-encounter them. He extends this idea (2007b) to the construction of grammars, and examines some evidence for the latter through an analysis of the collocates of "sixty", "60", "forty", and "40" in a corpus of *The Guardian* newspaper text. Hoey (2007b) concludes that his analysis of these lexical units provides some evidence for such priming and for the unexpected decisions made by writers (also see Hoey & O'Donnell, 2008).

Collocation in academic writing has also attracted interest recently. Collocations have also been called formulaic sequences, "chunks (…) multiword units (…) conventionalised forms, ready-made utterances" (Wray, 2002: 9), naturally co-occurring strings of words (Chan & Liou, 2005), and word partnerships (Mudraya, 2006). Many writers stress their importance: they have been called an essential organizing principle of language in use (Stubbs, 1995; Schmitt & Carter, 2004). Stubbs (1995), Mahlberg (2003) and Gledhill (2000a) emphasize that meaning develops across word clusters and not through single words, and Herbst (1996) that there is no doubt that language competence includes knowledge of collocations. They also let users express membership of a group, articulate ideas economically and reduce processing effort for readers (Jones & Haywood, 2004; Gledhill, 2000b). Gledhill points out (2000a) that collocations are fundamental units in texts, that they validate the existence of discourse communities, and that they are subconscious efforts to conform to discipline norms. Finally, they may be more quickly recognized than individual words (Cantos & Sanchez, 2001) and reduce processing effort for readers (Jones & Haywood, 2004). Schmitt and Carter (2004) note that there is a lot of evidence that collocations are stored and processed as unitary wholes, and Schmitt, Grandage and Adolphs (2004: 127) that writers use the same clusters repeatedly because they are "prepackaged in the memory".

Very little previous research seems to have investigated the high-frequency collocations of common nouns. Ward (2007) looks at common nouns and their collocations in Chemical Engineering textbooks, and compares the collocation frequency with that found in four other engineering disciplines. The three most common nouns were "gas", "liquid", and "heat". Ward asserts that while collocations are certainly discipline specific, this is not true of individual words. He observes that the important phrase is not "gas" but "gas +" and that collocations are a threshold to discipline membership. However, he does not give a list of common nouns apart from these three, report common nouns in the other disciplines, or report collocations (apart from a large number of "gas +" collocations). However, Ward's (2007) exploratory study is valuable as it pays attention to collocations within a corpus. Gledhill (2000a) researches salient words which he defines as words that occur significantly more often in one text or part of a text than another, though the research is not confined to nouns. He reports collocations in his Pharmaceutical Corpus of 150 RAs from 22 cancer and pharmacology journals, though not focusing on nouns. Some example collocations were "patients who had tumours" and "both accelerate and delay".

Rationale for research

The high-frequency collocations of common nouns may be an important part of academic English including RAs, and worth investigating further. There have been several calls for research into collocation, for example Groom (2005) suggests that disciplines can be differentiated by their favoured terminology and that this notion is well worth examining on a larger scale. Gledhill (2000a) says that looking at different disciplines is an intriguing possibility. Ellis, Simpson-Vlach and Maynard (2008) emphasize the importance of collocations, suggesting that they are common in academic discourse and that writers need to know them as a whole. Durrant (2009: 158) points out that the possible existence of sharp discipline differences in collocations imply that useful lists cannot be obtained by looking at any one discipline – "it is clearly misguided to seek any generic listing of academic collocations", adding that previous research has not attempted to describe disciplinary differences, and that it is important to undertake such research. Similar assertions regarding the importance of researching discipline differences (though in noun distribution) are made by Martinez, Beck, and Panza (2009) and Ward (2009).

If collocations are important, they must be acquired by aspiring research writers. Bhatia (2000) notes that a strong justification for genre research is that it informs the teaching of research writing, especially for writers who wish to join academic discourse communities, while Durrant (2009) suggests that learners need to acquire high-frequency collocations. There also appears to be some agreement that non-native speakers (NNS) find collocations difficult and/or misuse them. Wray (2002) asserts that collocations are hard for NNS, that NNS tend to use the right words in the wrong context, and are too creative with collocations. She also claims that NNS make overliberal assumptions about the use of collocations and that they are at a disadvantage with them, and predicts that NNS could end up with larger lexicons than native speakers (NS) but not know how to use collocation. Shei and Pain (2000) claim that it is commonly agreed that NS and NNS differ in their knowledge of collocations and that NS use them more, use a greater variety, and use them more accurately: arguments also put forward by Ellis and Simpson-Vlach (2009) and Ellis, Simpson-Vlach and Maynard (2008). Schmitt and Carter (2004), and Bahns and Eldaw (1993) agree that NNS misuse them. Other authors mention NNS research writing in more general terms. Paltridge (1993) contends that NNS need help in joining the discourse community of international academic research, and Yakhontova (1997) that NNS research writers tend to be unaware of genre conventions, which differ in the second language. Ahmad (1997) indicates that this is critical for NNS, who may not get published when their work is written in an incorrect rhetorical style. These difficulties might result from first language differences, which are very hard to overcome (Vassileva, 1997; Golebiowski, 1999). Wood (2001) adds that NNS writers of RAs have higher-level discourse problems and difficulties entering discourse communities and publishing.

There seems to have been very little research into the high-frequency collocations of nouns in RAs, or disciplinary variation in their use, and the area seems to be increasingly important due to the fast-growing numbers of research writers around the world, particularly NNS. Ward points out (2007) a problem affecting this area – language teachers lack knowledge of technical vocabulary and so cannot be expected to teach it. This present research can provide information that can be used to support the teaching of research writer competence across a number of disciplines. It is certainly possible that the correct selection of collocations is a vital part of the acquisition of competence in the skills of constructing scientific discourse; and if so, it will

be useful to provide a description of these collocations. It is proposed that the area has not received the attention it warrants and that further research is needed, to assess disciplinary variation across a number of disciplines. The results should reveal much more about the nature of RAs, and help teachers of research writing inform learners of appropriate collocations.

Methodology

This research investigated the distribution and frequency of the high-frequency collocations of nouns in 320 research articles across eight disciplines.

Research Aims

The aims of this research were, within the corpus, to:

(1) find and list the highest-frequency collocations of common nouns;

(2) investigate the frequency of these collocations; and

(3) investigate disciplinary variation.

The RA Corpus

The corpus was 320 published RAs, 40 from each discipline. The eight disciplines were selected because they represent a range of subjects and also have large numbers of research writers, mostly NNS, around the world. This increases the usefulness of this research regarding recommendations for teaching. Four leading refereed journals were selected from each discipline. Visits were made to the relevant departments and two sources from each were asked to name principal journals from their field.

Ten RAs from 2007/2008 were randomly chosen from each journal by giving each a number and drawing numbers from a box. Only empirical data-driven RAs with the Introduction-Method-Results-Discussion (IMRD) format were chosen, as this is an important genre (Hyland, 1998). The size of the disciplinary corpora, and the use of discipline sources to choose journals, suggest that the corpora are sufficiently representative.

Investigating the Corpus

Analysis was done in the following steps:

(1) High-frequency nouns were identified using the WordList function of WordSmith Tools 4.0 (Scott, 2004). Many of these nouns (see Table 1), for example "study", "process", and "variable", sometimes function as verbs or adjectives. Every occurrence of these functions was excluded from the count: to do this it was necessary to manually examine each occurrence using the Concord function. At this stage the research was limited to the 16 highest-frequency nouns (excluding usage as verbs or adjectives), to make the research more manageable.

(2) High-frequency collocations were identified, along with disciplinary variation, using the Concord function of WordSmith Tools plus the Clusters, Patterns, and Collocates sub-functions.

Regarding step 1, firstly, "function" means "operates" or "acts". Secondly, WordSmith Tools uses a measure of association called "mutual information" (MI) to define collocates, or more accurately to assess whether "co-occurrences" happen by chance or are statistically "significant". MI measures the strength of each collocation, eliminating those that appear by chance, and is thus a necessary statistical test of strength of association. Regarding step 2, the corpus was split into disciplinary corpora at times to check disciplinary variation. Individual manual checking of the function of every occurrence is vital. Many authors stress the importance of doing this, for example frequency can be obtained from statistical analysis but context is vital in understanding function (Tognini-Bonelli, 2004), and a "microscopic study" must be carried out before categorisation can be done (Williams, 2002: 60).

Two evaluators were involved in step 1: this writer and a local university lecturer. To measure inter-rater agreement, the second coder independently evaluated the function of every occurrence. To measure intra-rater agreement, this writer reassessed the function of every occurrence after one month. Inter-rater and intra-rater agreement were both 100%.

Results

The 16 highest-frequency nouns in the whole corpus, in order of frequency, can be seen below in Table 1. They are all abstract nouns. The highest-frequency collocates for each noun are also shown.

Noun	Collocations	Percent of all occurrences
study/ies	present ~, previous ~, case ~, results (of) ~	19
result/s	~ show/ed, ~ indicate/d, ~ suggest/ed, ~ obtained	14
effect/s	significant ~, main ~, no ~, positive ~	24
model/s	-	0
information	~ management, ~ system/s, ~ technology, ~ processing	12
data	~ (were) collected, ~ collection, ~ analysis, ~ were obtained	10
analysis/es	factor ~, regression ~, ~ was/were performed, ~ revealed	11
process/es	business ~, learning ~, information ~, planning ~	9
research	previous ~, future ~, further ~, ~ has shown	17
sample/s	~ period, ~ size	6
experiment/s	results (of/in) ~ (1, 2, 3), present ~, participated in ~, previous ~	8
relationship/s	~ between, customer ~, positive ~, causal ~	38
factor/s	~ analysis/es, (1st-, 2nd-, 3rd-, higher-) order ~, ~ structure, key ~	14
variable/s	dependent ~, dummy ~, independent ~, explanatory ~	29
method/s	-	0
evidence	provide/d/ing ~, find/found ~, empirical ~, there is/was no ~	27

Table 1. High-frequency collocations, in order of frequency – All disciplines.

Table 1 does not show all the high-frequency collocates, only the four most common. Examples follow: "analysis" collocated with "factor ~", "regression ~", "~ was/were performed", and "~ revealed". "Evidence" collocated with "provide/d/ing ~", "find/found ~", "empirical ~", and "there is/was no ~". "Process" collocated with "business ~", "learning ~", "information ~", and "planning ~". The right column shows the percentage of all occurrences of the noun which these particular collocations make up – for example, the four collocates of "study/ies" make up 19% of all occurrences of the noun: the average over all nouns was 16%. The percentage varied by noun – the five nouns that correlated most often with the most common collocates were "relationship/s" (38% of occurrences), "variable" (29%), "evidence" (27%), "effect" (24%), and "study/ies" (19%). However, "model" and "method" had no high-frequency collocates, and "sample" only two. A large number of disciplinary differences were found. These are shown in Tables 2 and 3.

Noun	Chemistry	Computer Science	Materials Sci.	Neuroscience
study/ies	present ~	previous ~, results (of) ~, present ~, case ~	previous ~, present ~	present ~, previous ~, current ~, recent ~
result/s	~ obtained	~ indicated, ~ show/n, experimental ~, ~ obtained	~ indicate, ~ show/n, similar ~, ~ suggest	~ suggest, ~ show/n/ed, ~ indicated, ~ (in/of) experiment (1,2,3)
effect/s	-	significant ~, positive ~, no ~	-	main ~, significant ~, no ~
model/s	-	user ~, ~ order, research ~	-	~ analysis/es, direct ~, memory ~
information	-	~ system/s, provides ~, ~ extraction, quality of ~	-	~ processing
data	~ collected, ~ collection, crystal ~	training ~, ~ collected, consistent ~	~ (…) shown, experimental ~, ~ presented, ~ obtained	regression ~, individual ~, ~ were obtained
analysis/es	elemental ~	data ~, factor ~ , ~ results, further ~	thermal ~, ~ was/ were performed, reaction ~	~ of data/data ~, statistical ~, ~ revealed, model ~
process/es	-	software ~, business ~	corrosion ~	cognitive ~
research	-	previous ~, ~ model, qualitative ~, future ~	-	future ~, previous ~
sample/s	-	data ~	~ is/as shown, ~ tested, laboratory ~, observed (in all) ~	-
experiment/s	-	~ conducted	~ (…) performed	participated in ~, condition/s (in/of) ~, previous ~, present ~
relationship/s	~ between	~ between, causal ~, ~ among	~ between	~ between
factor/s	-	~ analysis, key ~, contextual ~	-	-
variable/s	-	controlled ~, value/s (of the) ~, dependent ~, independent ~	-	independent ~
method/s	solved by direct ~	(…)-based ~, evaluation ~, clustering based ~, common ~	sterilisation ~	-
evidence	-	-	-	there is/was no ~, provide ~
% differing from Table 1	38	52	59	39

Table 2. High-Frequency Collocations: Science Disciplines.

Comparison of Tables 2 and 3 with Table 1 reveals considerable disciplinary variation. Over all eight disciplines, no fewer than 157 collocations differ from those in Table 1, or 53%. For example, Computer Science authors collocated "information" with "provides ~", "~ extraction", and "quality of ~". Neuroscience authors collocated "analysis/es" with "~ of data/data ~", "statistical ~", and "model ~", and "process/es" with "cognitive ~". Economics authors collocated "factor/s" with "~ model", "~ productivity", and "controlling ~", while Psychology authors collocated "process" with "inference ~" and "cognitive ~". Management authors collocated "model" with "business ~", "portfolio ~", "measurement ~", and "structural ~".

Noun	Economics	Language	Management	Psychology
study/ies	previous ~, empirical ~, several ~	present ~, previous ~, case ~	case ~, empirical ~, results (of) ~, previous ~	present ~, previous ~, results (of) ~, current ~
result/s	regression ~, ~ suggest, ~ reported, empirical ~	~ of this study, ~ showed, ~ of the/this analysis, ~ reported	~ (of this) study, ~ indicate/d, ~ show, ~ suggest	~ (of this) experiment (1,2,3), ~ shows/ed, ~ indicated, pattern of ~
effect/s	positive ~, significant ~	~ of (non)correction, positive ~, ~ for/of feedback, significant ~	positive ~, interaction ~, significant ~, negative ~	main ~, significant ~, ~ of target, revealed (significant/main) ~
model/s	regression ~, probit ~, structural ~, theoretical ~	CARS ~	business ~, portfolio ~, measurement ~, structural ~	~ fit, (1-,2-,3-,4-,5-) factor ~, parallel ~
information	~ available, obtain ~	-	~ systems, ~ management, ~ technology, ~ acquisition	location ~, ~ processing, ~ sources
data	~ available, ~ source	~ analysis, ~ collection	financial ~, ~ collection, ~ collected, ~ analysis	~ of/from experiment (1,2,3), ~ suggests, ~ revealed
analysis/es	unit ~, regression ~, empirical ~, comparative ~	genre ~, data ~, needs ~, discourse ~	data ~/~ of (the) data, empirical ~, factor ~, organizational ~	factor ~, regression ~, confirmatory ~, ~ reveal/ed
process/es	production ~	writing ~, learning ~, language ~	business ~, planning ~, information ~, management ~	inference ~, cognitive ~
research	future ~, previous ~	~ question/s, further ~, ~ project, second language ~	future ~, previous ~, further ~, prior ~	previous ~, future ~, present ~
sample/s	~ period, ~ firms, during the ~ period, ~ selection	representative ~	firms in the ~/~ firms, ~ size, ~ selection	(non-) clinical ~, present ~, ~ size, ~ consisted of
experiment/s	current ~, single ~, previous ~	controlled ~	-	results (of) ~, identical ~, present ~, data (of/from/in) ~
relationship/s	~ between, long-run ~, positive ~	~ between, significant ~	~ between, customer ~, ~ portfolio/s, business ~	~ between, specific ~, current ~
factor/s	~ model, ~ productivity, controlling ~	learner ~, other ~	~ analysis, success ~	~ structure, (1st-, 2nd-, higher-) order ~, ~ analysis, ~ loadings
variable/s	dummy ~, dependent ~, independent ~, control ~	independent ~, dependent ~	dependent ~, independent ~, control ~, explanatory ~	dependent ~, independent ~
method/s	-	research ~	-	-
evidence	provides ~, empirical ~, strong ~, ~ suggests	provide/d ~, anecdotal ~, further ~	empirical ~	provide/d/s ~, stronger ~, empirical ~, further ~
% differing from Table 1	64	59	46	57

Table 3. High-frequency collocations: Non-science disciplines.

The bottom row in Tables 2 and 3 shows the percentage of collocations in each column which differ from those in Table 1. The percentage varies by discipline, with five disciplines showing over 50% variance: Computer Science, Materials Science, Economics, Language and Linguistics, and Psychology.

Discussion and Conclusions

The collocations in Table 1 may look very familiar to readers, who might therefore assume that these are the collocates most frequently associated with these nouns. However, this is not the case, as examination of Tables 2 and 3 shows: a large number of disciplinary differences may be seen. Durrant (2009) warned that it is unwise to produce generic or non-discipline-specific lists of collocations, and that useful lists cannot be constructed by looking at only one or two disciplines. Table 1 is just such a standard list, and is presented in this paper to act as a contrast with Tables 2 and 3, whose sharp discipline differences certainly imply that useful collocations are discipline specific.

In order to try to understand the reasons for these discipline differences, a closer examination of the corpus was then made. Examples of many of the discipline differences follow. Definitions of certain terms will be given to aid understanding:

Chemistry:

- "crystal data": Summaries of the fundamental *crystal data* and experimental parameters for structure determination are given in Table 1.
- "elemental analysis/es" (determining what elements are present in a sample): The white crystals, analysed by *elemental analysis* were not consistent with the Zn(Net2)2 minimal formula.
- "solved by direct method/s" (a mathematical process for determining crystal structure): All structures were *solved by direct methods* using SHELXS-97.

Computer Science:

- "experimental result/s": This *experimental result* demonstrates that our index scheme has a significant improvement on storage requirement.
- "user model/s" (description of a user: related to user behaviour): A POMDP is further improved with the addition of a *user model* which indicates how a user's goal SU changes over time.
- "information extraction" (retrieving information): We therefore used *information extraction* techniques to automatically identify and extract phone numbers directly from the transcript.
- "quality of information": High experience respondents showed significant differences across conditions for both their ratings of the *quality of*

information provided by the computer, and their openness to influence from the computer.

- "training data" (instructions for users): In this approach, *training data* is first generated as a by-product of trainers' interactions in a virtual environment, and models of empathy are induced from the resulting datasets.
- "software process/es" (the organization and management of software development): There is a widely held belief that a better *software process* results in a better software product.
- "clustering based method" (statistical analysis method used in Computer Science): We can see that the ELP based method outperforms the *clustering based method* in terms of average accuracy under the same experiment setting.

Materials Science:

- "thermal analysis/es" (studying changes in materials as the temperature changes): Dynamical mechanical *thermal analysis* (DMTA) showed a more significant increase.
- "reaction analysis/es": The characterization and optimization of the polymer structure via *reaction analysis* are paramount.
- "corrosion process/es": The above observations permit a synthesis of the essential sequence of events that occur on the metal surface as the *corrosion process* develops.
- "sterilisation method/s": The use of this *sterilisation method* upon this type of polyurethane when in combination with this specific scaffold fabrication technique has not previously been reported.

Neuroscience:

- "memory model/s" (description or theory of how memory works): Moreover, *memory models* have been developed, which can describe the variance of the recency effect during immediate, delayed and continuous distracter free recall within a single-memory store.
- "regression data": However, in this instance, calculation of the measures using individual *regression data* was complicated.
- "statistical analysis/es": Confirmatory *statistical analysis* showed a significant interaction between condition and number of scene objects.
- "cognitive process/es": It is hoped that future research will be undertaken to assess the *cognitive processes* by which eye movements influence component processes in memory.

Economics:

- "regression result/s" (common method of data analysis): From the findings of the *regression results* from the three key variables…
- "probit model/s" (from probability theory and statistics): Given the ordinal nature of the dependent variable we specify an ordered *probit model*.
- "factor model/s" (mathematical model used for stock analysis): Recently, a strand of the finance literature incorporates regime-switching behavior in *factor models* of the term structure.
- "factor productivity" (ratio of output to the input of labour and capital): The large differences in total *factor productivity* (TFP) between countries of the world at the present time are suggestive of a substantial disequilibrium.

Language and Linguistics:

- "genre analysis/es": The linguistic approach of *genre analysis* is defined and understood to be the study of linguistic behavior in both academic and professional settings.
- "discourse analysis/es": The second stage of the research will involve a *discourse analysis* of audio and video recordings of GCAE meetings.

Management:

- "interaction effect/s" (statistical term meaning the effect of variables on each other): However, the subtle difference in groups' affectedness is not large enough; in this case the *interaction effect* between group and treatment is not significant.
- "information acquisition" (the collection of primary information from organizational stakeholders): The 24 independent statements regarding one's preferred manner of *information acquisition* were scored on a 7-point Likert scale.

Psychology:

- "result/s (of this) experiment": The *results of Experiment 3* replicated those reported by Ivanoff and Klein.
- "(1-,2-,3-,4-,5-) factor model/s": A CFA was performed using the two independent factors (one from each of the *four*- and *five-factor* models).
- "parallel model/s" (processing items simultaneously rather than serially): The fixed-capacity *parallel models* assume that multiple visual objects can be selected and spatially tracked in parallel.
- "location information" (the location of a stimulus to which a response is

made): The system responsible for shifting of attention during tracking also obtains *location information* of moving objects in parallel.

- "(non-) clinical sample/s": Future work utilizing structured interviews, *clinical samples*, and multi-method assessment tools are advisable. These studies have used small, *non-clinical samples*, and have methodological limitations.
- "factor structure" (statistical term related to factor analysis): Again, principle factor analysis with Promax (oblique) rotation was used to delineate the *factor structure*.
- "factor loadings" (statistical term related to factor analysis): The remaining items were those with the highest *factor loadings* based on the prior factor analysis.

Careful reading and analysis of the above examples from the eight different disciplines, and of Tables 2 and 3, lead to the proposal that many or most of the collocations presented are standard terminology within the discipline. Among the many examples of this discipline-specific terminology are "crystal data" and "solved by direct method/s" (Chemistry), "software process", "clustering based method", and "user model" (Computer Science), "thermal analysis", "reaction analysis", and "corrosion process" (Materials Science), "memory model" and "cognitive process" (Neuroscience), "probit model" and "factor productivity" (Economics), "genre analysis" and "discourse analysis" (Language and Linguistics), "information acquisition" (Management), and "parallel model" and "(non-) clinical sample" (Psychology).

Examination of the differing collocations expressed in the examples and in Tables 2 and 3 shows that they appear to arise from the topics discussed: or more explicitly, it is apparent that the collocations express differing terminology, different topics, different research methods, and differing content of discussions across the eight disciplines. This being the case, these collocations are clearly a very important part of the meanings, and therefore of the functions, of these nouns. It is also evident that these meanings and functions often differ by discipline, and that these meanings and functions are expressed by the collocations.

These collocations are more common than those seen in Table 1, leading to the suggestion that this evidence of sharp discipline differences underlines the importance of discipline-specific collocation research. Furthermore, the sharp discipline differences presented here indicate that the high-frequency

collocations of common nouns are part of the favoured terminology (Groom, 2005) by which disciplines can be differentiated. Also, analysis of the corpus leads to the suggestion that these high-frequency collocations are an important part of RAs and certainly part of the defining code (Williams, 2002) of RAs. They therefore represent disciplinary norms, and it is suggested that the different patterns presented are accepted within different disciplines as recognized ways for writers to describe and discuss their research. And as Hyland (2000: 78) notes, writers need to "project an insider ethos". He also proposes (1999) that discipline differences reflect rhetorical constraints within a discipline. Schmitt and Carter (2004) state that if a sequence is frequent in a corpus, this indicates it is conventional within the discourse community. This study has revealed some of these conventional forms in various disciplinary corpora.

Examination of Tables 2 and 3 reveals that the differences are between individual disciplines rather than more broadly between the four science disciplines as a whole and the four non-science disciplines. However, there are fewer high-frequency collocations in Chemistry and Materials Science. This can be explained, for just ten of the nouns (but not other nouns, across the whole corpus), by their low frequency – the noun frequency appears to be too low to allow the occurrence of any high-frequency collocations. The Chemistry nouns are "information" (220 pmw), "research" (60 pmw), "experiment" (210 pmw), "variable" (110 pmw), and "evidence" (220 pmw). In Materials Science they are "information" (120 pmw), "research" (90 pmw), "variable" (50 pmw), and "evidence" (160 pmw). Finally, there is just one in Management, "experiment" (80 pmw). As noted above, the description and discussion of the distribution of the other high-frequency nouns themselves is beyond the scope of this research.

Implications for teaching

Collocations are an important part of language knowledge, and need to be included in syllabus content (Willis, 1990; Lewis, 1993). Lewis (1993: 125-128) provides a valuable list of teaching suggestions, as do a number of chapters in his later book (Lewis, 2000a – see Hoey, 2000; Lewis, 2000b & 2000c; and Hill, 2000). The present research provides discipline-specific lists of high-frequency collocations of common nouns. These collocations have to be learned, stored and processed as complete units (Schmitt, Grandage &

Adolphs, 2004) by students in each discipline, and here, these lists may be of use. As noted above, writers need to learn collocations as a whole (Ellis, Simpson-Vlach & Maynard, 2008); Durrant (2009) proposes that learners need to learn high-frequency collocations. Two implications of the present findings for teaching research writing are that awareness of the discipline variations presented here is important for teaching, particularly to students of research writing, and that discipline-specific teaching of these collocations is certainly advisable. This might be especially important for NNS, who may be unaware of genre conventions and need help in joining the discourse community of international research (Paltridge, 1993), and fail when their work is written in an incorrect rhetorical style (Ahmad, 1997). These collocations are important in academic English, and if NNS make errors, they must be taught to NNS. This research can inform the teaching of research writing, and this is part of the usefulness of the variation.

As Ward (2007) achieved for one discipline, this study has accomplished for eight. The present findings are in agreement with Ward's assertion that collocations are very discipline specific. Analysis of the corpus found a number of disciplinary differences in the collocates of high-frequency nouns.

It is suggested that this research has added to the understanding of disciplinary conventions, including discipline differences, and of collocation. The present findings should improve knowledge of RAs and have relevance for the teaching of research writing to NNS and to NS, and help teachers prepare discipline-specific materials to teach collocation.

[Paper received 29 July 2010]
[Revised paper accepted 25 March 2011]

References

Ahmad, U.K. (1997). "Research article introductions in Malay: Rhetoric in an emerging research community" in A. Duszak (ed.), 273-301.

Bahns, J. & M. Eldaw (1993). "Should we teach EFL students collocations?" *System* 21: 101-114.

Bhatia, V.K. (2000). "Genres in conflict" in A. Trosborg (ed.), *Analysing Professional Genres,* 147-161. Amsterdam: John Benjamins.

Biber, D., S. Conrad & V. Cortes. (2004). "*If you look at*…: Lexical bundles in university teaching and textbooks". *Applied Linguistics* 25: 371-405.

Cantos, P. & A. Sanchez. (2001). "Lexical constellations: what collocates fail to tell". *International Journal of Corpus Linguistics* 6: 199-228.

Chan, T-P. & H-C. Liou. (2005). "Effects of web-based concordancing instruction on EFL students' learning of verb-noun collocations". *Computer Assisted Language Learning* 18: 231-250.

Durrant, P. (2009). "Investigating the viability of a

collocation list for students of English for academic purposes". *English for Specific Purposes* 28: 157-169.

Duszak, A. (ed.) (1997). *Culture and Styles of Academic Discourse*. Berlin: Mouton de Gruyter.

Ellis, N.C. & R. Simpson-Vlach. (2009). "Formulaic language in native speakers: triangulating psycholinguistics, corpus linguistics, and education". *Corpus Linguistics and Linguistic Theory* 5: 61-78.

Ellis, N. C., R. Simpson-Vlach & C. Maynard. (2008). "Formulaic language in native and second language speakers: psycholinguistics, corpus linguistics, and TESOL". *TESOL Quarterly* 42: 375-396.

Firth, J.R. (1957). *Papers in Linguistics 1934-1951*. Oxford: Oxford University Press.

Firth, J.R. (1968). *Selected Papers of J.R. Firth 1952-59*. Edited by F.R. Palmer. Bloomington and London: Indiana University Press.

Gledhill, C. J. (2000a). *Collocations in Science Writing*. Tubingen: Gunter Narr Verlag.

Gledhill, C.J. (2000b). "The discourse function of collocation in research article introductions". *English for Specific Purposes* 19: 115-135.

Golebiowski, Z. (1999). "Application of Swales' model in the analysis of research papers by Polish authors". *IRAL* 37: 231-247.

Groom, N. (2005). "Pattern and meaning across genres and disciplines: an exploratory study". *Journal of English for Academic Purposes* 4: 257-277.

Herbst, T. (1996). "What are collocations: sandy beaches or false teeth?" *English Studies* 4: 379-393.

Hill, J. (2000). "Revising priorities: From grammatical failure to collocational success" in M. Lewis (ed.), 47-69.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hoey, M. (2000). "A world beyond collocation: New perspectives on vocabulary teaching" in M. Lewis (ed.), 224-245.

Hoey, M. (2007a). "Lexical priming and literary creativity" in M. Hoey et al. (eds.), 7-29.

Hoey, M. (2007b). "Grammatical Creativity: A Corpus Perspective" in M. Hoey et al. (eds.), 31-56.

Hoey, M. & M.B. O'Donnell. (2008). "Lexicography, grammar, and textual position". *International Journal of Lexicography* 21: 293-310.

Hoey, M., M. Mahlberg, M. Stubbs & W. Teubert (eds.) (2007). *Text, Discourse and Corpora*. London: Continuum.

Hyland, K. (1996). "Talking to the academy: forms of hedging in science research articles". *Written Communication* 13: 251-281.

Hyland, K. (1998). *Hedging in Scientific Research Articles*. Amsterdam: John Benjamins.

Hyland, K. (1999). "Disciplinary discourses: Writer stance in research articles" in C. Candlin & K. Hyland (eds.), *Writing: Texts, Processes and Practices,* 99-121. London: Longman.

Hyland, K. (2000). *Disciplinary Discourses: Social Interactions in Academic Writing*. Harlow, Essex: Longman.

Jones, M. & S. Haywood. (2004). "Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context" in N. Schmitt (ed.), 269-300.

Lewis, M. (1993). *The Lexical Approach*. Hove: Language Teaching Publications.

Lewis, M. (ed.) (2000a). *Teaching Collocation: Further Developments in the Lexical Approach*. Hove: Language Teaching Publications.

Lewis, M. (2000b). "Language in the lexical approach" in M. Lewis (ed.), 126-154.

Lewis, M. (2000c). "Materials and resources for teaching collocation" in M. Lewis (ed.), 186-204.

Mahlberg, M. (2003). "The textlinguistic dimension of corpus linguistics: the support function of English general nouns and its theoretical implications". *International Journal of Corpus Linguistics* 8: 97-108.

Martinez, I. A., S.C. Beck & C.B. Panza (2009). "Academic vocabulary in agriculture research articles: A corpus-based study". *English for Specific Purposes* 28: 183-198.

Mudraya, O. (2006). "Engineering English: a lexical frequency instructional model". *English for Specific Purposes* 25: 235-256.

Paltridge, B. (1993). "Writing up research: a systemic functional perspective". *System* 21: 175-192.

Schmitt, N. (ed.) (2004). *Formulaic Sequences*. Amsterdam: John Benjamins.

Schmitt, N. & R. Carter. (2004). "Formulaic Sequences in Action" in N. Schmitt (ed.), 1-22.

Schmitt, N., S. Grandage & S. Adolphs. (2004).

"Are corpus-derived recurrent clusters psycholinguistically valid?" in N. Schmitt (ed.), 127-151.

Scott, M. (2004). *WordSmith Tools Version 4*. Oxford: Oxford University Press.

Shei, C.-C. & H. Pain. (2000). "An ESL writer's collocational aid". *Computer Assisted Language Learning* 13: 167-182.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stubbs, M. (1995). "Collocations and cultural connotations of common words". *Linguistics and Education* 7: 379-390.

Tognini-Bonelli, E. (2004). "Working with corpora: Issues and insights" in C. Coffin, A. Hewings & K. O'Halloran (eds.), *Applying English Grammar: Functional and Corpus Approaches,* 11-24. London: Arnold.

Vassileva, I. (1997). "Hedging in English and Bulgarian academic writing" in A. Duszak (ed.), 203-221.

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

Ward, J. (2009). "A basic engineering English word list for less proficient foundation engineering undergraduates". *English for Specific Purposes* 28: 170-182.

Williams, G. C. (1998). "Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles". *International Journal of Corpus Linguistics* 3: 151-171.

Williams, G. (2002). "In search of representativity in specialised corpora: categorisation through collocation". *International Journal of Corpus Linguistics* 7: 43-64.

Willis, D. (1990). *The Lexical Syllabus*. London: Collins.

Wood, A. (2001). "International scientific English: The language of research scientists around the world" in J. Flowerdew & M. Peacock (eds.), *Research Perspectives on English for Academic Purposes,* 71-83. Cambridge: Cambridge University Press.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Yakhontova, T. (1997). "The signs of a new time: Academic writing in ESP curricula of Ukrainian universities" in A. Duszak (ed.), 103-112.

Matthew Peacock teaches in the Department of English at the City University of Hong Kong. His research interests include English for Specific Purposes, corpus analysis, research writing, genre analysis, and TEFL methodology. In 2001, he co-edited (with John Flowerdew) a collection from Cambridge University Press, *Research Perspectives on English for Academic Purposes*.