Magali Paquot
Centre for English Corpus Linguistics, University of Louvain, Belgium

TOWARDS A PRODUCTIVELY-ORIENTED ACADEMIC WORD LIST

1. Introduction

Most studies of vocabulary in English for Academic Purposes (EAP) have emphasized the importance of a 'sub-technical' or 'academic' vocabulary alongside core words and technical terms in academic discourse (cf. Nation 2001: 187-216). A number of word lists have been compiled to meet the specific vocabulary needs of students in higher education settings. The *Academic Word List* (Coxhead 2000) is the most widely used today in language teaching, testing and materials development. It consists of the vocabulary required for academic reading comprehension, which can be broadly defined as a vocabulary common to a wide range of academic texts but not so common in non-academic texts. L2 learners, as well as EFL teachers and textbook developers, would greatly benefit from the elaboration of a productive counterpart to the *Academic Word List as l*earners' needs and difficulties are clearly not the same in production as in reception. This should be reflected in the selection criteria of a productively-oriented academic word list. Although frequency remains an important criterion, it is only half of the story. A productively-oriented academic word list should also give L2 learners the lexical means necessary to do the things that academic writers do, e.g. stating a topic, hypothesizing, contrasting, exemplifying, explaining, evaluating, etc. Recent corpus-based studies of recurrent word combinations (Biber 2004; Biber et al. 1999), lexical phrases (Oakey 2002) and abstract nouns (Cowie 1997) in native academic writing have pointed to the existence of an EAP-specific phraseology. It is therefore particularly important that new words be introduced together with information on how to use them, especially their collocational and colligational environment. Descriptions of EAP words should then be refined on the basis of a careful analysis of learner corpus data that will highlight learners' attested 'difficulties' in using these words (cf. Flowerdew 1998; Granger 2004a). Learner corpora are the only type of data that can reveal learners' typical 'difficulties' in terms of underuse, overuse, misuse and idiosyncratic use of EAP words or multi-word sequences. These corpus findings will undoubtedly have pedagogical implications in the sense that teachers will, for example, be able to focus more on words that have been proved to be underused while at the same time warning their pupils against overusing certain words or multi-word sequences.

The aim of this paper is therefore twofold. First, it examines the major reasons why the *Academic Word List* is not ideally suited for productive purposes (section 2) and proposes a new extraction methodology for the design of a productively-oriented academic word list (section 3 and 4). Secondly, it illustrates how learner corpus data can be used to fine-tune the description of EAP words by providing valuable insights into learners' 'difficulties' in using them (section 5).

2. The *Academic Word List*

The *Academic Word List* (AWL) consists of 570 word families that are not in the first 2,000 most frequently occurring words of English as described in West's (1953) *General Service List* (GSL) but which have wide range and reasonable frequency of occurrence in a 3,500,000 word corpus of academic texts (e.g. *approximate, capacity, link, presume, summary, widespread*, etc.). Taken together, words of the GSL and the AWL plus discipline-specific items should approach the critical 95% coverage threshold necessary for reasonable reading comprehension (Nation 2001:197). While the AWL is certainly a good supplement to the GSL for receptive purposes, it is not ideally suited to productive purposes for four main reasons. First, a classification into word families, without information on frequency, is not particularly helpful to learners as not all members of a word family are likely to be as frequent and useful in EAP. For example, under the headword *item*, which has a relative frequency of 36 occurrences per million words in the Micro-Concord academic corpus (see section 3.1.), we find the noun *itemisation* and word forms of the verb *itemise*. However, the verb *itemise* has a relative frequency of 1 occurrence per million words and the noun *itemisation* does not appear at all in the same corpus. Secondly, as the AWL is based on word forms, meanings and parts of speech are not differentiated. For example,

we do not know if the word form *issue* is a noun or a verb. Thirdly, like any other EAP word list, the AWL is based on single words. This goes against current trends in language acquisition and foreign language learning that stress the importance of prefabricated language over 'slot-and-filler' models of language (cf. Lewis 2002; Wray 2002; Schmitt 2004). Finally, Coxhead's criterion of non-appearance in the GSL is not really appropriate when it comes to productive purposes as lexical items may be included in the 2,000 most frequent words but used differently in EAP. Nouns such as *example, problem, reason, argument* and *result* appear with particular high range and frequency in academic corpora but are not considered as EAP vocabulary by Coxhead as these items are already in the GSL.

3. Data and methodology

3.1. Data

The corpus used for this study is the *Micro-Concord corpus collection B* (MC) (Scott and Johns 1993), a 1,000,000 word corpus of published academic prose which consists of 33 texts.[1] The corpus is divided into five subcorpora of about 200,000 words each: arts, belief and religion, applied science, science and social science. The classification of texts into five broad academic disciplines is particularly well suited for the purposes of this study as it seeks to extract words that are used by all members of the 'academic discourse community' (Swales 1990).

3.2. Methodology

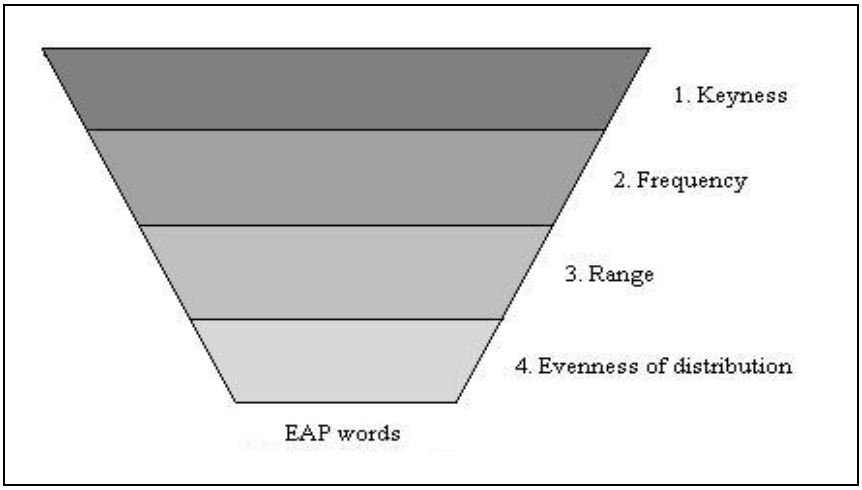
Coxhead (2000) selected word families for the AWL on the basis of three criteria:

1. Non-occurrence in the GSL,
2. Range: a frequency of at least 10 occurrences in each section of the corpus,
3. Frequency: a minimum frequency of 100 in the whole academic corpus.

The method proposed in this study is primarily based on keyness (cf. Scott 2001), a criterion that has not been used by Coxhead (2000). A series of quantitative filters involving frequency, range and evenness of distribution is subsequently used to narrow down the resulting list of EAP words (cf. Figure 1).

[1] The relatively small size of the corpus and the limited number of texts may be considered two limitations of this study. However, the selection procedure described in section 3.2 has recently been replicated on a larger corpus which represents a wide range of academic sub-genres, i.e. published articles, samples of books, student essays, etc. in different disciplines and the results are quite consistent with those presented here.

Figure 1: A four-layered sieve to extract EAP words



3.2.1. Keyness

The keyness method has been used in a variety of fields to extract distinctive words, e.g. business English words (Nelson 2000). Keywords are words that appear significantly more often in the corpus under study than in a comparison corpus. For the purpose of this research, the MC corpus was compared with a large corpus of fiction writing on the basis of the hypothesis that typical EAP words would be particularly under-represented in fiction. The comparison was made using the compare list option of *WordSmith Tools* (Scott 1999) which produces a list of all the words that present statistically significant differences in frequency between the two corpora.[2] It is based on lemmas and part-of-speech tags. Lemmas were chosen in preference to word forms as Granger and Paquot (2005) showed that, although the description of EAP words should be based on an analysis of key word forms, key lemmas give better results for the selection procedure. Part-of-speech tags made it possible to create lists of EAP nouns, verbs, adverbs, etc., and to distinguish between grammatical homographs, e.g. *use* as a noun or verb.

Used alone, keyness not only extracts highly frequent words in the academic corpus (e.g. *example, conclusion, view*) but also gives prominence to discipline or topic-related vocabulary, which is not necessarily frequent in the academic corpus but is typically under-represented in fiction writing (e.g. *bacterium, methane, DNA, penicillin, chromosome, enzyme, jurisdiction, rape, archbishop, martyr, rape,* etc.). The three criteria of frequency, range and evenness of distribution are subsequently used to refine the list of EAP keywords.

3.2.2. Frequency and range

Word families were selected for the AWL if they appeared at least 100 times in a 3,500,000 word corpus. The study presented here being based on lemmas in a 1,000,000 word corpus, for a keyword to be selected as an EAP word, its frequency should be equal to or higher than 30 occurrences. To distinguish frequent words that are liable to be found in most academic texts from others that are restricted to a specific discipline (e.g. *cell, gene, patient, treatment,* etc.), the criterion of range, i.e. frequency in terms of the number of texts, was also used. Range was calculated on the basis of the five subcorpora with the consistency analysis option of *WordSmith Tools*. Only words appearing in the five academic disciplines were retained as EAP words.

Used alone, range also has an important limitation: it gives no information on the frequency of a word in any text. Thus, the criterion of range dismisses the word *cell* as it only appears in two subcorpora but includes both the word *example*, which is intuitively regarded as an EAP word, and the lemma *law*, whose meaning is more discipline or topic-dependent (*the Canon Law, criminal law, the law of gravity*). Their sub-frequencies given in

[2] The significance of the statistical test was set at 0.01, which means that there is less than 1% danger of mistakenly claiming a significant difference in frequency.

table 1 confirm this distinction. The frequency of the word *example* ranges from 20 to 125, while that of the word *law* ranges from 6 to 334. The wider frequency range of *law* can be explained by the highly frequent occurrence of *law* in the social science sub-corpus.

Table 1: Distribution of the words *example* and *law* in the 5 sub-corpora

	Normalized frequencies / 100,000 words				
	Arts	Belief and religion	Applied science	Science	Social science
example	20	77	111	125	56
law	6	129	71	16	334

Differences such as these can be highlighted by a measure of the evenness of distribution of words in a corpus, the last criterion applied to further restrict the list of EAP words.

3.2.3. Evenness of distribution

The evenness of distribution or dispersion of a word is "a statistical coefficient of how evenly distributed a word is across successive sectors of the corpus" (Rayson 2003: 93). One such measure is Juilland's D statistical coefficient. Its values range from 0 to 1 and the closer a Juilland's D value for a word is to 1, the more equal is the spread of occurrences of that word across the sectors of the corpus.[3] For a word to be selected as an EAP word, its Juilland's D value should be greater than 0.5.[4] Thus, the noun *example* is selected as an EAP word as its dispersion value equals 0.73 whereas the noun *law*, with a Juilland's D value of 0.4, is not. Dispersion values make it possible to avoid the wrong conclusion that these two words behave in the same way in academic writing and reveal that only *example* is of widespread and general occurrence in this particular genre, while the noun *law* is over-represented in the social science sub-corpus.

4. A productively-oriented academic word list

4.1. Results of the automatic extraction

The successive application of the four criteria described in section 3 – keyness, frequency, range and evenness of distribution – automatically restricts the number of EAP words to 838. Table 2 gives the first 100 EAP words sorted on their keyness value, i.e. the preposition *of* is the EAP word that has the highest keyness in the MC corpus.

Table 2: First 100 EAP words

of, the, which, in, be, by, may, this, or, these, such, example, also, case, that, however, social, effect, development, its, form, principle, system, theory, between, term, develop, problem, result, per, will, individual, not, different, general, particular, century, change, thus, therefore, government, process, subject, require, an, as, produce, group, occur, period, establish, include, element, common, history, view, study, structure, involve, condition, many, use, chapter, reason, important, argument, cent, other, according, environment, rule, nature, issue, evidence, consequence, their, describe, base, action, likely, most, natural, idea, response, consider, early, role, source, interpretation, method, value, relation, person, similar, provide, during, policy, cause, apply.

[3] For more information on range and evenness of distribution, see Rayson (2003: 93-94).
[4] A Juilland's D value of 0.5 is arguably quite low; however, increasing this value would have resulted in a lower recall rate.

A high number of these words are typical of what would intuitively be regarded as EAP vocabulary, e.g. *example, principle, theory, develop, problem, result, view, study, reason, argument, issue, evidence, consequence, describe, source, interpretation, method, cause*, etc. Some words also stress the importance of considering multi-word sequences as well. For example, the lemma *term* almost always occurs in sequences such as *in general terms* or *in terms of* and *general* and *particular* often appear in the two-word sequences *in general* and *in particular*. Similarly, the two words *per* and *cent* are part of the lemma *per cent* and *according* does not stand alone but appears in the complex preposition *according to*.

4.2. EAP nouns

In this case study, we focus on EAP nouns as they have proved to be significantly underused in learner academic texts (cf. Granger and Rayson 1998; Hinkel 2002). Our list of EAP words initially contained 298 nouns, from which four categories of nouns were selected for their potential usefulness for productive purposes:

1. 'Shell nouns' are "abstract nouns that have, to varying degrees, the potential of being used as conceptual shells for complex, proposition-like pieces of information" (Schmid 2000:4). Thus, in the following sentence, the noun *argument* is a shell noun which stands for the underlined proposition:

   Sir Norman's **argument** was <u>that it should not be replaced but maintained</u> (ibid: 132).

   Examples of 'shell nouns' in the EAP list are *argument, consequence, favour, discussion, value, equivalent, example, suggestion, tendency, version, proposal, possibility*, etc. These nouns are particularly interesting as they constitute an important type of lexical cohesion (cf. Francis 1994) which has been shown to be pervasive in academic discourse and likely to be problematic for non-native, as well as native speakers, notably because they refer to abstract entities and introduce additional propositional density to a text (cf. Flowerdew 2003).
2. Quantifying collective nouns: *a **group** of, a **majority** of, a **set** of, a **series** of*, etc.
3. Species nouns: *a **kind** of, a **sort** of, a **type** of, a **form** of, a **class** of*, etc.
4. Metalinguistic text nouns (Francis 1994): *chapter, figure, article*, etc.

Examples of nouns that were not selected are *citizen, colleague, history, environment, education, work* and *welfare.*

A comparison of the final list of 237 EAP nouns with the *General Service List* and the *Academic Word List* shows that only one third of the nouns is shared with the AWL whereas two thirds are General Service words. Table 3 gives the distribution of EAP nouns in the GSL and the AWL together with examples.

Table 3: Distribution of EAP nouns in the GSL and the AWL

Lists	Number of EAP words	Examples
GSL	147 [62%]	act, addition, aim, argument, cause, change characteristic, choice, claim, comparison, difference, discussion, effect, example, exception, explanation, extent, fact, favour, group, idea, judgment, matter, measure, point, possibility, problem, question, reason, result, statement, study, subject, suggestion, term, variety, view.
AWL	83 [35%]	alternative, analysis, approach, aspect, assumption, chapter, conclusion, consequence, contrast, criterion, debate, definition, evidence, factor; implication, instance, interpretation, issue, method, proportion, research, summary, theme, theory, version
Off-list	7 [3%]	appeal, criticism, opposition, reference

These results highlight the important role played by General Service words in academic prose and justify their inclusion in a productively-oriented academic word list.

5. Fine-tuning the list: the contribution of learner corpus data

The value of a productively-oriented academic word list or of any other EAP pedagogical tool is greatly increased if findings from learner corpus data are also used to select the words that should be given special attention in EAP courses. By way of illustration, the frequencies of the EAP nouns described in section 4.2 are compared in a subpart of the *Louvain Corpus of Native Speaker Essays* (LOCNESS) which consists of 150,000 words of argumentative essays written by American university students and a comparable corpus of essays written by higher-intermediate to advanced French-speaking university students. The learner corpus, henceforth referred to as ICLE-FR, is a sub-part of the *International Corpus of Learner English*, a large corpus of English as a foreign language produced by learners from eleven different mother tongue backgrounds (Granger et al 2002). The comparison shows that while a restricted set of EAP nouns are massively overused by French learners (e.g. *action, difficulty, conclusion, development, example, importance, instance, possibility, problem, question,* etc.[5]), the majority of them are typically underused by this learner population (e.g. *argument, issue, claim, evidence, support, effect, type, research, debate, case, method, value, reason, approach, emphasis, consequence, alternative*, etc.). A very high number of overused and underused EAP nouns belong to West's GSL, which provides further justification for their inclusion in a productively-oriented academic word list.

Learner corpus data can also be used to fine-tune the description of the phraseology of EAP words by providing useful information on learners' difficulties in terms of underuse, overuse and misuse of multi-word sequences and use of nonnative-like sequences. French learners' use of the noun *example* is a case in point. A comparison of the use of the EAP noun *example* in native and French learner writing first shows that learners significantly overuse the conjunct *for example* and sometimes even use it to introduce what is definitely not an example:

a) *As far as the European culture is concerned, I could say that to a certain extent the unification of Europe is leading to a loss of identity.* ***For example****, the question of the language that will be spoken in the community arises, and many problems are still not solved.* (ICLE-FR)
b) *As soon as you practice or simply like music, novels, poetry, theatre, etc., you give a place for dreaming and imagination. Some people might then infer that* ***for example*** *children's imagination is threatened precisely because they sit most of the time passively in front of the television.* (ICLE-FR)

By comparison, French learners typically underuse the two productive frames with the verb *to be* commonly found in native academic writing, i.e. 'X *is a(n)* (adj.) *example of* Y' and '*a(n)* (adj.) *example of* Y *is* X', as illustrated in the following two sentences:

c) *However, the microwave uses radiation to excite water particles in food, thus creating friction, which creates heat. This* ***is a prime example of*** *thinking that does not follow in old footsteps but breaks away from convention and forges new routes.* (LOCNESS)
d) *Because the case against teaching New Ages ideas reaches more people, the argument is more powerful.* ***An example of this is*** *the magazine Christianity Today.* (LOCNESS)

Non-native tendencies to overuse the fixed formula *for example* (even when another phrase would be more appropriate) and underuse productive frames are clearly in line with Granger's (1998:156) conclusion that "learners' repertoires for introducing arguments and points of view are very restricted and they therefore 'cling on' to certain fixed phrases and expressions which they feel confident in using." An analysis of adjective and verb collocates also reveals learners' lack of feeling for what are native speakers' preferred ways of saying things. Learners use fewer typical collocations such as *set an example, cite an example, prime example* or *classic example* but tend to favour less typical or clearly erroneous combinations such as *a little example, an opposite example, state an example,* etc.

[5] It should be noted that the nouns *conclusion, example* and *instance* are not overused alone but in the multi-word sequences *for example, for instance* and *in conclusion*.

French learners repeatedly use the sequence *let's take the example of* (13% of the occurrences of the noun *example* in ICLE-FR), a frame that is not found in native formal writing. Academic writing is a genre characterized by high degrees of formality and detachment and the speech-like nature of this frame leads to an overall impression of stylistic inadequacy:

e) *One form of this is nationalism. To show what I mean,* ***let's take the example of*** *an Englishman in Belgium.* (ICLE-FR)
f) *One of them is the loss of contacts in families. (..)* ***Let us take an example:*** *many people eat while watching TV. I personally think that this is a pity.* (ICLE-FR)

Whereas the conjunct *for example* is significantly overused in most ICLE subcorpora (from 44% to 66% vs. 35% in LOCNESS), the frame *let's take the example of* is quite rare in other interlanguages and seems to be characteristic of French learners. This over-representation may well be L1-induced as the frame has a congruent counterpart in French, i.e. *Prenons l'exemple de* …, which is found in student and professional formal writing.[6] Similarly, Nesselhauf (2004) suggested that German learners' erroneous combination *X is an example* ***for*** was most probably modelled on German *ein Beispiel für*. For more examples of the potential influence of the mother tongue on phraseological patterns in EFL academic writing, see Granger and Paquot (2005) and Paquot (2005).

Descriptions of EAP vocabulary in native corpora are becoming available (cf. Verdaguer and González 2004) but information derived from learner corpora analysis, and more crucially, L1-specific information, is currently sorely lacking (cf. Granger 2004b). Although L1-specific findings may be more difficult to incorporate into generic EAP textbooks, patterns shared by all learner populations could be used to increase the pedagogical value of these materials: lexical means found to be typically underused by all learner populations could be presented as useful alternatives to learners' preferred patterns. Thus, instead of presenting *for example*, 'X *is a* (adj.) *example of* Y' and '*a* (adj.) *example of* Y *is* X' as interchangeable means to introduce an example, EAP textbooks could draw learners' attention to the fact that EFL students generally tend to overuse *for example* and advise them to use the productive frames as well.

6. Conclusion

This study highlights the importance of a rigorous and empirically based selection of vocabulary and presents a new methodology for the identification of lexical items that should be part and parcel of a productively-oriented academic word list. The results suggest that while a distinction between General Service words and Coxhead's academic words is valuable for receptive purposes, it is not totally adequate for productive purposes. Numerous General Service words have important discourse functions in EAP and their phraseological uses are not fully mastered by L2 learners, even at an advanced level. Learners need pedagogical tools to help them use these words to produce accurate and stylistically appropriate language. The trend towards more productive tools has already been highlighted by Rundell (1998) in the field of pedagogical lexicography. The present paper shows that although foreign language teaching, and more especially EAP, has already started its own 'productive revolution', much remains to be done. It also brings out the high value of insights gained from learner corpus data in the design of productive tools for learners and supports Flowerdew's (1998) suggestion that EAP textbooks should pay more attention to aspects of overuse, underuse, misuse and learners' idiosyncratic use of lexico-grammatical means to fulfil rhetorical functions. Finally, a comparison of the phraseology of the noun *example* in several interlanguages reveals that if a number of features are shared by most learner populations and are therefore more likely to be developmental or teaching-induced, other aspects of learner phraseological use in academic writing are characteristic of one mother tongue background and therefore probably L1-dependent. In contrastive rhetoric, L1 influence has already been shown to manifest itself "in the writer's choice of rhetorical strategies and content" (Connor 2002:494). These findings have important implications for EAP teaching as they suggest that the mother tongue deserves a place in the academic writing class.

[6] First person plurals of the present imperative and indicative are much more frequent in French formal writing.

Acknowledgements

I gratefully acknowledge the support of the Communauté française de Belgique, which funded this research within the framework of the 'Action de recherche concertée' project entitled 'Foreign Language Learning: Phraseology and Discourse' (No. 03/08-301). I would also like to express my deep gratitude to Sylviane Granger for her guidance and support.

REFERENCES

Biber, D. (2004). "Lexical bundles in academic speech and writing." In: Lewandowska-Tomaszczyk, B. (ed.). *Practical Applications in Language and Computers* (PALC 2003). Frankfurt am Main: Peter Lang, 165-178.

Biber D., Johansson, S., Leech, G., Conrad, S. and E. Finegan (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Connor, U. (2002). New directions in contrastive rhetoric. *TESOL Quarterly* 36(4): 493-510.

Cowie, A.P. (1997). "Phraseology in formal academic prose." In: Aarts, J., de Mönnink, I. and H. Wekker (eds.). Studies in English Language and Teaching In Honour of Flor Aarts. Amsterdam/Atlanta: Rodopi, 43-56.

Coxhead, A. (2000). "A new Academic Word List." *TESOL Quarterly* 34 (2): 213-238.

Flowerdew, J. (2003). "Signalling nouns in discourse." *English for Specific Purposes* 22: 329-346.

Flowerdew, L. (1998). "Integrating 'Expert' and 'Interlanguage' Computer Corpora Findings on Causality: Discoveries for Teachers and Students." *English for Specific Purposes* 17(4): 329-345.

Francis, G. (1994). "Labelling discourse: an aspect of nominal-group lexical cohesion." In: Coulthard, M. (ed.). *Advances in written text analysis*. London/New-York: Routledge, 82-101.

Granger, S. (1998). "Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae." In: Cowie, A. (ed.). *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, 145-160.

Granger, S. (2004a). "Practical Applications of Learner Corpora." In: Lewandowska-Tomaszczyk, B. (ed.). *Practical Applications in Language and Computers (PALC 2003)*. Frankfurt: Peter Lang, 291-301.

Granger S. (2004b) "Computer learner corpus research: current status and future prospects." In: Connor U. and T. Upton (eds). *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam/Atlanta: Rodopi, 123-145.

Granger S. and P. Rayson (1998). "Automatic profiling of learner texts." In: Granger, S. (ed.). *Learner English on Computer*. Harlow: Longman, 119-131.

Granger, S., Dagneaux, E. and F. Meunier (eds) (2002). The International Corpus of Learner English. CD-ROM and Handbook. Louvain-la-Neuve: Presses universitaires de Louvain. Available from http://www.i6doc.com

Granger, S. and M. Paquot (2005). "The phraseology of EFL academic writing: Methodological issues and research findings." Paper presented at ICAME 26 – AAAACL6 (International Computer Archive of Modern and Medieval English - American Association of Applied Corpus Linguistics), 12-15 May 2005, University of Michigan, USA.

Hinkel, E. (2002). Second Language Writers' Text. Linguistic and Rhetorical Features. London: Lawrence Erlbaum Associates.

Lewis, M. (2002). *Implementing the Lexical Approach. Putting Theory into Practice*. Heinle: Boston.

Nation, P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nelson, M. (2000). *A Corpus-Based Study of Business English and Business English Teaching Materials*. Unpublished PhD Thesis. Manchester: University of Manchester.

Nesselhauf, N. (2004). "Learner corpora and their potential for language teaching." In: Sinclair, J. (ed.). *How to use corpora in language teaching*. Amsterdam/ Philadelphia: Benjamins, 125-152.

Oakey, D. (2002). "Formulaic language in English academic writing: A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines." In: Reppen, R., Fitzmaurice, S.M. and D. Biber (eds.). *Using Corpora to Explore Linguistic Variation*. Amsterdam/Philadelphia: Longman, 111–129.

Paquot, M. (2005). "EAP vocabulary in learner corpora: A cross-linguistic perspective." Paper to be presented at *Phraseology 2005: The Many Faces of Phraseology*, 13-15 October 2005, Louvain-la-Neuve, Belgium.

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Unpublished PhD dissertation, Lancaster University.

Rundell, M. (1998). "Recent trends in English pedagogical lexicography." *International Journal of Lexicography* 11(4): 315-342.

Scott, M. (1999). *WordSmith Tools version 3*. Oxford University Press.

Scott, M. (2001). "Comparing corpora and identifying key words, collocations, frequency distributions through the WordSmith Tools suite of computer programs." In: Ghadessy, M., Henry, A. and L. Roseberry (eds.). *Small Corpus Studies and ELT.* Amsterdam: Benjamins, 47-67.

Scott, M. and T. Johns (1993). *MicroConcord*. Oxford: Oxford English Software.

Schmid, H. (2000). *English Abstract Nouns as Conceptual Shells: From Corpus to Cognition*. Berlin/New York: Mouton de Gruyter.

Schmitt, N. (ed.) (2004). *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: Benjamins.

Swales, John M. (1990) *Genre analysis: English in academic and research* settings. Cambridge: Cambridge University Press.

Thurstun, J. & C. Candlin (1997). *Exploring Academic English: A workbook for student essay writing*. Sydney: Macquarie University Press.

Verdaguer, I. and E. González (2004). "A Lexical Database of Collocations in Scientific English: Preliminary Considerations." In: Williams, G. and S. Vesssier (eds.). *Proceedings of the Eleventh EURALEX International Congress*. Lorient: Université de Bretagne-Sud, 29-934.

West, M. (1953). *A General Service List of English Words*. London: Longman.

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

Keywords: English for Academic Purposes (EAP), EAP vocabulary, academic word list, methodology, phraseology, learner corpora, multi-word units, teaching materials, reception vs. production