Encoding the Past, Decoding the Future

Encoding the Past, Decoding the Future: Corpora in the 21st Century

Edited by

Isabel Moskowich and Begoña Crespo

CAMBRIDGE SCHOLARS PUBLISHING

Encoding the Past, Decoding the Future:
Corpora in the 21st Century,
Edited by Isabel Moskowich and Begoña Crespo

This book first published 2012

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-3581-1, ISBN (13): 978-1-4438-3581-7

TABLE OF CONTENTS

PREFACE

Till the Future dares Forget the Past.
—Shelley, *Adonais* i, (1821)

From very early in history the study of written records has been one of the main concerns of scholarly activity. Therefore, philology constitutes a systematic attempt to obtain linguistic, cultural and historical information from documents. From the more primitive literal readings to the interpretations of texts and manual collecting of data to the retrieval of information about languages and the different peoples using them, many centuries have passed. In less than a century, however, we have witnessed how different technological devices have made it possible to reach more reliable conclusions in a much shorter period of time. The empirical study of language by (semi)automatically retrieving data from texts is the essence of corpus linguistics.

We have decided not to participate as authors here and, instead, we have preferred to use our experience, if any, to try and edit a volume we are sure will have something to offer to the scientific community. Our work as compilers of the *Coruña Corpus of English Scientific Writing* in its different sub-corpora provides us with a hint about what corpus-compilation is and the time-consuming tasks it entails. As users of corpora (our own, the *Coruña Corpus*, as well as some of the other available in these days) we can also appreciate the tools modern technology offers researchers and what the possibilities of exploitation are.

As a twentieth-century innovation in our field, corpora and corpus linguistics have been present in research and study for a reasonably long time now. Once the new century has already witnessed conferences, new publications and all sorts of events related to the field, it seems most convenient not to revise the evolution of the discipline but rather, to try to offer an outline of the advances made in the past decade as well as to try and make a guess as for what is yet to come.

The same way the first collections of texts, linguistic data, were not at first thought of as corpora in our modern sense, the discipline we now easily recognise as Corpus Linguistics may probably not have been considered a branch of linguistics in itself but only a good way to process those linguistic data. Nowadays, Corpus Linguistics as a methodology has

proved to be an impeccable one, probably the most elaborate way to approach empirical studies on languages which are the first step to formulate more general theories or hypotheses about practically all aspects of languages at all historical recorded periods.

The works here collected aim at offering a good representation of the type of research carried out in these first years of the twenty-first century. Therefore, they range from the presentation of new exciting projects as the *Cadis Corpus* by Maurizio Gotti (chapter 7) to very specific studies such as the one presented by María de los Angeles Gómez González who, in chapter 3 offers a corpus-based contrastive analysis of the behaviour of tag questions in English and Spanish. More than one language are also present in Gómez-Guinovart's work on bilingual terminology extraction (chapter 6) and information retrieval is also the concern of Julia Lavid's chapter 8 when dealing with the identification and extraction of metaphor from computerised corpora.

Different theoretical approaches to grammar and language pervade the pages of this volume and are adopted by contributors to it as well as different languages are used to express their ideas. In chapter 2, Jose María García-Miguel beautifully describes the lights and shadows of verbal valency as seen through corpora and applies it to the particular case of present-day Spanish.

Other contributors aim at establishing some kind of rank or taxonomy on linguistic/mental artefacts that have not yet been defined satisfactorily enough or on constructions that have been a little bit slippery for scholars all along history… In fact, the opening chapter of this book by Averil Coxhead and Patricia Byrd, "Collocations and Academic Word List: The strong, the weak and the lonely", offers such a brave attempt to provide a classification for a loose term referring to a strong structure. Similarly, Martínez-Insua and Pérez-Guerra's work on meta-informative strategies, presented in the middle pages of this volume, means a step forward in linguistic study and opens new perspectives for future research.

We have wanted to include the work not only of those who are at present compiling new corpora or developing their research in the opening of the twenty-first century but also that of those who paved the way for us when computers were not so quick and statistics was something only the initiated resorted to. Therefore, chapter 4 by Manfred Markus is here offered as a symbol of the union of what has been done so far and what can be done yet on corpus linguistics. His work on spokenness as a feature of late Modern English and the way in which corpus linguistics methodology can help modify pre-conceived ideas or general assumptions on languages and stages of languages occupies also a central position.

As is obvious, our choice of authors from different countries and continents as well as the use of different languages has been a conscious one. It is true that corpus linguistics as a methodology as well as a branch of Modern Linguistics was first developed in English speaking countries and, when not, to analyse different aspects either synchronic or diachronic of English. However, the development of the discipline is such that it is now applied to almost all languages in the world, either spoken or written, and to almost any historical stage in their development. The lexicon is not the only protagonist any more and many other aspects, almost bordering extra-linguistic areas, are also the object of researchers' interest.

After a careful compilation, the pages that follow contain, we hope, a wide variety of the trends at present being developed in the world. If this volume shall anyway contribute to the end proposed, let the authors have the glory, and the editors the good wishes and gratitude of readers.

—The Editors

CHAPTER ONE

COLLOCATIONS AND ACADEMIC WORD LIST: THE STRONG, THE WEAK AND THE LONELY

AVERIL COXHEAD AND PATRICIA BYRD

1. Introduction

An ever expanding body of research is demonstrating the various strong relationships among words as they are used for communication for particular purposes (Baker 2006; Biber 2006; Carter and McCarthy 1995; O'Halloran 2005; Shin &Nation, 2008; Ellis and Simpson-Vlach, 2010). These relationships exist at various levels of system in a language. Words form strong relationships in at least these patterns: (a) Connected sets at the phrase level of analysis (often termed *lexical bundles* or *clusters*) (see Pickering and Byrd 2008; Biber, Conrad, and Cortes 2004; Hyland 2008; and Byrd and Coxhead 2010 for examples); (b) discontinuous sets (often termed *frames*); (c) adjacent and non-adjacent pairs (often discussed as *collocations*. See Shin and Nation (2008) for a discussion of high frequency collocations in spoken English); (d) clauses and lexicogrammatical relationships; and (e) semantic connections as well as syntactic ones.

Nation (2001: 27) highlights form, meaning and use as three aspects of knowledge important for language learners (see Table One below). These categories start with the basic concepts of how a word is written and spelled for written purpose or pronounced for spoken purposes. It moves on to more complex areas of knowledge associated with a word's meaning, including its associations and collocations, and constraints on its use such as frequency and register. The key question related to collocations for learners is 'what words or types of words must we use with this one?'

Table 1. Knowledge required for production of a word in writing (adapted from Nation, 2001: 27)

Form		**How is the word written and spelled?**
Meaning	Form and meaning	What word form can be used to express this meaning?
	Concepts and referents	What items can this concept refer to?
	Associations	What other words can we use instead of this one?
Use	Grammatical function	In what patterns must we use this word?
	Collocations	What words or types of words must we use with this one?
	Constraints of use (register, frequency…)	Where, when and how often can we use this word?

The Academic Word List (AWL) (Coxhead 2000) is a well known and widely used list of 570 word families developed using a written academic corpus of 3,500,000 running words. An example of a word family is *benefit*, *beneficial*, *beneficiary*, *beneficiaries*, *benefited*, *benefiting*, and *benefits*. The corpus was divided into arts, commerce, law, and science, with approximately 875,000 running words each. It contained 414 texts, balanced for length when possible, including textbooks, articles, book chapters, and laboratory manuals. The selection of word families for inclusion in the AWL was guided by four key principles. The first was that the 2,000 most frequent word families of West's *General Service List of English Words* (GSL) (1953) would be excluded. This decision was made because the focus of the list was not general vocabulary. The GSL is considered old and does not include some current everyday words such as *computer*and *television*, but this list has still not been replaced. Frequency, range, and uniformity were the three other principles for selecting the AWL words. Word families had to occur 100 times or more in each of the four discipline areas of the corpus, in 15 or more of the subject areas, and ten times or more in the four disciplines. The AWL is made up of ten sublists. Sublist One contains the 60 most frequent word families in the AWL, Sublist Two contains the next 60 most frequent word families, and

Table 2. The coverage of the AWL over different academic and non-academic corpora (adapted from Coxhead 2011)

Type of texts	Study	Corpus	Number of running words	Coverage of the AWL
University level Texts	Cobb and Horst (2004)	*Learned* section of the Brown corpus (Francis and Kucera,1979)	14, 283 words	11.60%
	Konstantakis (2007)	Business	1 million words	11.51%
	Ward (2009)	Engineering	271,000 words	11.3%
	Vongpumivitch, Huang and Chang (2009)	Applied linguistics research papers	1.5 million words	11.17%
	Hyland and Tse (2007)	Sciences, engineering, and social sciences, written by professional and student writers	3,292,600 words	10.6%
	Li and Qian (2010)	Finance	6.3 million words	10.46%
	Chen and Ge (2007)	Medical research articles	190, 425 words	10.073%
	Martínez, Beck, and Panza (2009)	Agricultural sciences research articles	826,416-words	9.06%
	Coxhead and Hirsh (2007)	Science	1.5 million words	8.96%
Secondary school level texts	Coxhead, Stevens, and Tinkle (in press)	Pathway series of secondary science textbooks	279, 733 words	7.05%
Non-academic texts	Coxhead (2000)	Fiction	3,500,000 words	1.4%
	Coxhead (unreported)	Newspapers	1 million words	4.5%

so on. Coxhead (2011) has downloadable versions of the headwords and sublists of the AWL. On average, the AWL covers 10% of the AWL corpus. In subsequent work by a variety of researchers and academic corpora (see Table Two above), the AWL coverage is usually around 10%. Contrast those figures with fiction (1.4%) and newspapers (4.5%) respectively.

The more we know about the words that commonly occur in academic discourse and the lexical and grammatical patterns in which they occur, the more we can look for similarities and differences and work to inform students, researchers, teachers, and materials designers. Durrant (2008: 163) explains two pedagogical benefits of lists of collocations for language learning. The first is that learners' attention can be drawn to patterns that learners need. The second is, "they may draw attention to productive patterns which are tied to specific lexis in a way that can lead them to be overlooked by traditional grammars".

Durrant's research illustrates a careful approach to developing a list of collocations based on an academic corpus. Table Two below lists ten of the top academic collocations from Durrant's list, along with their mean frequency per million words. Note the drop in mean frequency from *between and* at 935 occurrences per million to *related to* with 190 occurrences per million.

Table 3. Ten of the top 100 key academic collocations adapted from Durrant (2008: 166-168)

Words 1 and 2	Mean frequency/million words
between and	935.56
can be	857.4
number of	634.6
based on	404.64
due to	374.12
associated with	315.52
according to	267.36
and respectively	249.68
in addition	204.72
related to	190.72

Durrant (2008) points out that most of the collocations he finds in his study of academic texts are not in the AWL.

For more on collocations in a particular academic subject area, see Ward's (2007) work on engineering which includes a sample of practical learning tasks for the classroom. See also Gledhill (2000) for an interesting

discussion and exemplification of the discourse function of collocation in a corpus of cancer research articles, which builds on earlier work in specialised corpora by the same researcher.

A recent corpus-based analysis of patterns formulas in academic and non-academic corpora of speaking and writing is Simpson-Vlach and Ellis's (2010) Academic Formulas List. This major study directly addresses pedagogical concerns. The researchers sorted the data from their study "according to major discourse pragmatic functions [which] allows teachers to focus on functional language areas which, ideally, will dovetail with functional categories already used in EAP curricula" (497). The researchers present a 'core' list of written and spoken academic formulas, data from primarily spoken academic English, and data from primarily written academic English. For example, formulas of contrast and comparison (p. 499) such as *and the same* and *as opposed to* are from the core AFL, *(nothing) to do* and *the same thing* are primarily from spoken data, and *be related to the* and *is more likely* are primarily from the written data. Like Durrant (2008), these researchers used comparison corpora to establish the academic nature of their formulae. A limitation of the present study is that no comparison corporaares used.

In this approach to language analysis, language is seen as being organized by repeated use of relatively set wordings in particular discourse settings. The study of language as it has been used for various communicative and discourse purposes has developed rapidly with the practical, applied use of computers to analyze corpora; a theory base for corpus linguistics has been developing with similar speed. Work such as Sinclair (2004), Halliday (2002), Hoey (2005), Stubbs (2005), Biber (1988) and Nation (2001) provides a theoretical framework that explains, justifies, and builds on an empirical approach to linguistics and a definition of language in terms of lexicogrammatical patterns used for communication among members of social groups. Thus, we anticipate that the words in the AWL will exhibit the same interconnectedness found in other analyses of words-in-use-in-register-contexts. Firstly, however, we need to look at some challenges when investigating patterns in written language.

2. Theoretical and methodological challenges in the study of word patterns

A primary problem with *collocation* seems to be one of definition. *Collocation*, as a generally agreed upon definition, is the characteristic co-occurrence of words. Firth (1957, as cited in Xiao and McEnery 2006) is

credited with first having used it as a technical term. Variations of Firth's explanation abound, all seeming to share ambiguous terms such as *characteristic*, *greater than chance*, *habitual*, and *frequent*. This vague terminology seems the crux of a growing dilemma in collocational research in corpus linguistics. Does the definition of *collocation* require a quantifiable, statistical relationship between collocates? Perhaps such a relationship is implied, even assumed to be part how *collocation* is defined. According to Oakes (1998) the term *significant collocation* can be applied when the probability of two lexical items occurring together within a specified span is greater than could be expected by chance. Since he does not seem to be advocating two categories of non-significant collocation vs. significant collocation, the definition provided in this often cited work on statistics and corpus linguistics means that, for Oakes, a collocational relationship can only be found through the application of statistical procedures and not through the educated guessing of the researcher.

Hunston (2002: 12) defines *collocation* in two slightly different ways. In an introductory chapter she says "collocation is the statistical tendency of words to co-occur," (12), yet later calls collocation "the tendency of words to be biased in the way they co-occur" (68). While these definitions are far from contradictory, they do bring to attention an important detail; is a statistical relationship necessarily implied by the term *collocation*? Her examples using the statistical data from the Bank of English suggest that she expects a statistical process to be applied in the search for collocations.

McEnery, Xiao, and Tono (2006) state that collocates are identified by using statistical approaches, perhaps meaning that while the notion of statistical analysis may not be implicit in the definition of collocation, it is the way in which collocates are determined. This close examination of the definition may seem pedantic, but there are instances in the literature where extracting collocates seems to involve little more than finding words near the node that intuitively seem as though they are related.

2.1. Selecting a statistical approach

For the most part the statistics for recognition and validation of collocations rely on some form of comparing the observed frequency of a pair of words with the frequency expected by chance. These formulae are new incarnations of statistics used in experiments of the social sciences, most of which evolved from statistics used in hard sciences. The formulae needed to be manipulated when they were applied to groups of people in social science use. Now further adaptations are needed to try to make them fit language. This attempt to use statistical approaches created for other

types of research has left formulae that are not a good fit with linguistic data. These statistics were never intended to measure language. Many, such as t-score, z-score, and MI score, include standard deviation as part of the calculation. In order to determine standard deviation, it is necessary to assume a normal distribution, and in order to have a normal distribution, it is further necessary to assume the possibility of randomness. However, language is not random (Kilgarriff 2005). Language does not distribute normally. Stubbs (1995) addresses many of the problems associated with collocational statistics and seems to conclude that one should proceed with caution. In other words, the statistics can still provide some useful information, but we need to be aware of their inherent limitations when applied to language data.

Considering these limitations, we have chosen to report log likelihood (LL) scores for this project. Log likelihood does not assume a random distribution of the data and therefore more meaningful results can be obtained from less data (Dunning 1994). The assumption behind the use of LL analysis with language data is that massive amounts of data should yield something closer to a normal distribution. This decision is in line with the practice in introductions to corpus linguistics such as Baker (2006) and McEnery, Xiao, and Tono (2006). Because the field is in such flux, we are also reporting as much raw frequency data as possible so that the data that we present about the AWL can be re-analyzed in the future.

2.2. Methodology

The corpus used in this study is the 3,500,000 word written academic corpus from Coxhead's AWL study (2000). It is briefly described in the introduction above and detail on the corpus is available on Coxhead's (2011) website and in Coxhead (2000).

2.2.1. Principles to guide selection and presentation of data

Working with 570 word families and a 3.5 million-word corpus involves problems that result from having substantial amounts of data, especially for the higher frequency words. For example, *assessment*, the most frequent member of the word family *assess*, occurs 684 times in the AWL corpus. Wordsmith Tools 4.0 (Scott 2007) reports collocational relationships with 92 words in positions following *assessment* with log likelihoods ranging from 1007.57 (with *of*) down to 3.14 (with *have*) For the word *assessment*, that software package also provides 138 3-word clusters that occur at least three times in the AWL corpus. Thus, decisions

must be made about which data are of most importance for understanding how *assessment* functions in academic writing since reporting all of this data would be some combination of un-wieldy and un-useful for most readers.

Because many word families include four or more related words, the AWL totals 2538 words. Rather the studying all members of the word families, this current study focuses on the most frequent member of each word family, limiting the task to the analysis of the use-in-content of 570 words. Where two words in a family have similar frequency, the study presents information about both members of the family. However, generally one member of such sets is much more commonly used and, thus, the focus of our study.

As indicated earlier in this discussion, we have chosen to use WordSmith Tools 4.0 for our analysis and to report the log likelihood statistic for collocational information. These two decisions about software and statistical approach led to the development of extensive data sets. For example, the word *concept* occurs 580 times in the AWL corpus. Wordsmith Tools reports 487 possible collocates for *concept*, most of which are well below any reasonable standard for statistically significant relationships. In the case of *concept* (and many other words), the highest log likelihood score for the relationship between *concept* and *of* is followed by much smaller scores for other pairs. Additionally, Wordsmith Tools 4.0 reports collocational data in two ways, the first being the data for the node and words that follow it and the second report being for relationships between the node and words that come in front of it: for example, *concept of* vs. *the concept*. To give readers a sense of these numerical (and linguistic) patterns, we decided to report the top ten relationships in each direction. Tables Four and Five show the data as reported by Wordsmith Tools 4.0 and copied into Microsoft Excel. As can be seen in the columns for the log likelihood data, the fall off in values is steep.

Since no cut off values exist for log likelihood, principled decisions must be made and then consistently followed in reports of data. To indicate the strength of collocational relationships for the AWL words, we decided to report the top five scores in each pattern as shown in Appendix A.

Table 4. Collocational data in Log Likelihood order, for *concept* followed by possible collocates

Word 1	Freq. of Word 1	Word 2	Freq. of Word 2	Gap between the words	Joint use of the words	Log L.
concept	580	of	146,362	1	375	1,674.97
concept	580	oppression	144	2	18	206.92
concept	580	is	51,830	1	74	201.09
concept	580	a	72,595	2	61	108.51
concept	580	has	9,606	2	24	88.14
concept	580	the	248,132	2	105	82.56
concept	580	in	81,262	1	54	75.24
concept	580	citizenship	147	2	6	54.91
concept	580	it	22,334	4	24	51.51
concept	580	that	40,856	4	31	48.89

Table 5. Collocational Data in Log Likelihood order for *concept* preceded by possible collocates

Word 1	Freq. of Word 1	Word 2	Freq. of Word 2	Gap between the words	Joint use of the words	Log L.
the	248,132	concept	580	1	448	1,804.28
a	72,595	concept	580	1	91	228.39
to	88,802	concept	580	2	77	142.57
of	146,362	concept	580	2	88	112.3
this	20,153	concept	580	1	34	100.64
marketing	1,068	concept	580	1	14	96.35
is	51,830	concept	580	3	47	88.74
matching	64	concept	580	1	5	52.42
brookers	5	concept	580	3	3	45.76
law	4,468	concept	580	4	12	45.48

3. Results and discussion

In this section, we report on data from several words in the AWL and illustrate what we mean by 'the strong, the weak, and the lonely'. Let's start with the strong.

3.1. The strong

In this section, we report on case studies of words from the AWL which present different patterns of collocation.

3.1.1 The case of create

Create is taken as an example of the 'strong' in this chapter because an analysis of this word provides a great deal of collocational data. One of the difficulties with dealing with an avalanche of corpus-generated data is deciding whether to limit the analysis to particular kinds of words such as nouns, verbs, or adjectives or to cast the net wider. If we take *create* as an example, a narrow analysis would show two main categories of nouns: concrete and abstract creations. The concrete uses in the AWL corpus include *document, environment, database, record,* and *field*. Most of these uses seem to come from computer science. On the other hand, the abstract uses include *impression, difficulties, reasons, problems*, and *rights*. While we can use the word *create* to communicate about the creation of just about anything, there's a strong tendency in academic prose for the word to be used in these patterns with these words. This strong tendency makes a principled beginning place for teaching students how to use the word.

When we take a wider view of this high frequency word, we see a number of relationships. The strongest relationship is between *create* and *a/an*. This set makes the phrases that begin "create a …" and "create an …" As we saw above, several words suggest computer language either with word processing or working with databases: *document, database, record,* and *field*. The nouns that follow *create* include a subset with negative meanings: "create difficulties" and "create problems."

Other nouns include *impression, environment, reasons,* and *rights*. *Create* is also used in compounds with *or* or *and*. For example: …*the difficulty of trying to create and maintain an ideology.*…. The adjectives that collocate with create fall into two subgroups. One is focused on the number of things created, for example new and additional. The other subset focuses on the quality of the new creation, for example, *create a new document* or *create a good curriculum*. Several prepositions collocate

with create. The highest on the list is for in phrases such as *create … for …*: *It can create major problems for those who have to find the finance. Create* is followed by nouns, and those nouns are like many nouns in academic prose in that they are part of long, complicated noun phrases. Having prepositions and relative pronouns on the list suggests the presence of these long noun phrases as in … *create other problems that make its use undesirable*. This little pattern suggests at least two concerns:

(a) English uses the singular nouns with *a/an* to refer to categories, for generic meaning, in sentences such as "She's a lawyer" or "I need a pencil" that refer to general rather than particular items. The use of generic in academic prose should be checked.

(b) The set of words creates a short frame that can be completed with many different nouns. What patterns exist in the kinds of items that are created in academic communication". Is there any patterning that suggests a way to prioritize learning of new words that go in the frame?

3.1.2. An analysis of *analysis*

Another example of a high frequency word in the AWL with strong collocational relationships is *analysis*. The strongest log likelihood relationship is between *analysis* and *of*. This relationship spans other strong log likelihood relationships, including *an analysis* (as in *an analysis of*) and *for … analysis* (as in *for an analysis of*). *Analysis of* is followed by nouns such as *interaction, language, system, programme, the effect, the changes,* and *data.* A key point to make here is that the relationship between *analysis* and *of* goes both ways. *Of* also precedes *analysis*, and *of analysis* often occurs in patterns such as *method/unit type/level of analysis.* Another example of a two-way connection between words is *analysis and* along with its reverse *and analysis.* While the data for the pairs are similar, the collocates for these patterns are quite different. Examples of *analysis and* collocates include *assessment, evaluation, interpretation, management,* and *results.* Examples of *and analysis* collocates include *description, data collection,* and *representation.*

It is important also to note that some lexical items with relatively high log likelihoods, such as *data analysis* and *factor analysis*, do not occur uniformly across the four disciplines of the corpus. *Data analysis* does not occur in the Law subcorpus and *factor analysis* does not occur in the Law or Science subcorpora. Furthermore, *factor, data* and *analysis* all occur in Sublist One of the AWL, meaning that they are all high frequency items in

the list. Much more work needs to be done to find out more about collocations in academic texts and their distribution across different academic disciplines (Durrant 2008: 159; see also Hyland 2008; Hyland and Tse 2007).

3.1.3 An assessment of assessment

We include *assessment* here as another example of an AWL word with interesting patterns in its data. The noun *assessment* is much the most frequent member of the *assess* word family, making up very close to half of the occurrences of the family. The strongest relationship shown by log likelihood is the lexicogrammatical relationship between *assessment* and *of*. Very often the noun that names the area to be assessed is generic in meaning, for example *assessment of a firm's export success*, *assessment of ambivalence*, *assessment of beneficiary income*, and *assessment of change*. Even when the following noun phrase uses the definite determiner *the*, the meaning is clearly general rather than specific, for example *independent and professional assessment of the child's position in her home*. As this example suggests, *assessment* is often involved in the long complex noun phrases fundamental to academic prose with several words/phrases used to give details about the type of assessment coming before the word and several words/phrases coming after *of* to characterize the thing being assessed.

The use of the coordinating conjunction *and* with *assessment* creates highly frequent pairs with assessment in particular relationships to other parts of a process: NP(s) *and assessment* or *assessment and* NP(s). The ordering of the combination reflects the place of assessment in a larger process: *objectives, content, delivery, and assessment* [assessment comes at the end of the process] but *environmental assessment and protection* [assessment must come first in the process]. Thus appropriate use of the word *assessment* will involve packaging it in a longer phrase that specifies the nature of the assessment and might involve having *assessment* as a sub-component in a longer process.

The types and purposes of assessments in academic text are suggested by the adjectives or nouns that come before the word: for example, *accurate, appropriate, arbitrary, careful; brief, broad-based; competency, impact, internal; project*, and *employee* along with a few others. While none of these words is used frequently enough for the combination with *assessment* to rise to statistical significance, there is a definite pattern of use. Some adjectives specify the quality of the assessment with an emphasis on characteristics that make for a good assessment. Other

adjectives give the extent of the assessment. Another set of words describes the time sequence possible for assessment. The fourth set specifies the type of assessment (*competency, skill*). A fifth combination focuses on the group or entity to be assessed (*employee, environmental*). A very small sixth group gives the assessor (*feminist, self-*).

Like the other examples given above, log likelihood analysis shows a strong relationship between *concept* and *of* and also with *the* and *concept*, suggesting a longer chaining with *the concept of*. Another pattern with *concept* is the string NP + *of the concept of* + NP, for example *analysis of the concept of unconscionable conduct, definition of the concept of performance, examination of the concept of justice.* The string begins with an abstract noun that refers to some action or process applied to the concept. The phrase ends with the naming of the concept. However, two interesting features of this word involve its close relationship with *the* and its use as what Schmid (2000) has termed a shell noun. Because *the* is so frequent in English, it forms many combinations with nouns that are not statistically significant but the result of the numerous uses of *the*. The data on *concept* show that this word bonds tightly with *the*, when *concept* is used then we can expect that the word *the* is highly likely to be included in the phrase. Of the 594 uses of *concept*, 309 are the collocation *the concept*. Of those 309, 251 are of the collocation *the concept of*. The use of *concept* as a shell for complex concepts can be seen in the following sequence of sentences [numbers added to make reference to the sentences easier for the reader]:

1. The results shown here suggest that an alternative **mechanism** could account for the beneficial effects of nicotine in AD.
2. **This mechanism** involves a nicotine inhibition to amyloidosis.
3. A major weakness in **this concept** relates to the lack of physiological data to support a role for nicotine.

The chain involves “the results” that lead to “an alternative mechanism” which is defined in the second example sentence. The whole definition is then packed into this concept so that the writer does not have to repeat the definition again in the third sentence.

The words associated with *concept* fall into two large categories. In the first set, the words are used to describe and evaluate the concept: *associated, base, basic, broad, broader, central, cyclical, dangerous, definitive, descriptive, detailed, different, difficult,* and others. In the second set, the combination points to the source of the concept or the

special area in which it is used: *anthropological, Anglo-American, astrological, economic, ecological, ideological, legal,* and a few others.

3.1.4 The case of *indicate*

While most word families (and lemmas) have one member that is much more frequently used than other members of the group, some word families include two (rarely more than two) members of almost equal frequency. Since *indicate* occurs 445 times and *indicated* 405 times, we decided to include both forms in our study. The strongest log likelihood relationship in the data for both forms is with *that*. Thus, there are both present tense and past tense sentences with *that*-clause complements:

(a) Present tense: Between-site variations indicate that the sample sites primarily reflect local vegetation patterns.
(b) Past tense: Estimates based on dilutions indicated that the final concentrations would be around 400ppm.

However, *indicated* is often used for a passive meaning as a passive participle or in a passive verb phrase. The importance of the passive uses of *indicated* points to an implication of the data. If the strongest relationship for *indicate* is with *that* so that the most common complement of *indicate* is a *that*-clause, where do all these passives come from? One answer might be that *indicate* has two patterns: a version that uses *that*-clauses as its complement and a transitive version with a noun phrase in the object position. The transitive version can have a parallel passive formation; the version with a *that*-clause is unlikely to have a parallel passive version. A detailed study is needed to tease out the differences between versions of *indicate* that can be seen as related to the passive and versions of *indicate* that do not have parallel passive forms.

The examples in this section illustrate the strength of collocational patterns with some AWL words and some of the kinds of patterns we see in the data. We move on now to some examples of AWL words that do not seem to have the same strength of patterning.

3.2. The weak

In this section, we will investigate *ongoing* and *straightforward* as examples of words with collocations with weaker associations

3.2.1 The case of *Ongoing*

A highly frequent word like *create* is likely to have a longer list of statistically significant collocates than a low frequency word like *ongoing*. *Ongoing* from Sublist 10 occurs only 97 times in the AWL corpus (compared to the 365 times for *create* and 1747 times for *analysis*). *Ongoing* has a few strong collocational relationships, most of which are grammatical rather than lexical. Durrant (2008: 163, following Gledhill 2000) comments on the usefulness of reporting grammatical and lexical relationships. Because *ongoing* is used in academic prose, the noun phrase in which it is used is likely to be on the long side with prepositional phrases attached using *of*. Examples of the noun phrases in which *ongoing* is used include the following:

> [They] had little commitment to the kind **of ongoing relationship** that might make mediation both more suitable and more effective.

> The telephone script in general demands polite exchange **and ongoing** conversation.

> [The university] provided fellowship assistance for **part of the ongoing** research on this project during a period of sabbatical leave.

The phrase *part of the ongoing* is used six times out of the total of 97 uses of *ongoing* in the AWL corpus.

3.2.2 *Straightforward* and *required*

While some words just do not have strong collocational relationships, others have strong ties to words to one side or the other, but not on both sides. For example, *straightforward*, from sublist 10 occurs 86 times in the AWL corpus. Ties to left that show this word as a predicative adjective (as the complement of a linking verb) and as an attributive adjective (coming before a noun) as in *is straightforward* and *a straightforward*. *Straightforward* has strong ties to the left, as in is/a/relatively/more. Words to the right are statistically significant but not as strong as the relationships to the left.

Another example of a word with strong ties to one side but not so strong on the other is *required*. As would be expected from other studies of academic prose, for example Biber's application of multidimensional analysis to academic data (1988) where passive voice is shown to be highly characteristic of academic prose, *required* is most characteristically

used in passive verb phrases. Both the statistical data about the collocations with *required* and the most frequent three-word phrases converge around this use. Thus, *required* is like *indicated* in the importance of passive voice uses both with and without a *by*-phrase. *Required* differs in its complements, having its strongest relationship with a *to*-infinitive clause and rarely having a *that*-clause. This difference seems to be a result of the heavy use of passive voice with an infinitive clause (599 uses out of the total 1377 uses of *required* or 44%) compared to the rarity of active voice followed by a *that*-clause (16 uses out of the 1377 or 1%):

(a) Passive + infinitive clause: the maximum amount of accrued interest a purchaser can be required to pay is 6 months.

(b) Active + that-clause: this objective required that governmental authority and administrative officials [do something].

The purpose of this section has been to illustrate how some words may have strong collocational relationships only on one side, rather than on both, while others may have only one main pattern to comment on.

3.3. The lonely

In this section, we consider the words *nonetheless* and *reluctant*.Word families in Sublist 10 just meet the frequency and range requirements for entry into the AWL. Analyzing their uses presents different challenges from those associated with highly frequent words like *analysis*. *Nonetheless* has weak collocational relationships based on the log likelihood statistic and equally weak patterning in the set of three-word phrases. The data do suggest what is confirmed by the concordance lines: *nonetheless* patterns like other transition words by tending to be sentence initial or immediately after a coordinating conjunction. However, about 40% of the uses come after the subject or some element of the verb (after *be* or the first auxiliary), a location that delays the making of the connection between the two sentences, for example *but they may nonetheless be documents that, in the opinion of Ministers, ought not….* While this delay might be a more sophisticated location for a transition word, the placement could be confusing for less skillful readers of English.

Lower frequency of a word is not a predictor of low collocational pull. For example, *reluctant* from Sublist 10 appears 67 times in the AWL corpus. While that is a large number compared to the many words that

appear only one or two times, *reluctant* certainly appears much less frequently than words like *analysis* or *approach*. Along with statistically significant collocations with 11 other words, r*eluctant* has a very strong grammatical relationship with *to*. *eluctant* appears 61 times with *to.*…that is, 91% of the uses of *reluctant* are with *to*.

Working with low frequency words also requires careful analysis of data presented by systems like Wordsmith Tools 4.0. The Mutual Information (MI) statistic provides information about relationships between words with the MI scores often getting higher based on the rarity of the words in a particular corpus. For example, *nonetheless* occurs 103 times and *purports* occurs 22 times and they occur together 2 times; their relationship in MI terms is a strongly significant 11.51 (most studies use 3.0 as the level of significance for MI scores). This information means that when *purports* is used, it is fairly likely that *nonetheless* will be used in the same neighborhood. Their log likelihood score is only 28.53 (compared to the log likelihood of 2232.99 for *analysis* and *of*). This number means that the two words are not highly characteristic of the discourse covered in the AWL corpus. That is, the MI score tell us that *nonetheless* and *purports* are fairly closely related to each other; the log likelihood score tells us that *nonetheless* and *purports* are not strongly characteristic of academic prose. These two approaches to statistics suggest the care with which data about language must be read. An MI of 11.51 is analyzed in terms of its distance from a 3.0 level of significance; this statistic tends to be stronger for low frequency words that are often used together. No level of significance has been agreed upon or even suggested for log likelihood scores; that data must be viewed comparatively so that a 28.53 log likelihood score is very low compared to the scores achieved by high frequency words like *analysis* and *of*.

4. Implications and considerations dor language analysis

The data from the AWL reported above show a number of similarities. The first is the strong log likelihood between four of the five words and *of*, as in *assessment of, analysis of, indication of* and *concept of*. In the case of *analysis*, *assessment*, and *concept*, the log likelihood is far greater with *of* than any other word. The strength of the connection indicates a highly productive collocation that forms the basis of some of the most frequent three word clusters, such as *the analysis of* and *an analysis of*. This finding connects the AWL words to studies such as Biber *et al*. (1999) that demonstrate the importance in academic writing of long complicated noun phrases with post-modification as well as pre-modification. The high

frequency of the word *of* also points to the studies that build from Sinclair's discussion how *of* differs from other words in the category preposition (Owen 2007; Sinclair 1991).

The data sets show that AWL words interact a great deal with GSL words (West 1953), something Durrant (2008: 164) also points out, and that AWL words also interact with each other. Long complex phrases, such as *the **domestic**-based **evidence indicates** a **link** between information use and performance* or ***data analysis** and **assessment methods*** (AWL words are in bold), often involve several AWL and GSL items. Even simple patterns in the data with coordinating conjunctions such as noun *and* noun often involve two AWL words, as in *analysis and assessment* and *analysis and interpretation*. Such interaction shows how important it is that words lists such as the AWL as not seen as isolated lexical items but are viewed in the common collocations and phrases in which they occur. Hoey (2005: 8), in his book on *lexical priming* states,

> As a word is acquired through encounters with it in speech and writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered, and our knowledge of it includes the fact that it co-occurs with certain other words in certain kinds of context.

By providing examples of actual use of collocations in academic context, we might learn more about the cumulative loading of words.

The near impossibility of making completely satisfying decisions about statistical approaches to analysis of language data must always be kept in mind by those carrying out studies such as the one reported here. To date, no standards have been agreed upon about reporting the settings that have been used in programs like Wordsmith Tools 4.0 so that replicating or verifying studies is made very difficult and often impossible[1].

Finally, numerical data about patterns in corpora are just one initial step in understanding language-in-use. After collecting information about statistically significant patterns, linguists need to return to the corpus to understand communicative purposes for the linguistic patterns. Statistical data about the words and phrases in a corpus can easily not reveal patterns that involve words with similar meanings. As shown in the analysis of *assessment* given above, study of the adjectives that appear before the word can reveal patterns of meaning and use that are not shown in the reports of statistically significant relationships. Thus, in addition to generating lists based on statistical analysis, we need to look carefully at concordance lines to seek additional patterns through sorting and re-sorting the concordance lines. This process belies the expectation that computer based analysis of language data will be fast and easy; at some

point in an analysis, a linguist must take a long thoughtful careful look at how words and their typical phrases are being used in context.

4. 1. Implications for language teaching

Students and teachers of English for Academic Purposes (EAP) need to work actively with the information about AWL words and their collocations. Students can be helped to achieve this goal by having opportunities to examine the words in the context of textbooks they are using currently or will be required to use in the future and then calling on the data we provide to see how else the target word operates in other contexts (i.e. Does a word change in any way in a new context? If you have the verb, what is the noun and how does it work in context?) For example, teachers and learners could investigate strong log likelihood relationships between words such as *indicate* and *that*. Not only does *that* score the highest log likelihood for the most frequent word in the family, *indicate*, it is also the highest with *indicated* (the second most frequent family member) and *indicates*. Frequency data show differences between words. Comparing the highest log likelihood relationships of the most frequent words in the AWL, including *indicate + that*, *assessment of*, and *required to*, would encourage direct focus on the words themselves, an activity that consistently shows itself to be valuable for the development of vocabulary knowledge (Nation 2001). The principle of learning high frequency words first (Nation 2001) can serve not only beginner and lower level learners who are focused on the higher frequency items in English, but also for the higher level, more academically minded learners who need to know which words, collocations, and phrases will give them good return for their learning effort and time. Another application of such data sets is a comparison of near synonyms that may cause interference for learners or between high frequency words that are generally well known, such as *need*, and its more formal, seemingly synonymous *require* (see Coxhead 2006).

Another important point is that these patterns allow teachers and learners focus on the usual rather than the unusual. By highlighting that *required* most commonly occurs in the passive and is followed by the infinitive, as in *to be required to seek approval*, teachers and learners can expand knowledge beyond the meaning of the word alone. This knowledge may allow learners to build a bridge between knowing the meaning of a word to knowing how to use a word in writing (see also Coxhead and Byrd 2007). Further, the precise use of such words as *require*

in an academic context demands higher attention and deeper knowledge than knowledge of the word in general or everyday use.

Teaching applications using data sets such as we present here must be mediated. Taking raw data and linguistic techniques into the classroom requires a great deal of care. It may the case that corpus-based dictionaries and grammars are a wise approach at this time, rather than bringing about a major shift in the language curricular if sets of words are the foundation for language use and vocabulary is the starting point for curriculum design. Not all teachers have access to computer-based tools, just as not all teachers excel at materials and curriculum design. However, all teachers should be skillful at the implementation of activities and lessons and need to understand the nature of language in use as background to teaching toward student needs. Keeping in mind Nation's (2007) four strands when creating a vocabulary programmes is one way to think about student needs. The four strands are meaning-focused input (learning through reading and listening), meaning-focused output (learning through writing and speaking), language-focused learning (deliberate study of aspects of words such as their pronunciation, spelling, meaning, and grammar for example), and fluency development. Nation suggests that equal opportunities for learning vocabulary are needed across all four strands. Hirsh and Coxhead (2009) apply the four strands to activities using their science-specific list for English for Academic Purposes.

5. Conclusions

A principled vocabulary list is a useful starting point for research and study. A word list is not the end point for teaching and learning but is the beginning. Lists such as the AWL are a probe, meaning it is a tool that students and teachers can use to notice lexical, lexicogrammatical, and semantic patterns in their own reading, and develop their knowledge and ability to being to use these words in their own writing. It is important for teachers and learners to understand that the goal is to find the words in their own contexts and to use them for their own needs. It is also important to know that not only are all words not created equal (as Paul Nation often says), but that not all collocational relationships are created equal. That is, some collocations might be the strong, others the weak, and somewhat sadly, some might just be the lonely.