Available online at www.sciencedirect.com

ScienceDirect

English for Specific Purposes 27 (2008) 459–479

English for Specific Purposes

www.elsevier.com/locate/esp

Through the looking glass and into the land of lexico-grammar

Nilgün Hancioğlu[a], Steven Neufeld[b,*], John Eldridge[a]

[a] Eastern Mediterranean University, Famagusta (Gazimağusa), North Cyprus, via Mersin 10, Turkey
[b] The English Language Consultancy Association, P.O. Box 29, Lapta-Girne, North Cyprus, via Mersin 10, Turkey

Abstract

This article describes two complementary research projects into lexical patterning and frequency in general and academic English. The research suggests that treating current popularly used wordlists such as the General Service List (GSL) and the Academic Word List (AWL) as distinct constructs is of questionable merit. Rather, there are strong arguments for revising general lists of word frequency in order to ensure maximum utility to any language learner, regardless of specialization. In this respect, the construction of a new general list of word families is described. The article then proceeds to illustrate the difficulties involved in isolating specifically academic lexis and describes corpus-informed research which strongly indicates that what ESP practitioners require in addition to general frequency lists are complementary banks of lexico-structural items and collocates with genre-specific attributes and functions. Consideration of this data then leads to conclusions concerning the types of lexico-grammatical elements that might best serve different types of learners, and also the kind of methodology that might be appropriate, particularly for those involved in the teaching of English for general and specific academic purposes.

1. Introduction

Although the use of wordlists in language teaching has a long pedigree, it is only recently that they have again taken centre stage. Ellis (2002, p. 143), perhaps succumbs

[*] Corresponding author. Tel.: +90 533 849 7553 (mobile); +90 392 825 2403 (home).
E-mail addresses: nilgun.hancioglu@emu.edu.tr (N. Hancioğlu), steven.neufeld@gmail.com (S. Neufeld), john.eldridge@emu.edu.tr (J. Eldridge).

0889-4906/$34.00 © 2008 The American University. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.esp.2008.08.001

to hyperbole when referring to "40 years of exile" for frequency profiling, but his point is succinct enough. For many years, West's (1953) General Service List (GSL), designed to be a maximally useful list of word families, remained sidelined. Vocabulary teaching in general fared little better, and questions such as how much lexis learners already needed to know in order to infer meaning remained mostly unaddressed. Research suggesting, for example, that readers need to recognize 95% of the words in a text for the text to be of instructional use (Liu & Nation, 1985) was notable mostly for its novelty value.

The computer revolution however returned lexis to the limelight, and extensive analysis of corpora proved that West's (1953) efforts to establish a base vocabulary list had been remarkably successful. Very simply, the majority of text seemed to be compiled from a limited set of frequently used items (Nation & Waring, 2004). It was not long either before more specialized lists were developed. Perhaps the best known of these is Coxhead's (2000) Academic Word List (AWL), which consists of 570 word families, additional to the GSL, that occur with the greatest frequency across a corpus of academic texts.

2. The world of wordlists

Unfortunately, wordlist driven approaches are neither problem nor value-free (Folse, 2004). For, as Nation (2001, p. 23), points out, "words are not isolated units of language". Rather, lexico-knowledge is a complex phenomenon involving multiple interlocking systems and levels. Knowledge of a word can range from surface recognition only, to detailed knowledge of forms, derivations, synonyms, antonyms, hyponyms, collocations etc. Indeed, the more one looks at what knowledge of words involves, the more it becomes apparent that knowing any word in depth involves knowing *other* words, and quite a lot of them. Hence, the teaching and learning economy offered by wordlists alone is suspect, unless the purpose of instruction is to provide a superficial recognition of common lexical items for comprehension purposes only.

Discrete item frequency lists, though extremely useful in defining learning targets, are also by nature unrevealing of the subtleties of lexical phrases, multi-word units, and pre-formulated chunks. We may learn, for example, that *take* is a frequent item, but without reference to manifestations in the shape of *take off*, *take over*, *take into consideration*, and so on (Hancioğlu & Eldridge, 2007). Polysemantic items such as *bow* and *row* present another problematic case, as do primary and metaphorical usages, as in *driving into a ditch* and to *ditch one's girlfriend.*

The question of word forms and families meanwhile was addressed by Bauer & Nation (1993) who defined seven levels of family relationship. Strict application of these levels yields consistency in research, but not necessarily flexibility in pedagogy. For example, *courageous* is seen as a derived form of *courage*, but *famous* is not treated as a form of *fame*, the dropping of the final 'e' disqualifying it. Reference words are another contentious area since in texts containing numerous repetitions of referential items such as *it*, what is actually of concern is what is being referred to in each case and whether that is a known item or not, rather than the repetition of the reference item itself (Hancioğlu & Eldridge, 2007). In short, as Gardner's (2007) in-depth discussion of how to validate the construct of a word emphasizes, fundamental questions remain unresolved concerning what exactly is being counted in frequency-listing operations and why, and how precisely subsequent categorizing operations are being conducted.

For all these problems however, wordlists and corpora unquestionably offer a portal into the complex behaviour and intricate relationships of individual lexical items. For many learners, this is territory which has remained relatively unexplored. Issues of breadth and depth become even more critical when we consider productive skills. Good writing, for example, requires learners to develop textual cohesion through the delicate use of such lexically related techniques as synonymy, antonymy and hyponymy. It has also been argued that a key measure of proficient writing is its appropriate use of *lower frequency words* (Laufer & Nation, 1995). An additional concern thus surfaces, which is that a focus on high frequency words may in fact condemn learners to disfluency, and that instructional time may be sacrificed on items *most likely* to be learned through natural encounter. Krashen (2003) meanwhile reasserts the value of extensive reading programmes as the best foundation for natural vocabulary acquisition. In response, Cobb (2007b) describes research suggesting that such programmes do not have the capacity to promote effective vocabulary development in an ESL context. Again, much would seem to depend on what we mean by ‘word knowledge’. If comprehension-survival principles are paramount, practitioners may indeed want to consider whether higher frequency words will be naturally acquired through the methods Krashen describes. However, particularly in non-English speaking environments, it would be wise not to take too much for granted. In this regard, the following sections will examine, respectively, the General Service List (West, 1953) and the Academic Word List (Coxhead, 2000), not only because they are two of the most commonly used wordlists, but because for ESP practitioners, they offer a package more or less suggesting that the basis for survival in an academic environment is knowledge of the 2000 word families of the GSL plus the 570 word families of the AWL.

3. Standing up to the test of time: the General Service List

The General Service List (GSL) was developed by Michael West as a list of the 2000 most useful ‘general service’ headwords and families for English language learners. West used a written corpus of 5,000,000 words as the basis for his list, which was compiled not only according to frequency, but also on such criteria as range, learning ease, coverage, necessity and style (Nation & Waring, 2004).

Given the age of the GSL, it is not surprising that researchers have since expressed doubts about its composition and relevance. Engels (1968) agreed that the first 1000 words were good selections, but felt that many of the words beyond this could not be considered “general service words”, since their range and frequency were too low. The results of a subsequent study, detailed by Hwang and Nation (1995), supported Engels’ view that the lower frequency words needed some revisiting (see also Chujo & Utiyama, 2005). Nonetheless, the GSL is still used, most notably in the *Classic Vocabulary Profiler* (Cobb, 2006) available on the *Compleat Lexical Tutor Site* (Cobb, n.d.), and provides around 80% coverage of most written texts, which is remarkable considering the list evolved over several decades before its publication in 1953. By way of illustration, it may be noted that 82.84% of this article to this point is comprised of vocabulary from West’s list. This is an impressive return, and suggests that whilst the GSL might need restorative work, its foundations remain remarkably intact.

The following tables, however, illustrate some of the problems in the GSL (in the version published in Cobb’s Lextutor site) that were identified during the course of this research, using the classic vocabulary profiler (Cobb, 2006) (see Tables 1 and 2).

Table 1
Inconsistencies in the GSL

GSL problem areas:	Examples
US/UK spelling	*Litre* is in the GSL, but *Liter* is not
Word forms	*Rise* is in the GSL, but *risen* is not. Similarly, *hope* is in the GSL, but *hopefully* is not. *Motherhood* is in, but *fatherhood* is not. *Tour* and *tourist* are in, but *tourism* is not
Singular/plural	*Strength* is in the GSL. But not *strengths. Keepers* but not *keeper*
Archaic words	*Shilling* is in the GSL
Not up-to-date	*Radio* is in the GSL, but not *television, video, plastic, aircraft, airport, airlines, etc*

Table 2
Word family issues related to limitations of the vocabulary profiling tools and Bauer and Nation's (1993) levels

GSL problem areas:	Examples
Word families	*Pride* and *Proud* are treated as separate headwords.
Words that are spelled the same, but have completely unrelated meanings	*Canned* (tin) derives from *Can* (ability) *Saw* (the tool), and *Saw* (the past tense of see) are treated as one and the same.
Difficulties in deciding when prefixes indicate a separate word family	*Force* is in the *GSL*, but not *enforce*, which occurs in the AWL. Similarly, is *alive* a derivant of *live*? *Awake* of *wake*? Not according to the GSL.
Derivation or synonym?	On the other hand, *'Dad'* is listed as a derived form of *'Father'*.

As powerful then as the GSL may appear, it still contains residual problems. Thus, if subjected to various refining processes, a stronger list still should result. A fundamental objective of the project described here was to achieve just this.

4. The Academic Word List

When Coxhead's Academic Word List (2000) is used as an additional frequency band to the GSL with which to analyse this article, the resulting profile covers 92.71% of the tokens (number of words) used to this point. This too is an impressive outcome. In order to achieve it, Coxhead (2000) compiled a corpus of 3.5 million words from four broad academic subject groupings of arts, commerce, law, and science and then highlighted the 570 most common word families that occurred according to both frequency and range of occurrence throughout the corpus, excluding GSL words.

In adopting this procedure however, Coxhead was very much relying on the soundness of the GSL. Any flaws in the GSL were likely to be accentuated in the AWL, and the assumption that any high frequency word outside the GSL coverage in the academic corpus would be a de facto academic item perhaps accounts for the distinctly 'un-academic' texture of some of the items on the list.

Hyland and Tse (2007) criticize Coxhead for her approach to data collection, and argue that as a result, the AWL is biased towards certain fields, such as law and economics. Hyland and Tse further question the very notion of a "single academic literacy", and suggest that a solution to the deficiencies of the AWL lie in building more specialized lists, from field and genre-specific corpora.

Table 3 gives some examples of how AWL items seem to derive not only from general academia, but also from more specific academic fields, as noted by Hyland and Tse, and from general (non-academic) English, crucially however, what is evident from such intuitive categorizations is that there are in fact few items in any category that cannot be used quite freely outside academia. What makes text ‘academic’, then, is not the occurrence in isolation of certain specific items, but the ways in which certain items ‘collocate’ and ‘colligate’, in other words, the ways lexical items co-occur with other lexical and grammatical items (Hunston, 2002, pp. 12–13). To take just one example, the word *drama*, which is an AWL headword, actually behaves in a remarkably ‘un-academic’ fashion if a large corpus such as the Cobuild Bank of English is consulted (HarperCollins Publishers, 2004), common collocates including such unsurprising items as *school*, *series*, and *TV*. It is the derived form *dramatic* that lends *drama* its academic prominence through its co-occurrences with such items as *increase*, *decrease*, and *effect*.

The fact then that items such as *study* appear in the GSL (but not the AWL) and items such as *drama* in the AWL (but not the GSL), suggests that the division of vocabulary into mutually exclusive lists is likely to be an activity that for all its initial convenience may prove inherently problematic in the longer run. As Gilguin, Granger, and Paquot (2007, p. 324), caution, “Coxhead’s (2000) Academic Word List does not include the 2000 most common English words, with which non-native writers may still have considerable difficulties, especially in cases where their use in academic writing differs from their habitual use”. Paquot (2007) further questions the common practice of selecting EAP vocabulary based on words not appearing in the GSL.

Not only do distinctions need to be drawn between academic and ‘un-academic’ behaviour of individual words and word families, some thought should also be given to pedagogical issues. The work of Hyland and Tse (2007), for example, in identifying disciplinary variation, whilst obviously of significance, needs also to be balanced against the views of commentators such as McCarthy and O’Dell (2008, p. 6), who advise students that “specialist terms are often relatively easy to master – they will be explained and taught as you study the subject”, and furthermore “it is the more general words used for discussing ideas and research and for talking and writing about academic work that you need to be fully familiar with in order to feel comfortable in an academic environment”. Paquot et al. (2007) meanwhile points out that the AWL may be of greater value as a resource for receptive rather than productive teaching purposes, and argues for a more phraseological approach to productive teaching, a prescription supported more recently by Coxhead herself (2008).

Furthermore, not only are many academic programmes interdisciplinary, language support classes may often comprise groups of students from different disciplines (Eldridge, 2008). For many students, academic life may also encompass a range of other demands. The forthcoming Pearson Test of English, for example, is intended for “non-native speak-

Table 3
Examples of range of items in the Academic Word List

‘General’ items	‘Economics’ items	‘Academic’ items
adult bulk couple drama	currency corporate credit	criteria debate define
injure job odd sex	export economy estate	emphasis evaluate
somewhat tape	finance fund invest levy	illustrate interpret method

ers of English ... who want to study at institutions where English is the principal language of instruction" (Pearson Language Tests, n.d.). The test has been designed using an academic corpus compiled from not only a range of academic disciplines but also taking into account ''administrative and extra-curricular" needs (De Jong & Ackermann, 2008), a reminder that university students also need to cope with such matters as registration, rentals, and social adaptation, as well as formal study.

Meanwhile, Chung and Nation (2003, p. 1), conclude that the AWL basically comprises vocabulary that is common across a range of different academic fields, but that cannot be termed *technical* since it is "not typically associated with one field". Although such vocabulary has been termed *sub-technical* (Cowan, 1974) or *semi-technical* (Farrell, 1990), they suggest, finally, that the word families in the AWL are actually "more closely related to high frequency vocabulary than to technical vocabulary" (Chung & Nation, 2003, p. 1).

It is true that the AWL has proved quite successful in isolating vocabulary that is commonly used in academic discourse (Mudraya, 2006), albeit with inevitable disciplinary variations (see, for example Chen & Ge, 2007). Given however the difficulties in isolating a specific set of academic word families and clearly distinguishing them from general or non-academic English word families, logic would seem to dictate that a sensible step forward might well be to compile an expanded version of the GSL that reflects this. In this regard, it is telling perhaps that of the first one hundred items that Mudraya (2006) isolates as *academic* items in terms of engineering English, a full ninety-nine appear in the British National Corpus list of the 3000 most frequent words in English, as do 19 of the 20 most significant academic terms isolated by Chen and Ge (2007). Such overlaps are clear reason to avoid taking the GSL as any kind of 'given' in the compilation of more specialized wordlists, and suggest that a wiser approach might be to compile and work from wordlists that complement each other rather than constitute a mutually excluding sequence certain to produce anomalies and uncertainties.

5. Defining genres

As already noted, Hyland & Tse (2007) propose that field and genre provide a firmer foundation for corpus-informed work than the flawed concept of a "single academic literacy". It is important in this regard to understand what is meant by genre. Swales (1990) provides a comprehensive definition: "A genre comprises a class of communicative events, the members of which share some set of communicative purposes" (p. 58). Genres are composed of units of purpose, called *moves* or *move structures* (Swales, 1990), some of which are compulsory and some optional (Flowerdew, 2000; Halliday & Hasan, 1985). These constituent parts, or moves, represent the writer's communicative purpose (Flowerdew, 2000) and perform specific functions (Bhatia, 1993, cited in Henry, 2007). It should be noted however that Swales (2004) himself states that his 1990 definition of genres was "long and bold". Such definitional depictions, he cautions, may not be true "in all possible worlds and all possible times" (p. 61).

Swales (2004) stresses the importance of moves as functional units and suggests that a move "is better seen as flexible in terms of its linguistic realization" (p. 229). Henry (2007) emphasizes the significance of the lexico-structural features employed to fulfill moves; Flowerdew (2000) draws attention to key lexical phrases "representative of the move structures" and Tardy (Johns et al., 2006) also emphasizes the importance of lexico-structural features in generic moves. Moves can be realized in a number of ways, each of which is called a *strategy* or a *tactic* (Henry, 2007). Henry also stresses that each strategy has its own lexico-structural features that need to be identified. Chan and Foo (2001) meanwhile emphasize lexico-structural accuracy, organization and structure, and real world practices as major pedagogical concerns in a genre-analytic approach.

Academic texts may belong to different genres. As Swales (1990) points out, genres can share common features in terms of purpose, target audience, structure, style, and content. He gives research articles, research presentations, grant proposals, theses and dissertations, reprint requests, and abstracts as instances of academic genres. Specific parts of a genre, such as abstract, introduction, discussion and literature review sections of research articles and dissertations, have also been termed "part-genres" (Flowerdew, 2000; Hyland, 2005).

Much recent work (Charles, 2007; Hyland, 2008; Lee & Swales, 2006) has successfully focused on and identified linguistic variation between disciplines. Genres, however, cut across subject fields, and moves and many of their lexico-structural realizations are defined not only by specific subject matter (e.g. architecture), but by the conventions of the given genre (e.g. abstracts). Thus any suggestion that EAP should concentrate on compiling field-specific corpora and wordlists should be treated with care. As Hunston points out, "for many writers who are expert in their own field, . . .it is not the technical terminology, but what might be called the terminology of rhetoric that causes problems" (2002, p. 135). For many EAP practitioners dealing with classes of students from different academic fields, lists and banks that identify such generic lexico-structural commonalities are likely to remain of great value and utility.

Such discussions should also plainly maintain some contextual sensitivity. Differences in entry levels of learners can be very marked, and it is likely that learners who have not mastered the general lexico-structural building blocks of the language will struggle with either general or specific academic English courses. In other words, in addition to disciplinary variation, practitioners will equally need to contemplate learner and learning variation. In short, practitioners should be alert to the danger that discarding the AWL for more *specialized* lists may in fact result for some learners in the deconstruction of a critical section of the scaffolding of general English.

6. Research aims

The research studies described in this paper employ a "corpus-informed approach" (McCarthy, 2001). This approach allows the applied linguist to "mediate the corpus, design it from the very outset and build it with applied linguistic questions in mind, ask of it the questions applied linguists want answers to, and filter its output, use it as a guide or tool for what you, the teacher, want to achieve" (p. 129). With this in mind, the research described in the remainder of this article uses a corpus-informed approach with the aim of:

(i) Reviewing and revising the General Service List in a principled and effective way.
(ii) Reconceptualising the Academic Word List by integrating it into a revised General Service List.
(iii) Creating a model for lexico-structural banks for use in language teaching, and specifically thesis writing to non-native learners at post-graduate level.

7. Developing the Billuroğlu-Neufeld-List (BNL)

Since there has been a lot of activity recently in corpus-based, corpus-driven, and corpus-informed linguistics, a number of lists of commonly used words exist. In order to judge what kind of refinements the GSL might benefit from, Billuroğlu and Neufeld (2005) applied a simple method. All words from a basket of commonly used word lists were combined into one list, and filtered to obtain only the unique terms. The 4500 words in the resulting list were then ranked according to how many of the lists they occurred in, and controlled for homographs. The lists used were the GSL headwords and word family members (Dickins, Extended version of a General Service List of English words; Lextutor, 1000 families; Lextutor, 2000 families), the AWL headwords (Lextutor, AWL headwords) and most frequently occurring word family member (Lextutor, AWL sublists), on the basis that these were the most likely words to appear on the other lists, the first 2000 words of the Brown corpus (Edict, The first 2000 most frequent words from the Brown Corpus), the first 5000 words of the British National Corpus (Kilgarriff, Lemmatized BNC frequency lists), the revised version of the GSL (Bauman, About the GSL), the Longman Wordwise commonly used words (Longman, 2003), and the Longman Defining Vocabulary (Kennaway, The Longman defining vocabulary).

This approach was based on some simple premises. Although inconsistencies and overlaps seemed to occur in and between both the GSL and the AWL, their combined generative power suggested that there was a lot more right in the lists than there was wrong. Secondly, as shown in Table 4 below, there were clear indications that if the GSL was enlarged by even a relatively small degree, that much of the AWL would be absorbed into it. The methodology was thus based on accentuating the positive rather than the negative in prevailing lists and assuming provisionally that the combination of extant lists based on a number of extremely extensive corpora would naturally highlight common findings and isolate more singular and therefore questionable outcomes. The basic coverage that emerged from this combination of lists is given in Fig. 1 and Table 4.

The data show that the largest of the wordlists, the 5000 most frequent words of English as identified by the British National Corpus, encompasses a vast majority of the items in the other lists, including the Academic Word List, more evidence that the AWL is not necessarily as *academic* as has been assumed.

The classic vocabulary profiler on the Lextutor web site is based on dividing the GSL into two uniform frequency bands of approximately 1000 word families each (commonly

Table 4
Vocabulary profile of AWL and GSL in the first five thousand-word bands of the BNC

BNC frequency bands	Families		Band coverage%		Cumulative%	
	AWL	GSL	AWL	GSL	AWL	GSL
K1 words	91	864	15.04	47.36	15.04	47.36
K2 words	206	580	37.99	27.45	53.03	74.81
K3 words	96	345	14.21	14.83	67.24	89.64
K4 words	111	154	14.88	5.70	82.12	95.34
K5 words	75	64	9.61	2.36	91.73	97.70

Note: The British National Corpus (BNC) (BNC Consortium., 2005) consists of 100 million words collected from samples of written and spoken language from a range of sources, representing a wide cross-section of British English from the later part of the 20th century.

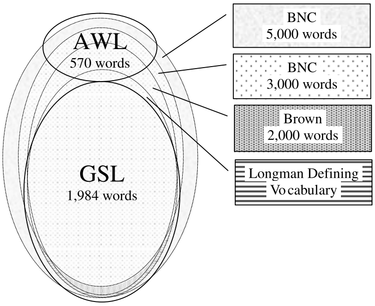

Fig. 1. VENN diagram of lists of commonly used words, showing areas of convergence.

referred to as K1 and K2), with the third frequency band being the AWL given as one list of 570 word families (arranged in uniform sublists of 60 words each, with 30 left over in the last). In contrast to the policy of mutual exclusivity followed in the GSL and AWL, Billuroğlu and Neufeld (2005), having combined all the wordlists noted above, then categorized the words according to the number of lists in which they were represented, producing a much finer resolution in the form of six distinct, ranked bands. The list was named the Billuroğlu–Neufeld list (BNL) with each frequency band assigned a BNL number, starting with '1' for the most frequent.

Table 5, illustrates how the bands were formulated, not as categories of arbitrary mathematical convenience but as bands that represented a genuine approximation of the natural vocabulary distribution of English texts. Hence Band 1 reflects the extremely high frequency of the top 765 word families, Band 2 the relative lesser frequency of the next 505 words, and so on. The reason for this departure from the convention of categorizing wordlists into even blocks was to provide users with information about differences in frequency between sublists, which can be quite significant, and potentially of pedagogical relevance Table 6.

As Table 5 additionally shows, the process also led to the emergence of a new set of 176 commonly used words that were outside both the GSL and the AWL.

The composition of the BNL ranking bands suggested that the GSL was not in need of any major surgery. The bands further reinforced the principle that commonly used words tend to continue to be used commonly over time. Also confirmed to a great extent were Engels' concerns about the range and frequency of items in the K2 band. Almost 100

Table 5
Breakdown of component constituents of the BNL, illustrating how the BNL bands approximate the natural vocabulary profile of English texts

BNL ranking	From K1	From K2	From AWL	Newly added	Subtotals
One	642	98	24	1	765
Two	192	274	38	1	505
Three	77	254	105	3	439
Four	46	212	145	26	429
Five	20	138	203	29	390
Six	2	8	55	116	181
Subtotals	979	984	570	176	2709

Table 6
GSL/AWL profile of key word examples from Hancioğlu's Target Abstract Corpus (TAC)

GSL (K1) (46)	GSL (K2) (13)	AWL (85)	Off-lıst (21)
apply, base, build, case, change, character, consider, describe, develop	aim, collect, combine, compare, critic, discuss, examining, explore, govern, improve	analyse, design, research, project, process, culture, thesis, construct, theory, environment	dissertation, interview, objectives, organisational, collaborate, correlate, quantitative, reform

of the K2 words ranked in BNL1, whilst the rest were quite evenly spread between BNL2 and BNL5. Finally, the bands that emerged indicated that the words in the AWL were not suitable only for academia, but included a large number of extremely commonly used words.

A medium-sized academic corpus of 730,000 words (Wordpilot., Academic Reports) served as a test to compare BNL and GSL/AWL text coverage. The text sampling in this corpus meets the criteria established by Chujo & Utiyama (2005). Topics varied from agriculture to volcanoes, pollution to economics, computers to political science and education to human factors in air traffic control, with texts derived from such varied sources as Harvard University, the US Securities and Exchange Commission, and the USGS Cascades Volcano Observatory.

Using RANGE (Heatley, Nation, & Coxhead, 2002), a vocabulary profiling software that provides information about frequency and occurrence across a range of two or more texts, the corpus was processed using the GSL/AWL lists as given in the Compleat Lexical Tutor site (Cobb, n.d.), and with the BNL, again using RANGE. The contrasting vocabulary profile in Fig. 2 below not only clearly reveals Engels' concerns about the coverage of K2 words, but also illustrates the natural distribution of words in an English text, which was used as the basis for producing and refining the banding system in the BNL.

In short, the new list of 2709 word families (Billuroğlu & Neufeld, 2007), using a basket of commonly used word lists to produce an improved unified perspective on commonly used words, better reflected the natural vocabulary profile of written texts, and also provided a more economic and meaningful approach to managing vocabulary development.

More importantly for the purpose of this discussion, the outcomes indicated that EAP practitioners should seriously consider putting aside the idea of a distinct discrete-item Academic Word List and instead focus on revisiting and recycling the most commonly used words in order to unravel the contexts, varied meanings, register, etc., that would help turn these words into powerful tools of understanding and expression. To give a brief illustration, if the text of this article to this point is profiled, the reader might note without any particular surprise, since this is an academic text, the frequent use of such sub-technical items such as *study, consider, describe, form, relationship, appear*. Yet, none of these items appears in the Academic Word List and therefore any practitioner choosing to teach from the AWL would run the risk of overlooking such items in academic English simply because they happen to have made a prior appearance in the GSL. In short, and rather ironically, the AWL excludes word families not just for the obvious reason that that they are not frequent enough in academic texts, but actually because they are *too frequent*. The argument made here is that the easiest solution to this problem is to avoid making the distinction between GSL and AWL in the first place.

Interested readers can view the BNL lists in full in a WIKI format (BNL, 2007) as well as test out the validity of these recommendations for themselves by visiting and using the

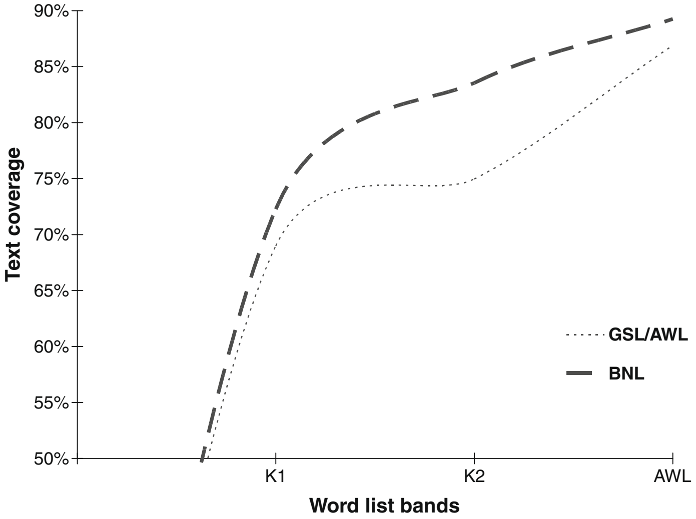

Fig. 2. Line graph of text coverage of a 730,000 word academic corpus, contrasting the natural distribution of words given by BNL to the GSL/AWL, highlighting the Engels K2 deficiency.

BNL version of the Vocabulary Profiler (Cobb, 2007a), and comparing results with the classic GSL and AWL version (Cobb, 2006). In the case of this article, analysed to this point (assuming world knowledge of proper nouns, abbreviations and acronyms), 92.93% is covered by the classic GSL and AWL combination, and 93.63% by the BNL, a marginal difference perhaps, but indicative of the main point – that the separation of lexis into general and academic is not necessarily as useful as might be thought.

8. Academic English and wordlists

As noted earlier, Hyland and Tse (2007) suggest that researchers should focus on developing and exploiting corpora from specific fields and genres. Taking the genre of thesis abstract, Hancioğlu did precisely this, compiling two corpora of thesis abstracts, (target and learner) and using the RANGE software and CONCORDANCE (Watt, 1999) for the purpose of analysis. The abstracts were taken from theses in the fields of Arts and Humanities, Sciences (including Engineering), Social Sciences, and Architecture and reflected the subject disciplines of postgraduate students taking Hancioğlu's advanced thesis writing course. During the course, participants are exposed to authentic samples of sections of a thesis, and after analysing them in terms of move structures and language, they then produce and develop their own work. Hancioğlu's aim was purely pedagogic, and therefore the corpora quite small. The target abstract corpus (TAC) comprised 174,093 running words of text compiled from 600 abstracts, with each of the four fields being represented equally. The learner abstract corpus (LAC) meanwhile was comprised of work by Hancioğlu's own students, and therefore necessarily smaller, totalling 21,575 running words from 100 abstracts compiled over six academic semesters.

Mudraya (2006) advocates the use of small corpora for language learning and teaching. She states that such corpora "...can be more useful as they are designed to represent the specific part of the language under investigation and are tailored to address the aspects of the language relevant to the needs of the learner" (p. 237). Observing the problems non-native writers from different disciplines, of different nationalities, and at different levels of proficiency faced in producing pragmatically acceptable text in their theses, Hancioğlu aimed to construct a pedagogic corpus that would improve teaching/learning outcomes by incorporating into her course corpus-informed data and tasks focusing on the lexico-structural features required to achieve specific moves. In this pedagogic study, academic abstracts were chosen for analysis since their basic move structure – (IMRD: Introduction–Method–Results–Discussion) (Swales, 1990) recurs throughout thesis and research writing in general. Therefore, they act as a kind of miniaturised version of the academic research genre as a whole, making them a powerful research and teaching device.

The Target Abstract Corpus (TAC) was compiled from universities in countries where English is the native language, though without attempting to make any distinction as to whether the authors were native or non-native speakers of English. The abstracts in the TAC, all published in a finalised form on the World Wide Web, were also produced by students, not 'experts'. Flowerdew (2000) draws attention to the importance of providing good 'apprentice' models rather than 'expert' generic models as these are more difficult to replicate due to learners' communicative and linguistic deficiencies. The corpus of abstracts (LAC) written by Hancioğlu's exclusively non-native postgraduate students living in a non-English medium country was collected separately. The purpose of the LAC was to observe the problems of the learners in detail and then, by identifying from the TAC a bank of patterns that would enable them to conduct moves with accuracy and appropriacy, help them develop their thesis writing skills.

The use of learner corpora has become increasingly common in EAP in recent years. Gilguin et al. (2007, p. 323), are strong advocates of the use of learner corpora in EAP research, and complain that "the overwhelming majority of corpus-based EAP studies are exclusively based on native corpora". They further cite Milton and Tsang (1991), who advocate the use of learner corpora to provide evidence that quantifies students' problems in written expression, and Flowerdew (2001, p. 364), who emphasizes that "insights gleaned from learner corpora need to be employed to complement those from expert corpora for syllabus and materials design" (2007, p. 322). The Cambridge Learner Corpus (CLC) (Cambridge University Press, 2008) and the International Corpus of Learner English (ICLE) (Centre for English Corpus Linguistics, n.d.) are two of the largest and better known learner corpora. In her discussion of the ICLE, Granger (2003, p. 543), states that evidence from learner corpora regarding learners' "under-, over-, and misuse can help materials designers and teachers select and rank ELT material at a particular proficiency level". Native corpus data, she says, does not give information about the degree of difficulty of structures and words for learners. Learner corpora, on the other hand, she adds, are "the resource par excellence to access this type of information" (p. 543).

Hancioğlu commenced her research by seeking to identify the most frequent content words (key words) of a sub-technical nature in her genre-specific target abstract corpus (TAC). The result was a list of 165 word families that seemed to be of fundamental importance in abstract writing. These words derived not only from the AWL (85 word families), but from the GSL (59), and included word families (21) not included on either list:

Again, what transpired was that the GSL contained words that were extremely common in academic usage and that the AWL contained words that were extremely common outside academia, as is clearly illustrated in Table 7 below in the BNL profile of the same words.

Needless to say, although these words are what would be termed by some authors as sub-technical, what was giving the list its academic texture was not really the items in isolation, but their *co-occurrence in the same environment.* And importantly, what seemed to be again confirmed by this exercise was the original thesis of Billuroğlu and Neufeld (2005) that there was little compelling rationale for the division between the GSL and AWL. It further seemed to confirm the initial hypothesis that teaching academic writing needed to be based on analysis of how lexical items collocated and combined to achieve specified moves for specified purposes within particular genres, and that many of the lexico-structural building blocks that were emerging were cross-disciplinary rather than disciplinary specific.

Analysis of Hancioğlu's Target Abstract Corpus (TAC) revealed the following breakdowns for the sub-corpora: (see Table 8).

Although the data show higher coverage of the Social Sciences sub-corpora, not only are the outcomes between the other three sub-corpora statistically very similar, in all four sub-corpora the BNL consistently provides higher coverage of the text than the GSL/AWL combination. The percentile differences though are not the major issue. The point that is made here is that corpus-informed pedagogy in EAP runs a grave risk of making

Table 7
BNL profile of key word examples from Hancioğlu's Target Abstract Corpus (TAC)

BNL1	BNL2	BNL3	BNL4	BNL5	BNL6	Off-list
base	aim	culture	critic	analyse	reform	dissertation
build	apply	govern	explore	construct		collaborate
case	collect	objectives		examining		correlate
change	combine	research		interview		
character	compare	theory		thesis		
consider	describe					
design	discuss					
develop	environment					
process	improve					
project	organisational					
	quantitative					

Table 8
Text coverage of the four sub-corpora in the Target Abstract Corpus

	GSL (K1, K2) and AWL (percentage of band text coverage)					BNL (percentage of band text coverage)								
	K1	K2	AWL	Total	Off-list	One	Two	Three	Four	Five	Six	Total	World knowledge	Off-list
Social Sciences	67.94	5.58	16.11	89.63	10.37	40.54	24.63	6.86	7.72	4.46	5.58	89.79	1.40	8.81
Humanities	66.06	4.78	13.81	84.65	15.35	41.68	21.96	6.54	6.31	4.01	5.09	85.59	0.95	13.46
Architecture	65.47	5.07	14.76	85.30	14.70	40.68	23.42	6.54	6.57	4.14	5.36	86.71	1.36	11.93
Sciences	63.24	5.74	16.06	85.04	14.96	39.03	23.31	6.92	6.09	4.35	5.45	85.15	1.82	13.03

serious teaching and learning omissions if it in any way assumes a category of pre-learned or taught general words and then selects for specific attention a group of what will in many cases be actually lower frequency items for particular focus in an EAP environment.

9. Moves and functions

The TAC offered ample data with which to analyze thesis abstracts according both to moves and discourse functions. Table 9 shows examples of the types of lexico-structural patterns that emerged and laid the basis for subsequent course design.

As can be seen, the basic patterns are almost exclusively of cross-disciplinary utility, for the simple reason that in each case they are used to fulfill the same generically driven function. To take just one example, that of *specifying the objective of the study* and looking in this regard only at lexico-structural patterns making use of the word *aim* and its derivations, the target corpus reveals that *aim* as a noun has three main lexico-structural realizations:

Table 9
Examples of categories from the TAC data (Lexico-structural patterns in bold)

MOVE	FUNCTIONS	EXAMPLES
Introduction	Introducing the Field	US fisheries legislation requires National Marine Fisheries Service (NMFS) to attend to the critical social and economic issues surrounding. . .
	Referring to previous work in the field	**Some researchers have found that** the influence of flowers promotes people positive emotions. (sic)
Need for the Study	Opening up a Research Gap	**This thesis posits the need to** integrate the design of landscape with the design of architecture
	Stressing the value of the study	...**this thesis demonstrates** an integrated design strategy and **its value and significance in** contemporary environmental design
	Stressing the challenge of the study	. . . **The solution of** linear systems **is an ancient and inexhaustible problem.**
	Opening up new links and relationships	**This thesis is dedicated to the study of** two seemingly unrelated problems...
	Stressing the originality of the study	**Ironically, few have attempted to use** pragmatism to articulate methods for ameliorating social difficulties. **This dissertation attempts to do just that**
Aims of the Study	Giving an overview of the thesis	**This thesis explores the integration**, through ideas of reciprocity, **of** landscape and architecture
	Specifying the objectives of the thesis	. . . **The intention is to demonstrate the usefulness of** a pragmatic **approach** to applied ethics
	Describing links and relationships	... **it is also central to this thesis that** this “reciprocal” relationship should be one of mutuality and interdependence,
Methodology of the Study	Giving information about research methods	Research methods included participant-observation, semi-structured ethnographic interviews (both in-person and on-line), and content analysis of text and visual data from Falun Gong books, pamphlets, and websites
	Giving information about research site	**Research sites included** Tampa, Washington DC, and cyberspace. . .
	Justifying choice of material and data	**The site and project were selected because they offered a good opportunity to explore the issues of** designing...
Conclusions	Stressing significance or novelty of findings	My **findings are contrary to** the allegations made by the Chinese Government and Western anti-cultists in many ways

(i) The (central/primary/main etc.) aim of this (thesis/dissertation, study etc.) is to (address/investigate etc.). Some variations to this pattern include use of plural ('aims') and use of the past tense ('aim was').
(ii) The (central/primary/main etc.) aim of this (thesis/dissertation, study etc.) is the (construction, design etc.) of ...
(iii) In chunks such as With the aim of ...ing and similar phrases such as In pursuing this aim, To meet this aim...

In total, these three lexico-structural realizations using *aim* as a noun recur 35 times in the corpus, the first being the most common with 23 realizations. *Aim* as a verb, again used to specify the objectives of the study, occurs another 43 times, in such realizations as:

(i) This (thesis/dissertation etc.) aims to (interpret/examine/describe etc.)...
(ii) This (thesis/dissertation etc.) aims at (...ing/the + noun).

Interestingly, *aim*, which occurs in total 86 times in the TAC, is another word that does not appear in the AWL, but in the second (K2) band of the GSL, meaning it too would run the danger of being excluded in any programme of study that took the AWL as its sole basis.

Whilst Hyland and Tse (2007) seem to suggest that specialist corpora should include products by both target discourse community and learners, Hancioğlu's approach was to keep the target and learner corpora distinct since comparison of the corpora would indicate not only the distance the learners needed to travel to reach target community standards, but would provide specific information about what they would need to do to achieve this. To take another example, the data below provide extracts from both target and learner corpora concerning *study*, the most common word in both the TAC and LAC.

Although the learner corpus was smaller than the target corpus, nonetheless as the data unfolded, it revealed that what typified the learner abstracts was a limited range of vocabulary, and an apparently limited productive knowledge of the collocations and colligations of even relatively common items. To illustrate this, Table 10 provides adjectival and verbal collocates (one to the left and one to the right) of *study*. In the case of the target corpus, verbs have been listed only up to the letter 'e'. With the learner corpus the same space has sufficed to list the entire repertoire of verb collocates employed by different learners over a full six academic semesters, amounting to a total of a mere 20 different types, in comparison to 102 types used in the target corpus.

In the case of adjectives used to the left of *study* meanwhile, the learner corpus revealed the use of descriptive adjectives only 19 times (7 different types), as opposed to 149 times (45 different types) in the target corpus. This gives a glimpse of the limited lexical resources available to this particular group of learners in comparison with the more sophisticated output emerging from the target corpus.

The advantage of genre-based corpus compilation and the use of concordancing tools is that lexico-structural relationships can be studied by researchers, instructors, and learners alike, and thus help narrow such gulfs. The potential for syllabus and materials design both for formal courses of instruction and for self-access learning is considerable, and furthermore is based on the authentic models to which learners are aspiring.

Table 10
TAC–LAC excerpt from collocates of ‘study’

TAC			LAC		
Adjectives		Verbs	Adjectives		Verbs
archaeological	study (*n*)	adds	case	study	aims
architectural		addresses	descriptive		applied
case		aims/aimed	documents		argues
cash-flow		analyzes/analyze	experimental		attempts
close		applies	field		based
comparative		argues	present		can / could
comprehensive		arises	time		consists
corpus-based		assessed/assesses			explores
cross-cultural		assumes			focus
current		attempts			has
disciplined		began			indicate / indicates
ethnoarchaeological		claims			investigates
ethnographic		clarifies			is/was/are/were
experimental		combine			offers
exploratory		compares			preferred
field		complements			provides
further		concludes			showed
in-depth		considered/consider			will
independent		consisted/consisting			
intensive		constitutes			
present		discerns			
		explores/explored			

Where then does this leave the wordlists? Interestingly, if all the items in Table 10 above are profiled, it emerges that 92.76% of the items come from the BNL 2709 list. This provides yet more evidence that practitioners and learners would well benefit from extending and deepening their knowledge of these general items through recycling and exploration. Looking at the LAC output for *study* and comparing it with the target corpus output in Table 11, we see that not only do the learners make no use of off-list words, they also make use of a restricted range of on-list words with a marked preference for words higher up the lists. A full 90% of the collocations with *study* in the LAC derived from the first three bands of the BNL, compared with under 60% in the TAC. All the evidence in this case indicates a deficit in productive knowledge not only of more specialized and less frequent lexis but also of what has been identified as general English lexis. Furthermore, each profiling exercise conducted with smaller and larger extracts from the corpora again revealed the lack of flexibility and options at the disposal of the learners in comparison with the writers of the texts in the target corpus.

10. Some pedagogical considerations

This last exercise returns us to our initial assertion that really ‘knowing’ a word means knowing a lot of other words. As long as some caution is exercised to avoid computer generated frequency lists being misappropriated by schools for rote-memorization and mechanical testing, both wordlists and collocation banks have the potential to liberate learners and contribute to autonomy in learning as well as vocabulary and language development. And certainly if 2709 word families can be shown to account for around 85–90%

Table 11
Comparison of frequency profiles of collocates of ‘study’ in Learner and Target Corpora according to BNL

	BNL cumulative totals (%)	
	Learner	Target
BNL-1 words	56.67	32.08
BNL-2 words	76.67	50.95
BNL-3 words	90.00	58.50
BNL-4 words	96.70	67.93
BNL-5 words	100.00	86.69
BNL-6 words		88.69

of most texts, then it would seem reasonable to integrate this lexis into the lexical syllabus of courses of instruction in a principled way, as well as continuing to refine the lists themselves.

At the same time, it is important to establish that whilst word families in general frequency lists may be of central importance, they do not offer the option of discarding the rest of the English lexicon. As has been illustrated, knowledge of individual word families involves a level of sophistication and breadth and depth of reference that is extremely high. This requires that instructors devote considerable time to helping learners develop lexical awareness by providing sufficient exposure for genuine acquisition to take place, and by offering plentiful opportunities for production. Principled approaches in this regard have been suggested, for example, by Cobb (2007b) in his approach to computer adaptive data-driven learning support for reading and Nation (2007) in his work on the four strands of meaning-focused input, meaning-focused output, language-focused learning and fluency development.

The development and use of disciplinary specific wordlists also still requires some thought, since the majority of the lexical items that students need to either understand or produce will inevitably derive from more general purpose lists. Further, much key specialist terminology may just be acquired through definitions, glossing and frequency of natural encounter in lessons with subject specialists. If the ESP instructor is not intending to explore the collocational and colligational behaviour of such items in further depth, it is questionable whether any bona fide language support is really being provided.

11. Conclusion

The main purpose of this collaborative study has not however been to gainsay the practical usefulness of specific wordlists, and nor has it been to suggest that there are no major linguistic differences between academic disciplines. We do, however, argue that:

(i) The distinction between general and so-called academic lexis is not clear-cut enough to sustain an Academic Word List attached as a third and distinct band to the GSL.
(ii) There is strong evidence to show that combining wordlists such as the GSL and AWL into a more comprehensive general list and banding them according to natural frequency of occurrence will provide teachers and learners of both general and academic English with a more useful resource, less likely to lead to the gaps and omissions in learning that may result from their separation and the artificial categorization of items into *either* general *or* academic.

(iii) Although there is evidence that suggests that there is considerable variation between academic disciplines and fields, there is also considerable evidence that is also suggestive of high levels of commonality, particularly when texts are examined generically.
(iv) The complexity and richness of such cross-disciplinary "terminology of rhetoric" (Hunston, 2002, p. 135) suggests also that there is still great benefit for students of academic English in following corpus-informed programmes that have been designed on the basis of cross-disciplinary studies.
(v) Clear distinctions need to be drawn between pure research findings and practical pedagogic implications. Single disciplinary features may be dealt with far more effectively by subject specialists than by language specialists. Further study would be required to show the practical implications of disciplinary variation for language teaching specialists.
(vi) Disciplinary variation is also accompanied by learner variation. Learners enter academic institutions at different levels and with different needs and aspirations. Many require multiple literacies and have linguistic needs that cannot be disciplinarily restricted. In many cases a more wide-ranging and general approach may serve their interests better.

The teaching of both general and specific academic English may therefore still be profitably organized around the use of relatively broadly-based wordlists, supplemented by more specialized genre-based banks of lexico-structural patterns. As already noted, from the foundation of a frequency-based general vocabulary, further vocabulary acquisition, targeting knowledge of the way words collocate and colligate, can then be promoted. Basically, the learning of vocabulary is facilitated by prior vocabulary knowledge. Mastering the most frequent words in English opens up texts, providing the context through which to infer unknown vocabulary, increasing reading speed and efficiency and thus assists with further lexico-structural acquisition, both receptive and productive. This is yet more reason not to prematurely eschew in-depth study of frequent words in favour of specialized terminology, and to adopt an approach to ESAP that is additive rather than substitutive.

References

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*(3), 1–27.

Bauman, J. *About the GSL [Data file]*. Retrieved from <http://jbauman.com/gsl.html>.

Bhatia, V. K. (1993). Analysing genre: Language use in professional settings. London: Longman.

Billuroğlu, A., & Neufeld, S. (2005). *The Bare Necessities in Lexis: A new perspective on vocabulary profiling*. Retrieved from <http://lextutor.ca/vp/BNL_Rationale.doc>.

Billuroğlu, A., & Neufeld, S. (2007). BNL 2709 The essence of English (4th ed.). Nicosia: Rüstem Kitabevi.

BNC Consortium. (2005). *The British National Corpus*. Retrieved from <http://www.natcorp.ox.ac.uk/>.

BNL. (2007). *Retrieved from the Bare Naked Lexis Wiki*. <http://www.editthis.info/thebnl/>.

Cambridge University Press. (2008). *Cambridge Learner Corpus*. Retrieved from <http://www.cambridge.org/elt/corpus/learner_corpus2.htm>.

Centre for English Corpus Linguistics. (n.d.). *International Corpus of Learner English (ICLE)*. <http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm> Retrieved 10.05.08.

Chan, S. K., & Foo, S. (2001). Bridging the interdisciplinary gap in abstract writing for scholarly communication. *Paper presented at GENRE 2001 (Genres and Discourse in Education, Work and Cultural Life: Encounters of Academic Disciplines on Theories and Practices)*. Norway: Oslo University College (May).

Charles, M. (2007). Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes, 6*, 289–302.

Chen, Q., & Ge, G. (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.

Chujo, K., & Utiyama, M. (2005). Understanding the role of text length, sample size and vocabulary size in determining text coverage. *Reading in a Foreign Language, 17*(1). Retrieved from <http://nflrc.hawaii.edu/rfl/April2005/>.

Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language, 15*(2). Retrieved from http://nflrc.hawaii.edu/rfl/October2003/chung/chung.html.

Cobb, T. (2006). *Web VP Classic (Version 2.7) [Software]*. <http://lextutor.ca/vp/eng/>.

Cobb, T. (2007a). *Web VP BNL (Version 2.3) [Software]*. <http://lextutor.ca/vp/bnl/>.

Cobb, T. (2007b). Computing the vocabulary demands of L2 reading. *Language Learning and Technology, 11*(3), 38–64.

Cobb, T. (n.d.). *The compleat lexical tutor for data-driven learning on the web [Software]*. <http://lextutor.ca/>.

Cowan, J. R. (1974). Lexical and syntactic research for the design of EFL reading materials. *TESOL Quarterly, 8*(4), 389–400.

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

Coxhead, A. (2008). Phraseology and English for academic purposes. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 149–161). Amsterdam: JohnBenjamins.

De Jong, J., & Ackermann, A. (2008). Lexical validity for academic English tests. *Paper presented at IATEFL annual conference*, Exeter, UK (April).

Dickins, J. *Extended version of a General Service List of English words [Data file]*. Retrieved from <http://www.languages.salford.ac.uk/staff/dickins/GSList2.xls>.

Edict. *The first 2000 most frequent words from the Brown Corpus [Data file]*. Retrieved from <http://www.edict.com.hk/lexiconindex/frequencylists/words2000.htm>.

Eldridge, J. (2008). No, there isn't an 'academic vocabulary', But: A reader responds to K Hyland and P. Tse's Is there an 'academic vocabulary'? *TESOL Quarterly, 42*(1), 109–113.

Ellis, N. C. (2002). Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143–189.

Engels, L. K. (1968). The fallacy of word counts. *IRAL, 6*, 213–231.

Farrell, P. (1990). Vocabulary in ESP: A lexical analysis of the English of electronics and a study of semi-technical vocabulary. *CLCS Occasional Paper No. 25 Trinity College.*

Flowerdew, L. (2000). Using a genre-based framework to teach organizational structure in academic writing. *ELT Journal, 54*(4), 369–378.

Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy & R. Roseberry (Eds.), *Small corpus studies and ELT* (pp. 363–379). Amsterdam: Benjamin.

Folse, K. (2004). Vocabulary myths: Applying second language research to classroom teaching. Ann Arbor, MI: University of Michigan Press.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28*(2), 241–265.

Gilguin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes, 6*, 319–335.

Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 37*(3), 538–546.

Halliday, M. A. K., & Hasan, R. (1985). Language, context, and text: Aspects of language in a social-semiotic perspective. Oxford: OUP.

Hancioğlu, N., & Eldridge, J. (2007). Texts and frequency lists: Some implications for practising Teachers. *ELT Journal, 61*(4), 330–340.

HarperCollins Publishers. (2004). *Cobuild concordance and collocations sampler [Software]*. <http://www.collins.co.uk/corpus/CorpusSearch.aspx>.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *RANGE and FREQUENCY programs [Software]*. <http://www.vuw.ac.nz/lals/staff/Paul_Nation> or as AntWordProfiler (Version 1.103, Windows and Macintosh) from <http://www.antlab.sci.waseda.ac.jp/software.html>.

Henry, A. (2007). Evaluating language learners' response to web-based, data-driven, genre teaching materials. *English for Specific Purposes, 26*(4), 462–484.

Hunston, S. (2002). Corpora in applied linguistics. Cambridge: CUP.

Hwang, K., & Nation, I. S. P. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System, 23*(1), 35–41.

Hyland, K. (2005). Perspectives on EAP. *ELT Journal, 59*(1), 57–64.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*, 4–21.

Hyland, K., & Tse, P. (2007). Is there an "Academic Vocabulary"? *TESOL Quarterly, 41*(2), 235–253.

Johns, A. M., Bawarashi, A., Coe, R. M., Hyland, K., Paltridge, B., & Reiff, M. J. (2006). Crossing the boundaries of genre studies: Commentaries by experts. *Journal of Second Language Writing, 15*(3), 234–249.

Kennaway, R. *The Longman defining vocabulary [Data file]*. Retrieved from <http://www2.cmp.uea.ac.uk/~jrk/conlang.dir/LongmanVocab.html>.

Kilgarriff, A. *Lemmatized BNC frequency list [Data file]*. Retrieved from <http://www.kilgarriff.co.uk/BNC_lists/lemma.num>.

Krashen, S. (2003). *Explorations in language acquisition and use: The Taipei lectures. Portsmouth*. NH: Heinemann.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307–322.

Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes, 25*(1), 56–75.

Lextutor. *1000 families [Data file]*. Retrieved from <http://www.lextutor.ca/freq/lists_download/1000_families.txt>.

Lextutor. *2000 families [Data file]*. Retrieved from <http://www.lextutor.ca/freq/lists_download/2000_families.txt>.

Lextutor. *AWL headwords [Data file]*. Retrieved from <http://www.er.uqam.ca/nobel/r21270/freq_lists/awl_heads.txt>.

Lextutor. *AWL sublists [Data file]*. Retrieved from <http://www.er.uqam.ca/nobel/r21270/freq_lists/awl_families_sublists.doc>.

Liu, N., & Nation, I. S. P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal, 16*(1), 33–42.

Longman. (2003). *The Longman wordwise dictionary*.

McCarthy, M. (2001). Issues in applied linguistics. Cambridge: CUP.

McCarthy, M., & O'Dell, F. (2008). Academic vocabulary in use. Cambridge: CUP.

Milton, J., & Tsang, E. (1991). A corpus-based study of logical connectors in EFL students' writing: Directions for future research. In R. Pemberton & E. Tsang (Eds.), *Studies in Lexis* (pp. 215–246). Hong Kong: The Hong Kong University of Science and Technology.

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

Nation, P., & Waring, R. (2004). *Vocabulary size, text coverage and word lists*. Retrieved from <http://www.wordhacker.com>.

Nation, P. (2001). Learning vocabulary in another language. Cambridge: CUP.

Nation, I. S. P. (2007). The four strands. *Innovation in Language Learning and Teaching, 1*, 1–12.

Paquot, M. (2007). Towards a productively-oriented academic wordlist. In J. Walinski et al. (Eds.), *PALC Proceedings* (pp. 127–140). Frankfurt: Peter Lang.

Pearson Language Tests. (n.d.). *Pearson Test of English.*<http://www.pearsonlanguageassessments.com/home/exams/pte> Retrieved 24.06.08.

Swales, J. M. (1990). Genre analysis. Cambridge: Cambridge University Press.

Swales, J. M. (2004). *Research genres. Exploration and applications*. Cambridge: CUP.

Watt, R. (2004). *Concordance (Version 3.2) [Software]*. <http://www.concordancesoftware.co.uk>.

West, M. (1953). *A General Service List of English words*. London: Longman, Green & co.

Wordpilot. *Academic Reports [Data file]*. Retrieved from <http://www.home.ust.hk/~autolang/AcademicReports.zip>.

Nilgün Hancioğlu teaches at the Department of General Education, Eastern Mediterranea n University, Northern Cyprus. She has an MA in ELT from the Middle East Technical University, Ankara and is currently working on her PhD dissertation at EMU. Her research interests include academic writing, corpus studies, data-driven learning, and lexical semantics.

Steven Neufeld is currently in charge of research, development and training for *The English Language Consultancy Association* in Northern Cyprus. He has a B.Ed. from the University of Saskatchewan and an M.Sc. from

Leicester University. His current interests include vocabulary profiling, data-driven learning and web-based learning environments.

John Eldridge also teaches at EMU, at the Department of General Education. He has an MSC in ELT from Aston University and an MBA in Educational Management from Leicester University. His current interests include semantic frequency, teacher autonomy, and the discourse of management.