RiCL Research in Corpus Linguistics

Corpus evidence for a discipline-specific phraseological approach to academic vocabulary

Geraint Paul Rees[1]
Universitat Pompeu Fabra / Spain

Abstract – This study examines the validity of the rationale underlying recent trends towards discipline-specific and phraseological approaches to vocabulary selection for English for Academic Purposes (EAP) courses. It examines the behaviour of Coxhead's (2000) New Academic Wordlist (AWL) using a 2,795,031 word corpus compiled from journal articles taken from the disciplines of History, Microbiology, and Management Studies. A two-stage method of analysis is employed. Firstly, coverage statistics for all AWL word families and their members are compared across the History, Microbiology, and Management Studies sub-corpora. This suggests difference in language use across disciplines. This difference is investigated further in a second stage of analysis which employs the Sketch Engine (Kilgarriff et al. 2004) Word Sketch Difference tool and Corpus Pattern Analysis (Hanks 2004) techniques to examine the collocational behaviour of a sample of 57 AWL headwords across the three sub-corpora. The results demonstrate that a large number of the AWL words have discipline-specific meanings, and that these meanings are conditioned by the syntagmatic context of the AWL item.

Keywords – corpus pattern analysis, English for Academic Purposes, language analysis, wordlist, specialist lexicography, Sketch Engine

1. Introduction

There is little doubt as to the importance of lexis in the teaching and learning of English for Academic Purposes (EAP). From the perspective of students, a number of works which employ student interviews and standardised language assessment measures present vocabulary knowledge as a limiting factor in academic achievement (Leki and Carson 1994; Evans and Green 2007; Berman and Cheng 2010; Evans and Morrison 2011). Similarly, from the perspective of instructors, lexis accounts for the greatest amount of variance in grading of writing (Astika 1993). University professors perceive lexical errors to be the most severe of all the errors they encounter in students' essays (Santos 1988). In more general terms, Nation (2013) regards direct vocabulary learning as key to non-native-speaker students being able to compete with native-speakers at an English medium university.

There is an abundance of studies highlighting how vocabulary might best be learnt and taught in an academic context. For example, reading or listening to academic texts (Bramki and Williams 1984; Flowerdew 1992). In a study involving EAP students, File and Adams (2010) find that isolated vocabulary instruction before reading leads to greater retention than vocabulary instruction integrated with reading. Both methods of instruction were statistically more effective than incidental instruction. As for speaking, activities should reflect academic presentations or discussions. Nikhat Shameem and Alison Hamilton-Jenkins (see Peacock and Flowerdew 2001: 259) developed the idea of issue logs with the aim of promoting fluency. Issue logs involve a learner choosing a topic on which they become experts on throughout the course of a pre-sessional EAP programme. They improve their fluency with specialised vocabulary by providing oral and written reports on their chosen topic throughout the course. As far as writing is concerned, practice

[1] This research was supported by an IULA-UPF doctoral research grant from Universitat Pompeu Fabra, within the project 'Significado léxico dependiente del contexto y polisemia, y las implicaciones para la representación lexicográfica' (FFI2015-70375- P, MINECO/FEDER).

writing formal texts, literature reviews, and written summaries all provide students the opportunity to use vocabulary in an academic context.

While there have been numerous and diverse responses to the question of how to teach EAP vocabulary, responses to the problem of which words to teach have been more limited in scope. The predominant approach to vocabulary selection in EAP involves wordlists. Coxhead's (2000) *New Academic Wordlist* (AWL) is perhaps the most well-known and most widely-used example. As such, it provides the ideal focus for the present evaluation of the wordlist approach.

The AWL is the first wordlist to be compiled using a machine-readable corpus. It is the attempt at developing an academic wordlist which has received the most attention over the last decade. It comprises a list of 570 headwords with the members of their word families, compiled from a 3.5 million-word corpus of written academic text. In common with many studies of specific language the AWL relies on frequency-based filters to distinguish between general and specific vocabulary. Coxhead (2000) applies filters based on the range, frequency, and specialised occurrence of words. To account for specialised occurrence only words outside the first 2,000 most frequently occurring words of English, as described by the *General Service List* (GSL) (West 1953) are included. To account for range it is stipulated that member of a word family must occur at least ten times in sections representing arts, commerce, law, and science; and in at least 15 of 28 more specific subject areas. To account for frequency an overall threshold of 100 occurrences in the corpus is set. The underlying idea is that a student with knowledge of the words in the AWL and GSL combined would have sufficient vocabulary to survive in an academic context (Hancioğlu et al. 2008). It is worth noting that a number of studies on specific-purposes language have used other criteria in addition to frequency-based criteria to differentiate general and specific vocabulary; for example, Rea Rizzo (2008) employs frequency and distribution measures along with an analysis of syntagmatic relations in an examination of specialised telecommunications vocabulary, while Alonso Campo and Renau Araque (2013) consider semantic and syntagmatic criteria to differentiate between general and specialised uses of verbs in their examination of environmental terminology.

Overall the AWL has been well received, although a number of its underlying assumptions are problematic. Firstly, the notion of a general-academic vocabulary is disputed. Trimble (1985) argues that academic and technical vocabulary takes on extended and distinct meaning in different technical contexts. For example, *fast* in Medicine has the sense 'resistant to' while in Mining it is used to describe 'solid ground'. Secondly, the meaning of a word is conditioned by its collocates; collocation is understood here in the broadest Sinclairian sense of any co-occurrence of words in a text. With this in mind, any useful approach to vocabulary selection must take into account phraseological concerns. To illustrate the importance of phraseology in conditioning meaning, Hanks (2013) argues that while the verb *switch* has a sense synonymous with the verb *change*, when accompanied by the particle *on* its sense is more closely related to the phrase 'to start a device'. Similarly, the meaning of the phrase 'to treat someone' which implies 'to apply a process or medicine to treat a disease' changes completely with the addition of the adverb *badly*. This is not merely a question of using parts of speech to disambiguate meaning.

(1) It is not possible to check the accuracy of the figures.
(1a) The DPK said that Kurdish guerrillas had checked the advance of government troops north of Sulaimaniya.

Sentences (1) and (1a) (Hanks 2013: 58) both contain the verb *check* + NP. However, it is clear from the context in (1) that *check* means 'inspect', while in (1a) it means 'stop'.

The idea that word frequency can be used as a measure of lexical specificity underlies the use of wordlists to select vocabulary for EAP courses and materials. Many researchers in English vocabulary studies conceive of words as belonging to one of three groups which can be placed on a cline according to the frequency with which they are used. At one pole lies the most frequent general vocabulary which comprises the first 2,000 to 3,000 most commonly occurring words. At the opposite end of the cline specialist technical vocabulary can be found. A third group of 'semi-technical' or 'academic' vocabulary lies between these poles (see Nation 2013: 16–20). Proponents of this approach point out that academic vocabulary is common to a wide range of academic texts (Barber 1962; Campion and Elley 1971; Praninskas 1972; Xue and Nation 1984; Coxhead 2000). They also point out that it accounts for a reasonable number of words in academic texts. For example, there have been many studies of the coverage of Coxhead's (2000) AWL in corpora representing different academic disciplines. Cobb and Horst (2004) find coverage of 11.6% in the Learned section of the Brown corpus (Francis and Kučera 1979); Hyland and Tse (2007) find coverage of 10.6% in a corpus of Science and Engineering texts by professional and student writers; Chen and Ge (2007) find coverage of around 10% in Medicine research articles; Konstantakis (2007) finds coverage of 11.5% in Business; Coxhead and Hirsh (2007) find 8.9 % in Science; Ward (2009) finds 11.3% in Engineering; Martínez et al. (2009) find 9% in Agricultural Sciences research articles; Vongpumivitch et al. (2009) find 11% in Applied Linguistics research articles; Li and Qian (2010) find 10.5% in Finance.

A consequence of the emphasis on frequency has been a proliferation of frequency-based general-academic wordlists which respect the general/sub-technical/technical distinction. Campion and Elley (1971) is the first attempt to create a list of academic vocabulary with reference to real academic language, and like Praninskas (1972) it is based on a manually compiled corpus of academic language. Coxhead's (2000) AWL is the first to be compiled using a machine-readable corpus. More recently, *The New Academic Vocabulary List* (Gardner and Davies 2014) employs a significantly larger machine-readable corpus and unlike the AWL it is determined on the basis of lemmas not word families.

In contrast, some researchers do not employ the general/sub-technical/technical distinction. Lynn (1973) and Ghadessy (1979) note those words which students have annotated in their textbooks. Xue and Nation's (1984) *University Wordlist* is the amalgamation of these notation lists and Campion and Elley's (1971) and Praninskas's (1972) manual corpus work. Ward (1999) argues that the three-way distinction is not necessary for learners who have clear specialised goals. More recently there has been a great deal of interest in the creation of discipline-specific wordlists, for example in Engineering (Ward 2009), Agriculture (Martínez et al. 2009), Environmental Studies (Liu and Han 2015), Nursing (Yang 2015), and Medicine (Lei and Liu 2016).

Given the doubt about the assumptions underlying general academic wordlists like the AWL, the trend towards more discipline-specific academic wordlists is a positive advance. However, single-word wordlists cannot take phraseological concerns into account, nor can they deal with a great deal of discourse organising vocabulary which plays a key role in academic language. These concerns have led to the creation of various lists of academic phrases; some deal with general academic vocabulary (Biber et al. 2004; Biber and Barbieri 2007; Ädel and Erman 2012; Pérez-Llantada 2014; Kashiha 2015). Others provide more discipline-specific lists of phrases. For example, lexical bundles in History and Biology (Cortes 2004), collocations in science writing (Gledhill 2000), lexical bundles in pharmaceutical articles (Grabowski 2015), collocational frameworks in medical research papers (Marco Luzón 2000), lexical bundles in Applied Linguistics research articles (Le and Harrington 2015), collocations in Engineering (Ward 2007), formulaic sequences for prospective EFL Business studies students (Hsu 2014), collocational networks in plant biology research articles (Williams 1998), and multiword terms in science texts (Yang 1986).

As is the case with wordlists in general, the hegemony of frequency in vocabulary studies limits the usefulness of many of these phrase lists. The majority of such lists are created using by extracting n-grams which occur above a given frequency threshold. In English vocabulary studies this is known as the lexical bundle approach (Biber and Conrad 1999). Although this approach is straightforward to implement and has resulted in interesting findings, it is limited in so much as it only returns strings of words of a given length which occur above a frequency threshold. Semantic considerations are not taken into account. In short, the consequences are phrases such as 'to do with the' or 'I think it was', which are intuitively not relevant, or in Simpson-Vlach and Ellis's (2010: 490) terms not "psycholinguistically salient".

Simpson-Vlach and Ellis (2010) mitigate the problem of nonsensical phrases and those not worth teaching time by applying an expert-judged teaching-worth measure. This results in more useful output but cannot account for the semantic information lost in the initial selection process. Similarly, *The Academic Collocation List* (Ackermann and Chen 2013) employs a mixed-method approach of statistical collocation and expert judgment to derive a list of lexical collocations which the authors claim is immediately useful to EAP teachers.

2. Method

2.1. Aims and hypotheses

This study has the principal aim of examining the assumptions underlying recent trends in the selection of vocabulary for EAP purposes; namely:

(i) academic vocabulary is discipline-specific.
(ii) the meaning of academic vocabulary is conditioned by phraseological context.

Additionally the study aims to point the way to a corpus-based methodology for the selection of phraseological vocabulary for EAP that, unlike the lexical bundle approach, can be feasibly employed with large amounts of data and take into account semantic concerns in addition to frequency.

Although the AWL is not the newest academic wordlist, it is the most well-known and most widely-used. For this reason it is chosen as the focus of the present study. The above assumptions are examined in systematic corpus-based experiment designed to test the following hypotheses:

(i) AWL words demonstrate distinct meanings in corpora from different academic disciplines.
(ii) The meaning of AWL words is conditioned by the words with which they co-occur.

2.2. The Academic Journal Article Corpus

The study tests the above hypotheses with regards to written language. This is not because these hypotheses only apply to this mode, but rather a result of practical constraints such as the difficulty inherent in obtaining sufficient academic gesture or speech to compile a corpus, and the large quantity of time which would be needed to encode this data. The written mode, in contrast, is not subject to these constraints; electronic academic journals provide a convenient source of machine-readable academic language which is already classified by discipline.

Determining what constitutes an academic discipline is a controversial issue. This is especially true given the greater emphasis which has been placed on interdisciplinary approaches in recent years. In addressing this issue, the present study follows the strategy adopted by Cortes (2004) and Hyland (2008), who limit their comparisons of academic language to customarily unrelated fields. For example, Cortes (2004) examines differences between History and Biology, and Hyland (2008) differences between Electrical Engineering, Business Studies, and Applied Linguistics. Following this example, the present study examines academic language across three sub-corpora: History, Microbiology, and Management Studies. The examination of three sub-corpora is methodologically advantageous since it allows space triangulation (Burns 2010). This permits more confidence in the results, since while it is possible that a difference in language between two sub-corpora could occur as a result of chance, it is less probable that this difference would occur by chance across all three disciplines.

The study employs a bespoke 2,795,031 token corpus; the Academic Journal Article Corpus (AJAC). The AJAC is split into three discipline-specific sub-corpora; History which contains 1,138,452 tokens, Microbiology which contains 678,721 tokens, and Management Studies which contains 985,486 tokens.

2.3. Procedure

2.3.1. Compilation of the corpus

A sampling frame is adopted to ensure that the corpus is balanced and representative of language in the three disciplines. Academic research is not immune to the quirks of fashion; certain topics, methods, or approaches become fashionable only to disappear after a few years. These trends influence the language employed in the discipline and, in this respect, they represent a deviation from the norm. In order for the range of language contained in the corpus to not be adversely influenced by such deviations, the corpus samples journal articles over a ten-year period spanning from 2001 to 2011. It is not only temporal concerns which affect the range of language the corpus contains, but also the scope of the journal from which the articles are sampled. For this reason the corpus does not include articles from journals limited to extremely narrow sub-disciplines, since the language they contain is likely to be atypical of the discipline as a whole. In addition to this exclusion, journal rankings and citation indexes provide a further means of ensuring that the articles sampled are typical. Starting from the assumption that those journals with a high average number of weighted citations contain language which is representative of their disciplines, the present study employs the SCIMago (2007) Journal and Country Rank to categorise journals according to their discipline and popularity. The corpus contains only those articles from journals with a Journal Ranking Indicator, a measure of a journal's impact, influence, or prestige, in the first quartile for their discipline. Another factor which might affect the typicality of the language sampled is the dialect of English used in the journal. For example, Liu (2012) claims that while writers from the US prefer the phrase *in general*, British writers prefer the phrase *as a whole*. To counter this reason, care has been taken to include articles from journals published in both the United States and the United Kingdom. However, it must be noted that academic publication is a global phenomenon and that as a result all the journals sampled accept contributions in both dialects of English. Table 1 provides a summary of the composition of the AJAC.

Journal sampled	Issues published per year	Sample periodicity	Number of articles sampled	Av. length of article sampled (tokens)	English
Historical Journal	4	1 article every issue	40	10,560	UK
Journal of World History	4	1 article every issue	40	10,648	US
Past and Present	4 (plus supplements)	1 article every issue	40	10,893	UK
Total History			120	10,700	
Cell	24	1 article every 6th issue	40	6,750	US
Developmental Cell	24	1 article every 6th issue	40	7,099	US
Nature Genetics	12	1 article every 3rd issue	40	6,910	UK
Total Microbiology			120	6,920	
Human Resources Management Review	4	1 article every issue	40	9,413	UK
The Leadership Quarterly	6	1 article from 4 issues a year	40	9,527	US
The Journal of Management Studies	8	1 article every 2nd issue	40	9,801	UK
Total Management			120	9,580	
Overall total			360	9,066	

Table 1: Characteristics of the articles in AJAC

In addition to aiding representativeness, sampling equally over a ten-year span also ensures the balance of each sub-corpus, which contains an equal number of journal articles from each year. Similarly, sampling an equal number of articles from three journals per discipline, and selecting the article from the sampled journal at random, ensures that the sub-corpora are more representative than they would be had only one journal been sampled, while mitigating the risk of confirmation bias.

2.3.2. Selection of AWL sample

Having compiled the corpus, the next step in the procedure involves deciding which AWL headwords to include in an in-depth examination of collocational patterns. The present study adopts a randomised sampling approach; each AWL headword is assigned a number from 1 to 570. 57 numbers, 10% of the AWL headwords, are selected using a random number generator. The figure of 10% is judged to be sufficient since the hypotheses set out above do not assume a significant difference between the number of discipline-specific and general AWL headwords. Instead, hypotheses are accepted if a reasonable number of AWL headwords are found to exhibit discipline-specific behaviour. If a significant difference were assumed, a larger sample size would be required. The random element is intended to counter any confirmation bias on the part of the analyst. The headwords sampled are listed in Table 2.

Headwords sampled			
access	establish	invoke	resolve
accompany	exhibit	label	role
accurate	export	lecture	route
aspect	expose	migrate	schedule
assemble	file	mode	sequence
assume	focus	orient	sex
attach	function	partner	shift
challenge	fundamental	phase	stress
coincide	identify	potential	structure
collapse	implicate	proceed	suspend
commit	individual	proportion	text
confer	infer	radical	trace
considerable	inhibit	recover	trigger
demonstrate	integrate	regime	uniform
distinct			

Table 2. AWL sample

3. ANALYSIS AND DISCUSSION

3.1. Coverage and type token ratio

	History	Microbiology	Management	Overall
Tokens	1,138,452	678,721	985,486	2,802,659
Types	43,357	20,409	25,732	89,498
Standardised Type-Token-Ratio (SD)	44.82 (54.71)	35.12 (64.23)	39.55 (60.29)	40.86 (59.06)
Overall AWL coverage[2]	5.61%	6.87%	10.73%	7.73%

Table 3: AJAC composition and AWL coverage

As a step prior to the examination of the AWL sample, it is worthwhile examining some characteristics of the corpus in general and coverage statistics for the AWL as a whole, that is, every headword in the AWL and its word family members. The standardised type-token-ratio shows the percentage of new types for every 1,000 tokens, and thus gives an indication of the variety of vocabulary in the corpora. The figures show that the History sub-corpus contains the greatest variety of vocabulary, while the Microbiology sub-corpus is the least varied. There are also differences in the AWL coverage across sub-corpora; overall the AWL covers 7.73% of tokens in the AJAC corpus, but this figure varies across sub-corpora; for instance, in the History sub-corpus the figure is 5.61% and in the Management Studies sub-corpus it is 10.73%.These figures are interesting for two reasons: firstly, with the exception of the figure for Management, they differ from the coverage statistics reported in previous studies of AWL coverage of academic language; secondly, they suggest that academic language is more lexically varied than Coxhead (2000) anticipates.

In her original evaluation of the AWL Coxhead (2000) reports coverage percentages between 9.3% in a corpus of academic language from the Arts and 12% in a corpus of academic language from Commerce. In an overview of studies on AWL coverage, Coxhead (2011) presents coverage as ranging from 1.4% in a corpus of fiction to 11.51% in a corpus of Business Studies language (Konstantakis 2007). As in previous studies, the coverage percentage of 10.73% obtained for the Management sub-corpus in the present study demonstrates that the AWL is biased towards the disciplines of Business Studies and Commerce. The coverage percentages obtained for the History and Microbiology sub-corpora are lower than those obtained for academic language in previous studies.

There are several possible reasons for this discrepancy. The first is that the methods used to calculate the coverage statistics reported in the present study are not appropriate. Although this is possible, it is worth noting that this simple procedure for calculating the coverage percentage employed here is the same as the one employed by Coxhead (2000) and a number of subsequent studies of AWL coverage.

The second possibility is that the AJAC corpus does not represent academic language. Given the measures taken at the corpus compilation stage to ensure a balanced corpus representative of academic language, this seems unlikely. In fact, the extent to which corpora employed to test AWL coverage in other studies are balanced and reasonably

---

[2] The Sketch Engine (Kilgarriff et al. 2004) wordlist tool is employed to calculate the sum of the frequencies of occurrence of AWL headwords and all their family members in the AJAC. This procedure is first conducted for the corpus as a whole and then for each sub-corpora. The relevant sum of the raw frequencies is expressed as the percentage of the total number of tokens in the whole corpus, or the sub-corpus to which it pertains.

representative of academic language can be questioned. For example, a number of studies employ academic language from a limited number of sources, and some, such as Hyland and Tse (2007), also include undergraduate projects in their test corpora. Limitations of representativeness are found to some extent in all corpora; representativeness is a question of degree. An important caveat here is that no corpus can ever be entirely representative of language employed in a natural communicative setting and that in absolute terms no corpus is representative of anything other than itself. Nonetheless, Coxhead (2000) has been criticised for confirmation bias while constructing her test corpora. Hyland and Tse accuse her of being "opportunistic" (Hyland and Tse 2007: 239) in her approach to sampling by including more texts from certain disciplines than from others.

The third possibility is that, in practice, academic language is richer and more varied than some researchers in EAP and lexical studies believe. The coverage statistics presented here suggest that it is not the case that a student with knowledge of the words in the AWL and GSL combined would have sufficient vocabulary to survive in an academic context.

A comparison of both standardised type-token-ratios and AWL coverage of distinct academic disciplines brings to light some differences in academic language across disciplines and suggests that there are problems inherent in single-discipline approaches to the selection of vocabulary for EAP. An examination of the collocational behaviour of AWL words reveals further differences and problems.

3.2. Comparison of collocational behaviour

An initial comparison of collocational behaviour takes place using *the Sketch Engine's* (Kilgarriff et al. 2004) Word Sketch Differences (WSD) tool. WSD allows the visual comparison of words in different sub-corpora according to their salient collocational context. The WSD tool arranges the collocates of a given word in sets according to their syntactic argument role. The user can see the keyword and collocate in context by clicking on the figure to the right of the collocate. For each of the 57 AWL headwords sampled, three WSDs are derived in total. This allows the comparison of the headwords' collocational behaviour between the History and the Microbiology sub-corpora, the History and the Management Studies sub-corpora, and the Microbiology and the Management Studies sub-corpora. An examination of the collocation behaviour in the different WSD measures (subject, object, modifier, etc.) elucidates differences in semantics across the different sub-corpora. Headwords that demonstrate such differences in three or more WSD measures are judged to be discipline-specific. The results of the comparison of the collocational behaviour of 57 headwords across sub-corpora are given in Table 4.

	Discipline-specific headwords
History/Microbiology	35
Microbiology/Management	29
Management/History	25
Overall	49

Table 4: Discipline-specific headwords across disciplines

A comparison of the History and Microbiology sub-corpora produced the most headwords with discipline-specific behaviour in three WSD measures. In this comparison, 35 out of 57 headwords were analysed as being discipline-specific. An analysis of headword collocational behaviour across the Microbiology and Management sub-corpora produced an almost even split between 29 discipline-specific and 28 non-discipline-specific headwords. An analysis of headword collocational behaviour across the Management and History sub-corpora resulted in the least variation, with only 25 headwords analysed exhibiting discipline-specific collocational behaviour. When the results of all three pair-wise comparisons are taken into account, only eight headwords, *considerable*, *fundamental*, *inhibit*, *distinct*, *lecture*, *migrate*, *schedule*, and *sequence*, show no discipline-specific collocational behaviour.

3.3. Discussion of collocational patterns

Due to space limitations a discussion of all 49 discipline-specific collocation patterns is not feasible. Instead, the discussion centres on the five verbal collocations which typify the differences encountered.

For explanatory ease the discussion makes use of the semantic type labels of Corpus Pattern Analysis (CPA) (Hanks 2004). CPA is a procedure for analysing meaning in keyword in context lines by combining syntax and semantics. It employs the SPOCA (Subject Predicator Object Complement and Adverbial) clause structure of Systemic Grammar (Halliday 1961). Each clause role is populated by a lexical set, i.e. a set of collocates which share a semantic meaning. Most lexical sets are assigned a semantic type label, enclosed in square brackets, which reflects this common meaning. Semantic types are often complemented by contextual roles, which are indicated after an equals sign. Curly brackets represent specific lexical sets which cannot be grouped into a semantic type. CPA patterns (2) and (2a) taken from the *Pattern Dictionary of English Verbs* (PDEV) (http://www.pdev.org.uk) exemplify typical CPA patterns. They are accompanied by implicatures and a typical example of the corpus evidence from which they are derived.

(2) Pattern: [[Human 1|Process|Drug]] treats [[Human 2|Animal|Disease|Injury]]
Implicature: [[Human 1 = Health Professional]] applies a [[Drug|Process = Medical]] to [[Human 2 = Patient]] for the purpose of curing the patient`s Disease or Injury
Example: *He was treated in hospital for facial cuts.*

(2a) Pattern: [[Human]] treats [[Inanimate]] ({with}[[Stuff]]{by}[[Process]])
Implicature: The chemical or other properties of [[Inanimate]] are improved or otherwise changed {by} [[Process]] or the application of [[Stuff]]
Example: *A section of 2-4 metres length should be thus treated.*

The discussion which follows employs the CPA notation for semantic types, semantic roles, and lexical sets. This is accompanied by the WSD argument position from which the evidence for the pattern found is taken and examples of the collocates which populate this position.

3.3.1. *Assemble*

CPA pattern	WSD argument position	Corpus evidence
assemble [[Group]]	Object	*chapter*, *delegation*, *parliament*, *district*, *crowd*, *regiment*, *spectator*, *onlooker*, *mujahidun*, *junta*
assemble {in} [[Location]]	Preposition	*in one place*, *in the city*, *in a space*
[[Person= Authority figure]] assembles [[Group]]	Preposition	*by the town clerk*, *by the king*

Table 5: Collocational behaviour of *assemble* in History

Assemble occurs 55 times in both the History and Microbiology sub-corpora of the AJAC. However, even a cursory glance at its argument structure in History and Microbiology, summarised in Tables 5 and 6 respectively, reveals that it exhibits a different sense in the two disciplines. In History the collocates in object position are of the semantic type [[Group]], whereas in Microbiology they suggest [[Physical Object]]. The collocates for the pattern *assemble* {in} also suggest a difference in sense. In History people assemble in [[Location]], whereas in Microbiology objects assemble in certain manners. A similar difference is apparent for the collocates of the pattern assemble {by}. In History a group is assembled by a person in authority, whereas in Microbiology objects, often parts of cells, are assembled by processes.

CPA pattern	WSD argument position	Corpus evidence
[[PhysObj]] assemble	Subject	*gene*, *ring*, *pore*
assemble [[PhysObj]]	Object	*unigenes*, *contigs*, *nucleoporins*, *scaffold*, *recombinosomes*, *draft*, *transcriptase*, *read*, *motor*, *heterochromatin*
assemble {in} [[Manner]]	Preposition	*in absence*, *in parallel*
[[Process]] assembles [[PhysObj]]	Preposition	*assembled by activation*

Table 6: Collocational behaviour of *assemble* in Microbiology

In the Management sub-corpus, *assemble* occurs 15 times. Its collocational behaviour is summarised in Table 7. A comparison with Microbiology reveals a key difference in subject position. In Microbiology physical objects and microbiological processes assemble objects, while in Management people or organisations assemble objects and groups.

CPA pattern	WSD argument position	Corpus evidence
[[Person]] assemble	Subject	*people*, *The Roundups*
[[Organisation]] assemble	Subject	*the company*, *the subsidiary*
assemble [[Group]]	Object	*workforce*, *actors*, *association*
assemble [[PhysObj]]	Object	*package*, *series*
-	and/or	*mobilise*, *work*, *deliver*, *leverage*

Table 7: Collocational behaviour of *assemble* in Management

There is a subtle difference between the pattern *assemble* [[Group]] in the History and Management sub-corpora. This is evidenced by the WSD 'and/or' measure, which presents those collocates which occur in close context to the keyword, but which are not necessarily closely syntactically related to it. In the History sub-corpus the words *arm* and *deploy* appear in this measure, thus suggesting that in this discipline *assemble* has a meaning similar to that of *convoke* or *mobilise* in the sense of forming an army or another group. In the Management sub-corpus the word *mobilise* also suggests this meaning; however, *work*, *deliver*, and *leverage* suggest a sense closer to ‘build’ or ‘construct’.

3.3.2. *Challenge*

Challenge appears 95 times in the History sub-corpus (Table 8), yet only nine times in the Microbiology sub-corpora (Table 9).

CPA pattern	WSD argument position	Corpus evidence
[[Human]] or [[Group=human]] challenge	Subject	*leader*, *woman*, *man*, *scholar*, *priest*, *parliament*, *philosopher*, *thinker*, *duke*, *pietists*, *poor*
challenge [[Human]]	Object	*stewards*, *jurors*, *constables*, *justices*
challenge [[Abstract]]	Object	*secondary scholarship on the history*, *ideas about the frontier and cultural contacts*, *the traditional male-centred history*

Table 8: Collocational behaviour of *challenge* in History

CPA pattern	WSD argument position	Corpus evidence
[[Process]] challenge	Subject	*repeat*, *identification*, *assay*
challenge [[Abstract]]	Object	*low copy repeats repetitive sequences*
challenge [[Process]]	Object	*assay identification*

Table 9: Collocational behaviour of *challenge* in Microbiology

In addition to this marked difference in frequency, an examination of the collocational behaviour of the verb illustrates differences in meaning according to discipline. This is particularly apparent in relation to subject position. A brief examination reveals that in History half of the 22 subjects are [[Human]] or [[Group=human]], while in the Microbiology sub-corpus they are all [[Process]]. The semantic types of the collocates in object position reflect the focus of each discipline. In History the object has the semantic type [[Human]], implying a sense of questioning, while in Microbiology the semantic type of the object is [[Process]], implying a sense of a process being made more difficult. In both disciplines there are also abstract entities in object position.

Challenge appears 78 times in the Management sub-corpus (Table 10). Its collocational behaviour is more varied than in the History or Management sub-corpora. [[Human]] or [[Group]] are present in subject position. While [[Abstract]], [[Human]], [[Process]] are present in object position.

CPA pattern	WSD argument position	Corpus evidence
[[Human]] or [[Group=human]] challenge	Subject	*researcher constituent*, *subordinates*
[[Abstract]] challenge	Subject	*concept*, *comparison*, *convergence*, *construction*, *feedback*
challenge [[Abstract]]	Object	*reliance*, *schemas*, *hegemony*, *over-emphasis*
challenge [[Process]]	Object	*operationalization*
challenge [[Human]]	Object	*caller*

Table 10: Collocational behaviour of *challenge* in Management

3.3.3. *Proceed*

Proceed occurs in the History sub-corpus 67 times. Its collocational behaviour here is summarised in Table 11.

CPA pattern	WSD argument position	Corpus evidence
[[Human|Process]] proceed	Subject	*king*, *husband*, *negotiation*, *debate*, *council*
proceed [Adv[Direction]]	Object	*north*, *south*, *east*, *west*
proceed [Adv[Manner]]	Preposition	*as if nothing had happened*, *around them*

Table 11: Collocational behaviour of *proceed* in History

In the majority of these occurrences its collocate in object position is a directional adverbial. This implies the sense ‘to follow’ or ‘to go forward or onward’. By contrast, there is only one occurrence in the Microbiology sub-corpus (Table 12), where *proceed* occurs 24 times overall, which suggests this sense, namely the presence of *initiate* in the ‘and/or’ measure. The same sense is found in the Management sub-corpus (Table 13), where *proceed* occurs 26 times overall.

However, here *proceed* is most frequently followed by an adverb of manner such as *consistently*, *quickly*, *slowly*, or a prepositional phrase denoting manner (*as follows*). It is worth noting that *proceed* also co-occurs with prepositional phrases denoting manner in History, where it co-occurs only twice with *as if nothing had happened* and *around them.* The discursive use of *proceed* in the sense 'to continue one's discourse' is present in Management. This is not true of the other two sub-corpora. The principal difference in subject position is that collocates of the semantic type [[Human|Process]] predominate in History and Management, whereas [[PhysObj]] predominates in Microbiology.

CPA pattern	WSD argument position	Corpus evidence
[[PhysObj]] proceed	Subject	*endoderm*, *MEFs*, *chromosome*, *cell*
proceed [[Activity]]	and/or	*initiate*

Table 12: Collocational behaviour of *proceed* in Microbiology

CPA pattern	WSD argument position	Corpus evidence
[[Human|Process]] proceed	Subject	*analysis*, *group*, *women*, *nation*, *speaker*, *scientist*
proceed [Adv[Manner]]	Object	*consistently*, *quickly*, *slowly*
proceed [Adv[Manner]]	Preposition	*as follows*
proceed {to} [[Activity]]	Object of to	*initiate*, *investigate*, *present*, *develop*, *suggest*

Table 13: Collocational behaviour of *proceed* in Management

3.3.4. *Stress*

Stress occurs 109 times in the History sub-corpus (Table 14). Its behaviour is very similar in the Management sub-corpus, where it occurs 56 times (Table 15).

CPA pattern	WSD argument position	Corpus evidence
[[Human]] or [[Document]] stress	Subject	*rubric*, *chronicle*, *argument*, *historian*, *bishop*
stress [[Abstract Entity]]	Object	*need*, *importance*, *opposition*, *danger*, *necessity*

Table 14: Collocational behaviour of *stress* in History

CPA pattern	WSD argument position	Corpus evidence
[[Human]] stress	Subject	*researcher*, *author*, *writer*
stress [[Abstract Entity]]	Object	*need*, *importance*, *consistency*, *exploitation*, *homogeneity*

Table 15: Collocational behaviour of *stress* in Management

In both sub-corpora the semantic types of its collocates in subject and object argument position imply a sense emphasising the importance of something. In both sub-corpora object argument position is populated by collocates of the semantic type [[Abstract Entity]]. In subject position in History there are collocates of the type [[Human]] or [[Document]], whereas in Management only [[Human]] is found. In the Microbiology sub-corpus, where *stress* occurs far less frequently with 13 occurrences (Table 16), this sense of emphasising the importance of something is also present in one argument in subject position. However, the dominant sense of stress in Microbiology is that of provoking a stimulus as evidenced by the presence of [[Entity = MB_Unit]] in object position and [[Process]] as an object of the preposition {by}.

CPA pattern	WSD argument position	Corpus evidence
[[Human|Document]] stress	Subject	*our findings*
stress [[Entity = MB_Unit]]	Object	*ER*, *organism*
stress {by} [[Process]]	Preposition	*by depletion*

Table 16: Collocational behaviour of *stress* in Microbiology

3.3.5. *Suspend*

In History (Table 17), where *suspend* occurs 15 times, and Management (Table 18), where it occurs three times, it exhibits a sense synonymous with *postpone*. In both of these sub-corpora, this sense is evidenced by lexical sets in object position of the semantic type [[Activity]].

CPA pattern	WSD argument position	Corpus evidence
suspend [[Activity]]	Object	*law*, *principle*, *debate*, *sentence*, *restriction*, *talk*
suspend [Adv[Position]]	Preposition	*in the air*, *in the middle*
suspend {from} [[Role]]	Preposition	*from parliament*, *from political office*

Table 17: Collocational behaviour of *suspend* in History

In Management, phrases such as *suspend the class/tutorial for a week* provide further evidence for this sense. In History, there is the sense of stopping someone from undertaking a job or position implied by {from} [[Role]]. Somewhat surprisingly, this sense is not found in the Management or Microbiology sub-corpora.

CPA pattern	WSD argument position	Corpus evidence
suspend [[Activity]]	Object	*class*, *tutorial*
suspend [[Activity]]	Preposition	*for a week*

Table 18: Collocational behaviour of *suspend* in Management

In Microbiology (Table 19), *suspend* occurs seven times with adverbs of manner or position as prepositional arguments which imply the sense ‘stopping something from moving’. This sense is also present in the History sub-corpus, where *suspend* occurs with prepositional phrases such as *in the air* and *in the middle*.

CPA pattern	WSD argument position	Corpus evidence
suspend [Adv[Manner]]	Preposition	*at concentration*
suspend [Adv[Position]]	Preposition	*in buffer*, *in µl range*

Table 19: Collocational behaviour of *suspend* in Microbiology

4. Conclusion

The present study aimed to test two hypotheses. The first hypothesis is that AWL words demonstrate distinct meanings in corpora from different academic disciplines. The veracity of this hypothesis is first suggested by a comparison of the text-coverage statistics for the entire AWL across three sub-corpora representing language from History, Microbiology, and Management Studies, respectively. An examination of the syntagmatic behaviour of a sample of AWL headwords confirms this suggestion along with the second hypothesis. This states that the meaning of AWL words is conditioned by the words with which they co-occur. In this regard, it has been demonstrated, for example, how *assemble* has a sense similar to *convoke* in History, while in Microbiology its sense isrange more akin to *build* or *construct*. The confirmation of both of these hypotheses provides evidence to support the assumptions underlying recent trends towards discipline-specific and phraseological approaches to vocabulary selection in EAP, empirically demonstrating that in EAP, as in General English, vocabulary is context specific and that its meaning is conditioned by phraseological context.

In examining the wordlist approach to the selection of vocabulary for EAP, this study has also made a methodological advance. As discussed above, previous phraseological studies of EAP lexis are usually concerned with lexical bundles and therefore take little or no account of semantic relations. The present study has demonstrated how a combination of the Sketch Engine, keyword in context, CPA and statistical comparisons are useful to deal with genuine usage-based academic language and offers the possibility of analysing both syntactic and semantic relations.

However, the present study does show certain limitations. Firstly, some of the examples of discipline-specific behaviour given above are based on a small number of corpus instances. The danger here is that such examples are outliers not representative of the language of a given discipline, but represent instead the idiosyncrasies of an individual author or editor. The procedures detailed in the present paper ensure that the results of this experiment are open to independent verification using different corpora. Moreover, the design includes three sub-corpora in order to allow triangulation. This permits more confidence in the results, since while it is possible that a difference in the behaviour of an AWL sample word between two sub-corpora could occur as a result of chance, it is less probable that such a difference would occur by chance across all three sub-corpora. Triangulation also plays a role in the conservative criteria adopted in the identification of discipline-specific headwords. A headword is only deemed to be discipline-specific if it exhibits differences in its collocational behaviour in regard to at least three different syntactic roles. In addition to the replicable design of the experiment and the use of triangulation, this experiment also combines qualitative and quantitative methods. Discipline-specific headword behaviour is first suggested by a comparison of word-coverage statistics, and then confirmed by qualitative analysis of the WSD.

The question of whether the differences in meaning across disciplines are relevant to EAP students is open to debate and beyond the scope of this paper. It is unlikely that when presented with the word *challenge* and a definition or translation into his/her L1, a student would later be able to recognise it in all the senses *to challenge a king*, *to overcome a challenge*, *a filtering process challenged by low copy receipts*. Nor is it probable that a student would need to recognise the word in all these senses. The issue for EAP instructors and lexicographers is, then, how to decide which vocabulary items to include in their resources and practice. It is clear that further corpus research has an important role to play in addressing this question.

References

Ackermann, Kirsten and Yu-Hua Chen. 2013. Developing the Academic Collocation List (ACL) – a corpus-driven and expert-judged approach. *Journal of English for Academic Purposes* 12/4: 235–247.

Ädel, Annelie and Britt Erman. 2012. Recurrent word combinations in academic writing by native and non-native speakers of English: a lexical bundles approach. *English for Specific Purposes* 31/2: 81–92.

Alonso Campo, Araceli and Irene Renau Araque. 2013. Corpus Pattern Analysis in determining specialized uses of verbal lexical units. *Terminàlia* 7: 26–33.

Astika, Gusti. 1993. Analytical assessments of foreign students' writing. *RELC Journal* 24/1: 61–70.

Barber, Charles. 1962. Some measurable characteristics of modern scientific prose. In Charles Barber ed. *Contributions to English syntax and philology*. Göteburg: Acta Universitatis Gothoburgensis, 21–43.

Berman, Robert and Liying Cheng. 2010. English academic language skills: perceived difficulties by undergraduate and graduate students, and their academic achievement. *Canadian Journal of Applied Linguistics* 4/1: 25–40.

Biber, Douglas and Federica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26/3: 263–286.

Biber, Douglas and Susan Conrad. 1999. Lexical bundles in conversation and academic prose. In Hilde Hasselggård and Signe Oksefjell eds. *Out of corpora: studies in honour of Stig Johansson*. Amsterdam: Rodopi, 181–190.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, Marie Helt, Victoria Clark, Viviana Cortes, Eniko Csomay and Alfredo Urzua. 2004. *Representing language use in the university: analysis of the TOEFL® 2000 Spoken and Written Academic Language Corpus*. Princeton: Educational Testing Service.

Bramki, Doudja and Ray Williams. 1984. Lexical familiarization in economics texts, and its pedagogic implications in reading comprehension. *Reading in a Foreign Language* 2/1: 169–181.

Burns, Anne. 2010. *Doing action research in English language teaching: a guide for practitioners*. New York: Routledge.

Campion, Mary and Warwick Elley. 1971. *The academic vocabulary list*. Wellington: New Zealand Council for Educational Research.

Chen, Qi and Guang-chun Ge. 2007. A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes* 26: 502–514.

Cobb, Tomm and Marlise Horst. 2004. Is there room for an Academic Word List in French? CHAP. In Paul Bogaards and Batia Laufer-Dvorkin eds. *Vocabulary in a second language: selection, acquisition and testing*. Amsterdam: John Benjamins, 15–38.
Cortes, Viviana. 2004. Lexical bundles in published and student disciplinary writing: examples from history and biology. *English for Specific Purposes* 23: 397–423.
Coxhead, Averil. 2000. A new Academic Word List. *TESOL Quarterly* 34: 213–238.
Coxhead, Averil. 2011. The Academic Word List 10 years on: research and teaching implications. *TESOL Quarterly* 45: 355–362.
Coxhead, Averil and David Hirsh. 2007. A pilot science-specific word list. *Revue française de linguistique appliquée* 7/2: 65-78.
Evans, Stephen and Christopher Green. 2007. Why EAP is necessary: a survey of Hong Kong tertiary students. *Journal of English for Academic Purposes* 6/1: 3–17.
Evans, Stephen and Bruce Morrison. 2011. The first term at university: Implications for EAP. *ELT Journal* 65/4: 387–397.
File, Kieran Andrew and Rebecca Adams. 2010. Should vocabulary instruction be integrated or isolated? *TESOL Quarterly* 44/2: 222–249.
Flowerdew, John. 1992. Definitions in science lectures. *Applied Linguistics* 13/2: 202–221.
Francis, W. Nelson and Henry Kučera. 1979. Brown Corpus manual. Providence, Rhode Island: Brown University. <http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM (accessed October 21, 2016)>
Gardner, Dee and Mark Davies. 2014. A new academic vocabulary list. *Applied Linguistics* 35/3: 305–327.
Ghadessy, Mohsen. 1979. Frequency counts, words lists, and materials preparation: a new approach. *English Teaching Forum* 17: 24–27.
Gledhill, Cristopher J. 2000. *Collocations in science writing*. Tübingen: Gunter Narr Verlag.
Grabowski, Łukasz. 2015. Keywords and lexical bundles within English pharmaceutical discourse: a corpus-driven description. *English for Specific Purposes* 38: 23–33.
Halliday, Michael. 1961. Categories of the theory of grammar. In Gunther Kress ed. *Halliday: system and function in language*. Oxford: Oxford University Press, 52–72.
Hancioğlu, Nilgün, Steven Neufeld and John Eldridge. 2008. Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes* 27: 459–479.
Hanks, Patrick. 2004. Corpus pattern analysis. In Geoffrey Williams and Sandra Vessier eds. *Proceedings of the Eleventh EURALEX International Congress*. Université de Bretagne-Sud, Lorient: EURALEX, 87–98.
Hanks, Patrick. 2013. *Lexical analysis: norms and exploitations*. Cambridge, Mass: MIT Press.
Hsu, Wenhua. 2014. The most frequent opaque formulaic sequences in English-medium college textbooks. *System* 47: 146-161.
Hyland, Ken. 2008. As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4–21.
Hyland, Ken and Polly Tse. 2007. Is there an "academic vocabulary"? *TESOL Quarterly* 41: 235–253.
Kashiha, Hadi. 2015. Recurrent formulas and moves in writing research article conclusions among native and nonnative writers. *Pusat Pengajian Bahasa dan Linguistik* 21/1: 47–59.
Kilgarriff, Adam, Pavel Rychlý, Pavel Smrž and David Tugwell. 2004. The Sketch Engine. In Geoffrey Williams and Sandra Vessier eds. *Proceedings of the Eleventh EURALEX International Congress*. Université de Bretagne-Sud, Lorient: EURALEX, 105–116.
Konstantakis, Nikolaos. 2007. Creating a business word list for teaching Business English. *Estudios de Lingüística Inglesa Aplicada (ELIA)* 7: 79–102.
Le, Thi Ngoc Phuong and Michael Harrington. 2015. Phraseology used to comment on results in the discussion section of applied linguistics quantitative research articles. *English for Specific Purposes* 39: 45–61.
Lei, Lei and Dilin Liu. 2016. A new medical Academic Word List: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes* 22: 42–53.
Leki, Ilona and Joan Carson. 1994. Students' perceptions of EAP writing instruction and writing needs across the disciplines. *TESOL Quarterly* 28/1: 81–101.
Li, Yongyan and David D. Qian. 2010. Profiling the Academic Word List (AWL) in a financial corpus. *System* 38: 402–411.
Liu, Dilin. 2012. The most frequently-used multi-word constructions in academic written English: a multi-corpus study. *English for Specific Purposes* 31: 25–35.
Liu, Jia and Lina Han. 2015. A corpus-based environmental Academic Word List building and its validity test. *English for Specific Purposes* 39: 1–11.
Lynn, Robert. 1973. Preparing word-lists: a suggested method. *RELC Journal* 4: 25–28.
Marco Luzón, María José. 2000. Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes* 19/1: 63–86.
Martínez, Lliana, Susana Beck and Carolina Panza. 2009. Academic vocabulary in agriculture research articles: a corpus-based study. *English for Specific Purposes* 28/3: 183–198.

Nation, Paul. 2013. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
Peacock, Matthew and John Flowerdew. 2001. *Research perspectives on English for academic purposes*. Cambridge: Cambridge University Press.
Pérez-Llantada, Carmen. 2014. Formulaic language in L1 and L2 expert academic writing: convergent and divergent usage. *Journal of English for Academic Purposes* 14: 84–94.
Praninskas, Jean. 1972. *American University Word List*. Harlow: Longman.
Rea Rizzo, Camino. 2008. El inglés de las telecomunicaciones: estudio léxico basado en un corpus lingüístico. PhD. Murcia: Universidad de Murcia.
Santos, Terry. 1988. Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly* 22/1: 69–90.
SCImago. 2007. SJR – SCImago Journal & Country Rank. <http://www.scimagojr.com (accessed October 21, 2016)>
Simpson-Vlach, Rita and Nick C. Ellis. 2010. An academic formulas list: new methods in phraseology research. *Applied Linguistics* 31: 487–512.
Trimble, Louis. 1985. *English for science and technology: a discourse approach*. Cambridge: Cambridge University Press.
Vongpumivitch, Viphavee, Ju-yu Huang and Yu-Chia Chang. 2009. Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes* 28: 33–41.
Ward, Jeremy. 1999. How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language* 12/2: 309–324.
Ward, Jeremy. 2007. Collocation and technicality in EAP engineering. *Journal of English for Academic Purposes* 6/1: 18–35.
Ward, Jeremy. 2009. A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes* 28: 170–182.
West, Michael P. 1953. *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Longmans, Green & Co.
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.
Xue, Guoyi and Paul Nation. 1984. A University Word List. *Language Learning and Communication* 3: 215–229.
Yang, Huizhong. 1986. A new technique for identifying scientific/technical terms. *Literary and Linguistic Computing* 1: 93–103.
Yang, Ming-Nuan. 2015. A nursing Academic Word List. *English for Specific Purposes* 37: 27–38.

Corresponding author
Dep. Traducció i Ciències del Lenguatge
Roc Boronat, 138
08018 Barcelona
e-mail: geraintpaul.rees@upf.edu

received: July 2016
accepted: November 2016