Introducing

LeXITronics

A Research and Development Project for English Language Learning

Compiled and Authored by:

Ali Billuroğlu
John Eldridge
Nilgun Hancioğlu
Steven Neufeld

Eastern Mediterranean University
Famagusta
Cyprus

September 2008
PART A: OVERVIEW

1. A Brief History of Lexitronics

Lexitronics began when a small group of English language instructors at the Eastern Mediterranean University, Famagusta, Cyprus started to investigate word frequency lists. Many of their students had difficulty in communicating, despite years of study. It seemed obvious that the students lacked, above all, knowledge of vocabulary. To address this deficiency, the instructors began to research the generative power of the General Service List (West, 1953) and Academic Word List (Coxhead, 2000). They found that not only were there anomalies with the lists, but also numerous issues concerning how best to exploit them. Meanwhile new and emerging technologies were presenting opportunities that led to a number of interlinking initiatives, both research-based and practical, that formed the core of the Lexitronics project.

Lexitronics has both research and practical dimensions. The submission described in this summary for a British Council Innovation Award is entered under the Research Category. However, as the research base has substantial practical applications, these will also be described. Since the research is the basis of a published book and a website for learners and practitioners, we believe that this submission cuts across the formal award categories.

2. Principles of Lexitronics

Lexitronics is based on four fundamental principles:

i. The acquisition of an extensive and in-depth vocabulary is the foundation of successful language learning. Skills and grammar are secondary to the enterprise, since nothing can be communicated without vocabulary.

ii. Learning the most useful vocabulary is fundamental to success. Text and corpus analysis software enable us to pinpoint the most frequent words and phrases in both general and more specialized fields. It is essential to use this information. Most learners have limited time to reach an adequate level of proficiency in English.

iii. Technology is to be embraced within a sound pedagogical framework. The pace with which computers have developed as instruments of learning and collaboration has forever changed the traditional, classroom-based, textbook-orientated approach to learning. To what extent this change improves the quality of learning is dependent on creative and appropriate use of the technology.

iv. Self-directed and collaborative learning are key concepts. The development of 'techno-frames' for both individualized and peer learning is not only of direct benefit to the learner, but can help develop autonomy, initiative and flexibility. Web2.0 tools provide opportunities for social constructivist methodology at low cost direct to the consumer. Lexitronics is founded on a principle of community learning.

In short, Lexitronics is concerned with the teaching and learning of vocabulary, making the maximum use of the new technologies within a primarily collaborative environment.

3. Who is the Project aimed at?

This project began at Eastern Mediterranean University, an English medium university with students from beginners’ level who needed English for both general and academic purposes. The findings and approaches described however are relevant for all learners of English, and have implications for teachers, teacher trainers, and administrators.

4. Project Contents: Primary Research Components

The research described here is available in expanded form and accessible in either published form
and/or on the Lexitronics website. The portfolio of primary research and documents constituting the official contents of this submission for the British Council Innovation Award comprises the following pieces of work, listed in the order we would suggest that a reader might best study them in order to get the best feel for how the project has developed:

i. Summary of Award Submission, as attached.
ii. The Lexitronics Website at: http://lexitronics.edublogs.org, including a research overview, and slidecasts summarizing our research, as well as links to a corpus of academic abstracts, project outcomes and feedback and evaluation of our work.

5. Project Contents: Published Books with CD-ROM

From the research basis, the following supplementary course book designed for systematic vocabulary development has been published and will be submitted. The book comes with a CD-ROM containing exercises and a full listing of the Billuroğlu-Neufeld list (BNL) of the 2709 most common words in English.


6. Project Contents: On-Line Courses, Workshops and Other Resources

The research is the foundation of a series of electronic readers currently under development that attempt to facilitate an in-depth knowledge of the most common 2709 words in English in an electronic and collaborative environment.

Teachers and students can also freely profile any text using the BNL web-based vocabulary profiler of the 2709 most common words hosted on Tom Cobb’s lextutor website.

At the more advanced end of the learning spectrum, the research underpins a corpus-informed approach to the teaching of thesis writing at postgraduate level. This work has also been developed using an on-line platform and the corpus of thesis abstracts (The Target Abstract Corpus) compiled as part of the research included as part of this award submission.

The emphasis on technology within ‘Lexitronics’ has also led to the development of teacher training materials exploring the use of Web 2.0 applications in language learning.


(Version 2.3) [Software]. Available from http://lextutor.ca/vp/bnl


vi. Eldridge, J. and Neufeld, S. (2007) ‘When it Comes to e-Learning, are you in a Muddle or a Moodle’, in IATEFL CALL review, Winter 2007 is supportive of our work rather than a primary artifact, but gives a flavour of our use of the MOODLE learning platform.


The Lexitronics Website at: http://lexitronics.edublogs.org provides an alternative organized route through these various offerings.

PART B: RESEARCH SUMMARY

1. Research Overview

Our research comprises a series of overlapping projects that have as their common denominator an interest in lexical patterning and frequency in general and academic English. In short:

i. We argue that treating the General Service List (GSL) and the Academic Word List (AWL) as distinct constructs is not sustainable.

ii. We suggest there are good reasons for revising general lists of word frequency to ensure maximum utility to learners. We argue that such a list should incorporate the majority of the items in the Academic Word List (AWL)

iii. We describe the construction and testing of such a list – the Billuroglu-Neufeld list (BNL).

iv. We illustrate the difficulties involved in isolating specifically academic lexis and describe research indicating 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, rather than an add-on Academic Word List.

v. We argue for regular use of vocabulary profiling tools in the work of English language instructors. In this regard, we describe a case study into text readability, and show how the information derived from vocabulary profiling tools may be usefully exploited.

vi. We discuss the use of such tools in analyzing extant course books and materials and show that most commercial coursebooks still do not satisfy the lexical needs of learners, and that supplementary vocabulary work remains essential for most learners.

2. Deficiencies of the GSL

The General Service List (GSL) was developed by West (1953) as a list of the 2000 most useful ‘general service’ headwords and families for English language learners. Although doubts have emerged about its composition, it is nonetheless 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.
The following tables illustrate some problems with the GSL:

**Inconsistencies**

<table>
<thead>
<tr>
<th>GSL problem areas:</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>US / UK Spelling</td>
<td><em>Litre</em> is in the GSL, but not <em>Liter</em>.</td>
</tr>
<tr>
<td>Word Forms</td>
<td><em>Rise</em> is in the GSL, but not <em>risen</em>. Similarly, <em>hope</em> is in the GSL, but not <em>hopefully</em>. <em>Motherhood</em> is in, but not <em>fatherhood</em>. <em>Tour</em> and <em>tourist</em> are in, but not <em>tourism</em>.</td>
</tr>
<tr>
<td>Singular / Plural</td>
<td><em>Strength</em> is in. But not <em>strengths</em>. <em>Keepers</em> but not <em>keeper</em>.</td>
</tr>
<tr>
<td>Archaic words</td>
<td><em>Shilling</em>, <em>Hurrah</em> are in.</td>
</tr>
<tr>
<td>Not up-to-date</td>
<td><em>Radio</em> is in, but not <em>television</em>, <em>video</em>, <em>plastic</em>, <em>airport</em>, etc.</td>
</tr>
</tbody>
</table>

**Word family issues related to limitations of the vocabulary profiling tools and Bauer and Nation’s (1993) levels:**

<table>
<thead>
<tr>
<th>GSL problem areas:</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word families</td>
<td><em>Pride</em> and <em>Proud</em> treated as separate headwords.</td>
</tr>
</tbody>
</table>
| 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) 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’. |

The GSL contains residual problems. If refined, a stronger list should result. This was a primary objective of our project.

### 3. Anomalies of the AWL

Coxhead’s Academic Word List (2000) comprises the 570 most useful word families from her 3.5 million word corpus according to frequency and range of occurrence, excluding GSL words. Flaws in the GSL were thus likely to be accentuated in the AWL, and the assumption that high frequency words outside GSL coverage in the corpus would be ‘academic’ perhaps accounts for the ‘un-academic’ texture of some items. It has also been argued (Hyland and Tse, 2007) that the AWL is biased towards certain fields, such as economics.

**Examples of range of items in the Academic Word List**

<table>
<thead>
<tr>
<th>‘General’ items</th>
<th>‘Economics’ items</th>
<th>‘Academic’ items</th>
</tr>
</thead>
<tbody>
<tr>
<td>adult</td>
<td>currency</td>
<td>criteria</td>
</tr>
<tr>
<td>bulk</td>
<td>corporate</td>
<td>debate</td>
</tr>
<tr>
<td>couple</td>
<td>credit</td>
<td>define</td>
</tr>
<tr>
<td>drama</td>
<td>export</td>
<td>emphasis</td>
</tr>
<tr>
<td>injure</td>
<td>economy</td>
<td>evaluate</td>
</tr>
<tr>
<td>job</td>
<td>estate</td>
<td>illustrate</td>
</tr>
<tr>
<td>odd</td>
<td>finance</td>
<td>interpret</td>
</tr>
<tr>
<td>sex</td>
<td>fund</td>
<td>method</td>
</tr>
<tr>
<td>somewhat</td>
<td>invest</td>
<td></td>
</tr>
<tr>
<td>tape</td>
<td>levy</td>
<td></td>
</tr>
</tbody>
</table>

| adult            | currency           | criteria |
| bulk             | corporate          | debate   |
| couple           | credit             | define   |
| drama            | export             | emphasis |
| injure           | economy            | evaluate |
| job             | estate             | illustrate |
| odd             | finance            | interpret |
| sex             | fund               | |
| somewhat        | invest             | |
| tape            | levy               | |
In fact, there are no items in these categories that cannot be freely used outside academia. For example, *drama*, an AWL headword, behaves in a remarkably ‘un-academic’ fashion if the Cobuild Bank of English is consulted (HarperCollins Publishers, 2004), common collocates including *school*, *series*, and *TV*. It is *dramatic* that lends *drama* academic prominence through co-occurrence with such items as *increase* and *decrease*. What makes text ‘academic’, then, is not the occurrence in isolation of specific items, but the ways in which certain items co-occur with other lexical and grammatical items (Hunston, 2002, pp. 12-13).

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 may be inherently problematic.

It is true that the AWL has proved quite successful in identifying vocabulary used in academic discourse (Mudraya, 2006), albeit with disciplinary variation. However, it is telling that of the first one hundred items that Mudraya (2006) isolates as academic in terms of engineering English, ninety-nine appear in the British National Corpus list of the 3000 most frequent words in English.

### 4. Development of the BNL

To refine the GSL, Billuroğlu and Neufeld (2006) combined all words from a basket of commonly used lists into one list, which was filtered to obtain the unique terms. The 4,500 words in the resulting list were then ranked according to the number of lists they occurred in. The lists used were: the GSL headwords and word family members, the AWL headwords and most frequently occurring word family member, the first 2,000 words of the Brown corpus, the first 5,000 words of the British National Corpus, the revised version of the GSL (Bauman, About the GSL), the Longman Wordwise commonly used words (Longman, 2003), and the Longman Defining Vocabulary. The idea was that the combination of lists would highlight commonalities and isolate more singular and therefore questionable outcomes.

Previously, the GSL and AWL had been treated by researchers such as Coxhead as separate and distinct. The classic profiler on the [Lextutor web site](https://www.lextutor.ca) treats the GSL and AWL as two mutually exclusive lists. For profiling purposes, the GSL is divided into two bands of around 1,000 word families each, with the third band as 570 word families of the AWL. The assumption that the GSL and AWL could be treated separately for profiling ignored the fact that the words from both lists originated from the same frequency bands of common use. As can be clearly seen below, the BNC 5,000 most frequent words of English encompass the majority of items in the GSL and AWL.
Billuroğlu and Neufeld (2006), however, having combined the wordlists noted above, categorized the words according to the number of lists in which they occurred, producing a finer resolution in the form of six distinct, ranked bands. The list was named the Billuroğlu-Neufeld list (BNL) with each band assigned a BNL number, starting with ‘1’ for the most frequent.

The bands were formulated to represent an approximation of the natural vocabulary distribution in texts and provide data about differences in frequency between sublists. Further, the bands that emerged indicated that the AWL includes a large number of very common words.
Breakdown of the BNL

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

A medium-sized academic corpus of 730,000 words (Wordpilot, Academic Reports) was used to compare BNL and GSL/AWL text coverage. Using RANGE (Heatly, Nation & Coxhead, 2002), a vocabulary profiling software, the corpus was processed using the GSL/AWL lists and the BNL. The contrasting profiles below, both reveal Engels’ (1968) concerns about the lower coverage of the original set of K2 words, and illustrate the natural distribution of words in an English text, as used for the BNL banding system:

Profile contrasting BNL and GSL/AWL bands, and highlighting Engels’ ‘K2 deficiency’.

The BNL thus provides an economic and meaningful approach to vocabulary acquisition. It indicates that practitioners should consider putting aside academic word lists and focus on more commonly used words and unravel the contexts, varied meanings, etc., that will turn these words
into powerful tools of understanding and expression.

To illustrate, in this text, the reader might notice sub-technical items such as study, consider, describe, form, relationship. Yet none of these items are in the AWL and any practitioner choosing to teach from the AWL would risk ignoring such items simply because they have made a prior appearance in the GSL. Ironically, the AWL excludes word families not just because they are not frequent enough in academia, but because they are too frequent.

Readers can view the BNL sublists in full in a WIKI format (BNL, 2007) as well as test out the validity of these recommendations by visiting and using the BNL version of the Vocabulary Profiler (Cobb, 2007), and comparing results with the classic GSL and AWL version (Cobb, 2006).

5. A Corpus Informed Approach to EAP

Hyland and Tse (2007) suggest researchers develop corpora from specific fields and genres. Taking the genre of thesis abstract, Hancioğlu compiled two corpora of thesis abstracts, (target and learner) and used RANGE and CONCORDANCE (Watt, 1999) for analysis. The abstracts reflected the subject disciplines of nonnative postgraduate students taking Hancioğlu’s thesis writing course. The purpose of the Learner Abstract Corpus (LAC) was to detail the problems of the learners, and the Target Abstract Corpus (TAC) to identify patterns to enable the learners to conduct the required moves in their own theses successfully.

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

**GSL/AWL profile of key word examples from the TAC**

<table>
<thead>
<tr>
<th>GSL (K1) (46)</th>
<th>GSL (K2) (13)</th>
<th>AWL (85)</th>
<th>OFF-LIST (21)</th>
</tr>
</thead>
<tbody>
<tr>
<td>apply, base, build, case, change, character, consider, describe, develop</td>
<td>aim, collect, combine, compare, critic, discuss, examining, explore, govern, improve</td>
<td>analyse, design, research, project, process, culture, thesis, construct, theory, environment</td>
<td>dissertation, interview, objectives, organisational, collaborate, correlate, quantitative, reform</td>
</tr>
</tbody>
</table>

Using the BNL meanwhile with the same 165 word families produced a finer resolution in terms of bands, and, notably, fewer offist words:

**BNL profile of key word examples from the TAC**

<table>
<thead>
<tr>
<th>BNL1</th>
<th>BNL2</th>
<th>BNL3</th>
<th>BNL4</th>
<th>BNL5</th>
<th>BNL6</th>
<th>OFFLIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>base build case change character consider design develop process project</td>
<td>aim apply collect combine compare describe discuss environment improve organisational quantitative</td>
<td>Culture govern objectives research theory</td>
<td>critic explore</td>
<td>analyse construct examining interview thesis</td>
<td>reform</td>
<td>dissertation collaborate correlate</td>
</tr>
</tbody>
</table>

What gives these lists academic texture is not though the items in isolation, but their co-
occurrence in the same environment. It confirms again that there is little rationale for the
GSL/AWL division and indicates that teaching academic English needs to be based on analysis of
how items collocate and combine to achieve specified moves for specified purposes within
particular genres.

To take the example of specifying the objectives of a study and looking at patterns making use of
aim, the TAC reveals that the noun aim has three main lexico-structural realizations:

- 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)
- The (central / primary / main etc.) aim of this (thesis / dissertation, study etc.) is the
  (construction, design etc.) of ......
- 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. Aim as a verb occurs another 43 times:

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

Interestingly, aim is not in the AWL, but appears in the second (K2) band of the GSL, meaning it
too would be excluded from any programme taking the AWL as its sole basis.

To take another example, the data below provide extracts from the corpora concerning adjectival
and verbal collocates of study, the most common word in both the TAC and LAC:

**TAC – LAC—Excerpt from collocates of study**

<table>
<thead>
<tr>
<th>Adjectives</th>
<th>Verbs</th>
<th>Adjectives</th>
<th>Verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>archaeological</td>
<td>adds</td>
<td>case</td>
<td>aims</td>
</tr>
<tr>
<td>architectural</td>
<td>addresses</td>
<td>descriptive</td>
<td>applied</td>
</tr>
<tr>
<td>case</td>
<td>aims/aimed</td>
<td>documents</td>
<td>argues</td>
</tr>
<tr>
<td>cash-flow</td>
<td>analyzes/analyze</td>
<td>experimental</td>
<td>attempts</td>
</tr>
<tr>
<td>close</td>
<td>applies</td>
<td>field</td>
<td>based</td>
</tr>
<tr>
<td>comparative</td>
<td>argues</td>
<td>present</td>
<td>can / could</td>
</tr>
<tr>
<td>comprehensive</td>
<td>arises</td>
<td>time</td>
<td>consists</td>
</tr>
<tr>
<td>corpus-based</td>
<td>assessed/assesses</td>
<td>explores</td>
<td>explores</td>
</tr>
<tr>
<td>cross-cultural</td>
<td>assumes</td>
<td>focus</td>
<td>has</td>
</tr>
<tr>
<td>current</td>
<td>attempts</td>
<td>time</td>
<td>has</td>
</tr>
<tr>
<td>disciplined</td>
<td>began</td>
<td>indicates</td>
<td>indicate / indicates</td>
</tr>
<tr>
<td>ethnoarchaeological</td>
<td>claims</td>
<td>investigates</td>
<td>investigates</td>
</tr>
<tr>
<td>ethnographic</td>
<td>clarifies</td>
<td>is/was/are/were</td>
<td>is/was/are/were</td>
</tr>
<tr>
<td>experimental</td>
<td>combine</td>
<td>offers</td>
<td>offers</td>
</tr>
<tr>
<td>exploratory</td>
<td>compares</td>
<td>preferred</td>
<td>preferred</td>
</tr>
<tr>
<td>Field</td>
<td>complements</td>
<td>provides</td>
<td>provides</td>
</tr>
<tr>
<td>further</td>
<td>concludes</td>
<td>showed</td>
<td>showed</td>
</tr>
<tr>
<td>in-depth</td>
<td>considered/consider</td>
<td>will</td>
<td>will</td>
</tr>
<tr>
<td>independent</td>
<td>consisted/consisting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
With the TAC, verbs have been listed only up to the letter ‘d’. With the LAC, the same space has sufficed to list the entire repertoire of verb collocates employed by different learners over six academic semesters, amounting to a total of 20 different types, in comparison to 102 types used in the target corpus. With adjectives used to the left of *study* meanwhile, the LAC revealed the only 19 instances (7 different types), as opposed to 149 (45 different types) in the TAC. Simply put, the learners’ command of the word ‘study’, a GSL 1 word, is by no means as secure and as sophisticated as might have been assumed.

Where does this leave the wordlists? Interestingly, 92.76% of the items in the table above come from the BNL 2709 list, yet more evidence that practitioners and learners would well benefit from extending and deepening their knowledge of these items. Looking meanwhile at the LAC output for *study* and comparing it with the TAC output in the table below, 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. All the evidence indicates that compared with their peers in ‘native-speaking countries’, the non-native learners in this study have a marked deficit in productive knowledge not only of more specialized lexis but also of what has been identified as **general English lexis**. We suspect that this is not an uncommon syndrome.

**Comparison of frequency profiles of collocates of ‘study’**:

<table>
<thead>
<tr>
<th>BNL Cumulative Totals (%)</th>
<th>Learner</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNL-1 Words</td>
<td>56.67</td>
<td>32.08</td>
</tr>
<tr>
<td>BNL-2 Words</td>
<td>76.67</td>
<td>50.95</td>
</tr>
<tr>
<td>BNL-3 Words</td>
<td>90.00</td>
<td>58.50</td>
</tr>
<tr>
<td>BNL-4 Words</td>
<td>96.70</td>
<td>67.93</td>
</tr>
<tr>
<td>BNL-5 Words</td>
<td>100.00</td>
<td>86.69</td>
</tr>
<tr>
<td>BNL-6 Words</td>
<td>88.69</td>
<td></td>
</tr>
</tbody>
</table>

6. THE WORLD OF WORDLISTS: Some Methodology and Pedagogy

Wordlist driven approaches are not problem free (Folse, 2004). Lexico-knowledge is complex, and involves knowledge of forms, derivations, synonyms, antonyms, hyponyms, collocations etc. Indeed, the more one looks at what knowledge of words involves, the more it is obvious that knowing any word in depth involves knowing other words.

Discrete item frequency lists are however unrevealing of the subtleties of lexical phrases, multi-word units, and pre-formulated chunks. We may learn that *take* is a frequent item, but what about *take off, take over, take into consideration*? Polysemantic items such as *bow* and *row* present problems, as do primary and metaphorical usages, such as *driving into a ditch versus to ditch one’s girlfriend*. Repetitions meanwhile of referential items such as ‘it’ or ‘this’ raise another issue since surely what is of import is what is being referred to, not the repetition of the reference item itself.

We do not claim to solve all such problems here. However, we have explored the relationship between ‘knowledge’ of individual vocabulary items and overall text comprehension.

In their analysis of five texts, using Vocabulary Profiling software, Hancioglu and Eldridge (2007) concluded that overall comprehension cannot be directly equated with passive knowledge of individual constituent items, but rather may be expected to be 5-20% lower than indicated by an idealised knowledge of frequent vocabulary. The methodology used to reach this conclusion was based on subjecting texts to simple manipulations called ‘filters’.  

11
Basically, these filters replaced reference words with the words they referred to, and conjoined multi-word units (came within a whisker) along with prepositions or particles that seemed to belong grammatically to a specific lexical item (accused of). Articles, auxiliaries and functors that did not carry semantic or referential weight were removed. Many words have multiple meanings and when assessing text, it is hard to know how far knowledge of these multiple meanings would apply. It would be hard to know also if students could uncover secondary and metaphorical applications, and of course, issues of background knowledge need to be addressed. Would they, for example, be able to make sense of the notion of a ‘UK Independence Party’? Items that it was felt might pose comprehension difficulties were thus marked, to ensure they went ‘off-list’.

To give an illustration of how these filters worked in practice:

Original: Have you got a pet? You can learn how to look after it at the Pet Care Centre.
Modified after filtering: you got pet? You can learn how lookafter pet at Pet Care Centre.

By the time these negatively-acting filters had been applied, the Vocabulary Profiler was now yielding results 5-20% below the unfiltered texts.

However, some texts employ intra-textual definition, in which terms are explicated:

They say natural chemicals found in red wine appear to protect against blocked blood passages. The chemical substances are called polyphenols. They come from the outer covering of grapes.

Scanning these texts again, clearly defined lexis was assigned high-frequency status. The texts were also scanned for cognates that students would be able to recognise in their L1, in our case Turkish. In the extract above, for example, English: ‘passage’ maps into Turkish: ‘pasaj’. The average impact on the texts was a 4.5% increase in tokens understood.

One final operation concerned the approximately 15% of words in a given text which are off-list. How many unknown items does this residue equal? This depends on the length of the text, and the number of repetitions of individual items. The longer the text, the more words it will contain, and the more difficult it may become. The greater the repetition meanwhile, the more acquired knowledge of individual items will positively affect comprehension.

We therefore used an ‘imaginary teacher’, to ‘teach’ six unknown headwords before introducing students to our texts. These items were selected on the basis of frequency of use in the texts.

This done, and the six headwords ‘taught’, and assigned high-frequency status, the vocabulary profiles of the texts were processed with this additional filter in effect. The average impact of this ‘teaching’ process was a substantial increase of 7.49% in token knowledge.

Using such tools, it is relatively easy for teachers to extrapolate statistical data about texts and assess their difficulty. Not only does this indicate if a text is suitable for a particular level, it also gives information about what elements should be focused upon in the teaching-learning process and when. As noted, one of the most helpful approaches is to ‘pre-teach’ a limited number of the most frequently occurring unknown words in a text. This can have an enormous effect on comprehension and drive students towards the 95% threshold recommended by Nation (2001), whose research shows that guessing meaning from context is most effective when the reader knows at least 19 out of every 20 words. In a subsequent post-reading phase, one may then move to the isolation of frequent words, the further exploration and assimilation of which would then constitute vocabulary learning in the long term. Again, readers are invited to test the practicality of these suggestions by using any of the profiling tools at: http://www.lextutor.ca/vp/.
Alternatively, a similar outcome can be obtained without the statistics, but rather in a visual map, by pasting text into the word mapping tool at: http://wordle.net. This tool generates a ‘word cloud’ that can be used as an easy reference for teaching and learning:

Word Cloud for a small academic corpus, generated by WORDLE:

---

7. Application of Lexical Principles in Commercial Course books

Most course books published these days purport to have a strong emphasis on lexis, derived in many cases from a corpus informed approach. Part of our project has been to test such claims. Our work in this area has been informed by commentators such as Coady (1997), who argues that in order to acquire **vocabulary incidentally**, learners need a core vocabulary to begin with.

Research has shown that students who successfully achieve band 5.5 in IELTS know at least 1,650 out of the 2,000 most commonly used words in English (Neufeld, 2008). At this threshold the **‘fast mapping’ principle** illustrated mathematically by McMurray (2007) takes effect, which is reflected by an extended study of the same students which shows that they **actually** know over 6,000 out of the 10,000 most common words. However, if students fall even 100 words short of the 1,650 word threshold, the McMurray model suggests that fast mapping is disabled and the chance of significant language development minimal. Cobb (1995) cites a threshold of 1,500 words, below which Arabic students in a university preparatory program failed the PET exit level. Billuroğlu’s unpublished study of remedial students at Eastern Mediterranean University yielded similar results. These students knew about 1,300 of the 2,000 most commonly used words and only about 3,500 of the 10,000 most commonly used words. In other words, a deficit of 300 of the most commonly used words meant that fast-mapping couldn’t take effect and as a result these students were 2,500 words short of the 6,000 word threshold.

That proficiency is defined by lexical competence is argued by Nation and Waring (1997), Nation (2001) and Meara (1995). Cobb (1995) analyzed three major course books and despite purporting to have a ‘lexical’ manifesto, all fell short of covering 1,500 of the most common 2,000 words. Thirteen years later, little has changed. The table below analyses the SUCCESS coursebook published by Longman and shows that while the authors have attempted to incorporate a ‘lexical approach’, the needs of the students are not actually met.
Lexical Profile of the SUCCESS Coursebook

<table>
<thead>
<tr>
<th></th>
<th>Beginner</th>
<th>Elementary</th>
<th>Pre-intermediate</th>
<th>Intermediate</th>
<th>Common words in SUCCESS word lists</th>
<th>Common words missing from SUCCESS word lists</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNL0</td>
<td>36</td>
<td>20</td>
<td>5</td>
<td>1</td>
<td>62</td>
<td>31</td>
</tr>
<tr>
<td>BNL1</td>
<td>194</td>
<td>255</td>
<td>102</td>
<td>36</td>
<td>587</td>
<td>79</td>
</tr>
<tr>
<td>BNL2</td>
<td>73</td>
<td>149</td>
<td>98</td>
<td>43</td>
<td>363</td>
<td>141</td>
</tr>
<tr>
<td>BNL3</td>
<td>46</td>
<td>87</td>
<td>66</td>
<td>55</td>
<td>254</td>
<td>186</td>
</tr>
<tr>
<td>BNL4</td>
<td>26</td>
<td>55</td>
<td>52</td>
<td>65</td>
<td>198</td>
<td>233</td>
</tr>
</tbody>
</table>

Of the first 2,000 commonly used words of English, only 1,400 appear in the word lists in the SUCCESS course. These findings are consistent with the research conducted by colleagues at Eastern Mediterranean University on other coursebooks, notably, Headway (OUP) and Pathfinder (Longman), confirming that lexical deprivation is not limited to one publisher or course book (A. Billuroğlu, E. Arkin, personal communications).

So what does this imply for instructors and institutions? Almost certainly, that they must find alternative means of driving up the lexical thresholds of their learners. For example:

i) Acquisition of lexis must be assessment driven

The use of BNL2709 at Zayed University, Abu Dhabi and the National Autonomy University in Mexico City, indicates that systematic vocabulary development is greatly enhanced through a positive backwash effect from an assessment driven approach. (K. Caldwell, personal communication, May 21, 2008).

ii) Assessment must continually and upwardly recycle lexis

The experience gained in the use of BNL2709 in Abu Dhabi and Mexico indicate that assessment of lexis has to be continually recycled throughout and across levels (F. Villaneuva, personal communication, May 24, 2008). The lexis in a previous level must be tested in subsequent levels, with opportunities to recycle basic meanings as well as meanings/collocations/lexical frames reflecting more depth of knowledge.

iii) Essential lexis not covered in course books must be dealt with through supplementary materials.

The wide range of topics covered in many course books provide little opportunity for students to explore target vocabulary to the depth required. It is essential to provide supplementary vocabulary development material.

PART C: PUTTING IT INTO PRACTICE

In this section, we will describe projects and products based on our research:
1. **Introducing Learners to the 2709 most frequent words in English.**


   i. BNL2709—the essence of English (ISBN 9789944968133), has sold over 10,000 copies in Cyprus to date. The book provides around 180 hours of self-directed learning for the 2,709 most commonly used words in English, and includes a three-stage learning system integrated as part of the companion CD-ROM. For the past two years, the English Preparatory School, Eastern Mediterranean University has based its lexical syllabus on BNL2709. BNL2709 has also been used in a freshmen ELT course and is on the supplementary book lists of five other universities in Cyprus. Pilot projects have taken place to test the Arabic version of BNL2709 at Zayed University, Abu Dhabi. A second project is underway at the National Autonomous University of Mexico to test the BNL2709 approach with Spanish speakers.

   ii. For practitioners, the principles underlying the BNL are the subject of a series of slidecasts, available at: [http://bnl2709.edublogs.org/colour-of-words/](http://bnl2709.edublogs.org/colour-of-words/)

   iii. Any text can be BNL profiled for free at: [http://www.lextutor.ca/vp/bnl/](http://www.lextutor.ca/vp/bnl/) (Part of Tom Cobb’s Compleat Lexical Tutor Site)

   iv. For self-directed learning, the BNL is integrated into a MOODLE for preparatory school students at [http://ied.emu.edu.tr/ehouse](http://ied.emu.edu.tr/ehouse)

2. **Learning Vocabulary in context and in depth: The E-Reader Series**

   The E-Reader series is being developed on the MOODLE platform and constitutes an attempt to embody our research outcomes within a medium of maximum use to learners. In this project, our primary aim is to facilitate **in-depth learning in context of the 2709 words on the BNL list.**

   The learning of vocabulary in-depth has been a continuing theme in our research. Such in-depth knowledge involves knowing meaning and basic forms, collocations, connotations and register, synonyms, antonyms, hyponyms, meronyms and collocates; it involves being able to use the word appropriately, knowing about cultural accretions, how the word is grammaticalised, and so on. Knowing a word, in short, means **knowing other words.** We have seen that contemporary text books provide only partial solutions to this problem. The main principles of the E-Readers are:

   i. **The E-Readers are intended to develop knowledge of the 2709 most common words in English.** The aim is for learners to acquire a thorough productive knowledge of all 2709 families – which constitute over 90% of running text in English.

   ii. **The E-Readers are intended to provide sufficient exposure to lexis.** Research has suggested that learners need 10-12 encounters with a word before it is absorbed. It is an operating principle to provide this number and drive students as close as possible to the 95% threshold for lexis to be acquired through such skills as inference (Nation, 2001).

   iii. **Lexis is studied in-depth.** By making use of such tools as on-line glossaries, wikis and online activities each of the target word-families is explored in detail.

   iv. **The E-Readers are the foundation of an E-Community.** Use of interactive software allows not only for materials delivery but for learners to discuss the material, interact, and in so doing productively use target lexis.

   The approach is not based on classical conceptions of level but on the observation that many students following traditional modes of learning have a lexical deficit. The E-Readers are designed to remedy this. Having said this, the readers are not designed for beginners, and it is anticipated that the e-reader concept and materials might best be marketed through a licence enabling institutions to change, adapt, and use the basic package according to their own needs.

   In this respect, the e-reader series is a flexible package, as well as being an embodiment of the
lexitronic principles discussed thus far. It may be viewed at: http://telcas.org/ereader.


The Advanced Thesis Writing programme at the Eastern Mediterranean University represents an attempt to realize the principles of lexitronics at a more advanced level. The course started as a classroom-based course with a lexical approach to teaching thesis writing skills. Since then however, it has made extensive use of a corpus-informed approach and concordancing software as well as being adapted to the MOODLE platform, where it is anticipated it can be developed into a stand-alone course of considerable value to non-native speaking academics. The current version, a hybrid classroom and Moodle-based programme can be viewed at: http://telcas.org/moodle along with the corpus of academic abstracts on which much of the programme is based.

4. Teacher Training, Orientation and Workshops

A number of workshops and short courses have been designed around the research, ideas, and tools discussed in this summary. The reader may explore these in more detail on the Lexitronics Website, but included are materials on:

i. Integrating BNL2709 into MOODLE-assisted hybrid or fully-fledged online courses
ii. Data-driven learning
iii. Corpus-informed approaches to materials development and syllabus design
iv. Practical ways of integrating Web 2.0 tools and instruments into the EFL classroom
v. Differentiating learning and community learning

5. Tools and Technology

As the label 'lexitronics' implies, our project makes extensive use of computer technology, in particular Web 2.0 applications. Of these, MOODLE should be singled out for special attention.

Although the new technologies are not a central aspect of our research, they are essential to our approach. We believe that Web 2.0 technologies in particular provide a platform for a collaborative and peer-based learning methodology with extraordinary potential to generate not only receptive but productive learning. With the assistance of appropriately designed 'techno-frames', it is our belief that learners can be driven up and beyond key learning thresholds, particularly within the key domain of vocabulary acquisition.

PART E: FEEDBACK AND EVALUATION

Our approach to feedback and evaluation is ongoing and dynamic. It is also necessarily multi-dimensional since the project involves both research and practical elements. As far as the research package is concerned, and since it is in the research category that we are entering this application, we have tried to adopt the following approach and principles:

i. All our practical projects should derive from a sound research basis.
ii. Therefore, as much as possible, our research should be published and open to scrutiny and response, both through peer review and response.
iii. Publication in serious journals represents positive feedback.

Direct feedback to our ongoing research work has therefore come in the shape and form of:
i. Publication of the research background to the BNL, and the BNL Vocabulary Profiler on Tom Cobb’s Compleat Lexical Tutor Site.

ii. Publication after extensive feedback of the main body of research in article form in English for Specific Purposes, and also in ELTJ.

iii. A response to some of our ideas in TESOL Quarterly, 42(1) by Ken Hyland and Polly Tse in a reader exchange.

iv. The adoption of the ELTJ article as course material on both the Eastern Mediterranean University and University of Hertfordshire MA programmes.

v. Correspondence, questions, and comments from professional peers, as well as invitations for further publications (including an article on e-readers for Language Learning and Technology), and conference and workshop presentations at a number of Cypriot and Turkish Universities. Steve Neufeld was also invited to participate in the 2007 American Association of Applied Linguists colloquium on large scale vocabulary projects regarding the use of BNL2709 as a lexical syllabus at EMU.

vi. Evidence of fast mapping supplied during BNL piloting in Mexico.

Feedback on the practical elements of our project includes meanwhile:

i. An archive of student responses to the BNL at: http://thebnl.org/archive/responses.doc

ii. Sales of over 10,000 copies of the BNL course book in Cyprus and Turkey, and its adoption onto syllabus lists.


iv. Karen Caldwell's feedback on piloting of the BNL in Abu Dhabi.

A feedback and evaluation section has also been included on the Lexitronics Website.

PART F: SOME CONCLUSIONS

1. Summary of findings

The essence of our findings, both theoretical and applied, may now be finally distilled:

i. Treating the Academic Word List as an ‘add-on’ to the General Service List creates irresolvable theoretical and practical problems.

ii. A revised wordlist – BNL 2709 – has been developed that provides greater text coverage with more discrete frequency bands, of greater use to teachers and learners.

iii. Given difficulties in identifying ‘academic’ lexis, a model is proposed in which lexico-structural banks are used for advanced and specialized language study taking as a prime focus lexical patterns within specified genres and moves.

iv. Vocabulary profiling tools are of value not only to researchers, but to teachers and should be standard equipment, used to underpin a systematic approach to teaching vocabulary.

v. Current course books largely fail to cover lexis to an adequate range and depth and need to be supplemented by materials with an intensive lexical focus. The E-Reader project and BNL materials aim to provide multiple contextualized encounters with lexis that will enable learners to pass the critical thresholds required to succeed with language learning.

vi. Web 2.0 applications provide opportunities for lexis to be studied more productively. Within the E-Reader project, for example, learners are encouraged not only to look and study but use the target lexis in a collaborative environment.

vii. The collaborative environments described are not merely faddishly humanistic. They multiply learning opportunities, are accessible, economic and more attuned to meeting the
needs of individual learners. They represent a revolution in learning.

2. Work in Progress: Where next?

In this section, we briefly describe the current state of play regarding Lexitronics projects:

i. We are nearing completion of a 'BNL for Beginners'--titled BNL709--as a 'primer' for the high school prep year. This covers the high frequency 1,000 words of English and is designed for the younger learner.

ii. A Common University Language Framework at EMU is being developed using a corpus-informed approach delimited by the BNL2709 word list. See http://bnl2709.edublogs.org/2008/04/22/a-common-university-language-framework/

iii. We continue to develop our e-reader series at: http://telcas.org/ereader

iv. The corpus-informed Advanced Thesis Writing programme is being developed as a fully online course at http://telcas.org/moodle

v. Workshops about vocabulary profiling have been requested for in-service training at a number of universities in Turkey and Cyprus as well as at upcoming conferences.

vi. As the projects take root, it should be possible to gather more data assessing to a greater extent the ways in which learners benefit from the practical aspects of the project.

3. Summary of Award Application

It has been fashionable for a while to bemoan the lack of attention paid to vocabulary learning over the years, and many commercial products and courses claim to have taken corrective measures. It is our argument however that the measures that have in many cases been cautious, conservative, and at best partial.

We believe that the artificial GSL / AWL separation is flawed, and the comprehensive BNL provides a better research and pedagogical vocabulary profiling tool. We further believe that if lexical approaches to language teaching and learning are to succeed, they must be far more intensive and far more in depth. Lexis cannot be considered a poor partner or even an equal partner to grammar and skills in the learning process. It is the essence of language learning, and should be the central hub around which all other elements revolve.

We hope that our research provides at least some stimulation for debate in this regard, and hope that the projects with which we have tried to put theory into practice provide some illustration of the potential of the approach.

________________________  __________________   __________________   __________________
A. Billuroğlu       J.Eldridge       N.Hancioğlu       S. Neufeld

September 2008
Bibliography


7743 words
POSTSCRIPT (December 3, 2008)

In this postscript, we provide some updated information regarding developments to our Lexitronics project subsequent to the submission of the original application to the British Council.

i) Publication of Through the Looking Glass and into the Land of Lexico-Grammar

Originally, the editor said it would take a year or so for our article to appear in print. However, they decided to publish it in their printed journal in the next issue of English for Specific Purposes. So, as well as being available online at http://tinyurl.com/5ldvcr, the article has appeared in the hard copy version of the journal:


ii) Publication Award from Eastern Mediterranean University

Award granted by the EMU Research Institute to N. Hancioğlu and J.Eldridge for publication of Through the Looking Glass and into the Land of Lexico-Grammar in a Social Sciences indexed journal.

iii) TWO abstracts accepted for the special "Technology and Learning Vocabulary" issue of Language Learning & Technology.

We were invited to submit abstracts by one of the guest editors of the LLT special Issue. The editors accepted our abstracts and have asked for a complete draft of our research articles by March, 2009.

The graded reader is dead, long live the electronic reader. (John Eldridge & Steve Neufeld)

This article will report the findings of vocabulary profiling and analysis of both coursebooks and graded readers and establish principles that will be used to revisit the graded reader concept and develop a series of 'eReaders' within a framework of corpus-based studies and Web 2.0 applications. Initial results suggest that the approach has the potential to help learners develop a thorough and productive knowledge of the most frequently used words in English.

The importance of being academic (Nilgun Hancioğlu & Steve Neufeld)

Many NNS postgraduate students find themselves ill-equipped lexically to present and publish their research. In this survey of the Advanced Thesis Writing course, we look at how the pedagogy has shifted towards a focus on lexico-grammar using a corpus-informed approach. Our findings show the positive impact of computer-mediated vocabulary support on performance in writing, and suggest a paradigm shift towards explicit application of data-driven principles in language development.

iv) Use of the BNL in the 'World of Reading'.

Joan Baker-Gonzalez and Eileen Blau, the authors of the World of Reading (an academic reading course series to be published by Pearson) have based their target vocabulary selection on BNL2709, rather than the GSL / AWL.

v) Publication online of the Learner Abstract Corpus (LAC)

Hancioğlu, N. 'Learner Abstract Corpus' available through the 'Advanced thesis writing skills' course at: http://www.telcas.org/moodle. The Learner Abstract Corpus (LAC) is a corpus of student abstracts written by students at Eastern Mediterranean University, and is used as an activity source, and research base for comparison with the abstracts written by students in English medium countries in the 'Target Academic Corpus', also available at the address above.
vi) **BNL Vocabulary Profiler updated with ‘EDIT-TO-PROFILE’ option.**
BNL Vocabulary Profiler at [http://www.lextutor.ca/vp/bnl/](http://www.lextutor.ca/vp/bnl/) now includes Tom Cobb's "Edit-to-Profile" facility which enables the user to directly edit text within the profiler providing a quick and efficient method for adapting and manipulating texts according to frequency-driven criteria.

vii) **CULF Project initiated at Middle East Technical University.**
The ELT Department at METU, North Cyprus Campus, has applied for a research grant to pursue in-service and pre-service teacher training in the use of a corpus-informed approach in language teaching. Steve Neufeld, a member of Lexitronics, has been asked to participate as one of the coordinators of the project.

viii) **Revised Lexitronics Principles**
- Learning the most useful vocabulary is fundamental to success.
- Vocabulary has its own grammar; grammar is vocabulary.
- The acquisition of an extensive and in-depth vocabulary is the foundation of successful language learning.
- Linguistic competence is defined by the depth of awareness and the use of lexico-grammatical patterns.
- Self-directed and collaborative learning are key concepts in language development.
- Technology is to be embraced within a sound pedagogical framework.
SUPPLEMENTARY MATERIAL

The Lexitronics Website

The hub of all the research and practical projects involved in the Lexitronics Project.

Homepage at:
http://lexitronics.edublogs.org

- Overview of the Lexitronics research as submitted for the 2009 British Council Innovation Award.

- The Colour of Words Slidecast giving background about the BNL.

- The full BNL 2709 list.

- The BNL Vocabulary Profiler

- The Target Abstract Corpus (TAC) Web Concordancer

- The Learner Abstract Corpus (LAC) Web Concordancer

- The Advanced Thesis Writing Course

- The E-Reader Series

- Links to Publications

- Evaluation and Feedback
The **BNL Vocabulary Profiler**

At: [http://www.lextutor.ca/vp/bnl/](http://www.lextutor.ca/vp/bnl/)

And available through [http://lexitronics.edublogs.org](http://lexitronics.edublogs.org)

Enables quick processing and analysis of texts according to the BNL 2709
The TAC WEB Concordancer
Available through:
http://lexitronics.edublogs.org

- 600 abstracts
- 174,093 words of running text
- Fields of Arts and Humanities, Sciences, Social Sciences and Architecture represented equally.
If you would like further information about this project or would like to have hard or soft copies of any of the research referred to, we would be happy to hear from you at:

lexitronics@gmail.com

This research summary is also available on our website for those who would like to read from the soft copy in order to follow the hyperlinks in the text.

http://lexitronics.edublogs.org

Ali Billuroğlu is an instructor at the English Preparatory School, Eastern Mediterranean University, Famagusta, Cyprus. Ali also serves as a trainer for in-service teacher training in educational technology, and his current interests encompass systematic vocabulary development with various web 2.0 tools and instruments. He was born near Larnaca. He lived in England for many years before returning to live and work in Cyprus.

John Eldridge teaches at the Department of General Education in the Eastern Mediterranean University. He has an MSC in ELT from Aston University and an MBA in Educational Management from Leicester University. His current interests include semantic frequency and teacher autonomy. He was born in London.

Nilgün Hancioğlu also teaches at the Department of General Education, Eastern Mediterranean University. She has an MA in ELT from the Middle East Technical University, Ankara and is currently completing her PhD dissertation on a corpus-informed approach to teaching academic writing at EMU. Her research interests include academic writing, corpus studies, data-driven learning, and lexical semantics. Nilgün was born in Nicosia.

Steve Neufeld is a freelance researcher, trainer, and educational consultant. 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. Steve was born in Saskatchewan, Canada, and now lives in Cyprus.

We would like to thank the Eastern Mediterranean University, Famagusta, Cyprus, for its support during the course of this project.