Building Corpora and Compiling Pedagogical Lists for University Medical Students

Simon FRASER
Institute for Foreign Language Research and Education
Hiroshima University

Elsewhere in this journal (Davies et al., 2013), my colleagues and I describe the design and implementation of an intensive medical English course for third-year students in the Faculty of Medicine at Hiroshima University. A key component of this course was the Medical Vocabulary List, a 380-word glossary of the lexical items compiled from a small medical corpus and considered to be most relevant for the learners. We hope to build on the success of the course, and the vocabulary list in particular, by constructing a larger corpus of the most representative texts which will eventually inform the design of a lexically-based syllabus for medical students. To this end, it will be helpful to look at the work that has been done to date in the creation of corpora and word lists for learners of medical English. Although researchers such as Cowan (1974), Salager (1983, 1985), Baker (1988), and Chung & Nation (2003, 2004) have greatly increased our understanding of the types of lexis making up medical texts, there have been few attempts at compiling pedagogical lists of the most useful words. Recently, however, Fraser (2007, 2009) and Wang, Liang, & Ge (2008) have taken up the challenge by building corpora and providing word lists for pharmacology and general medicine respectively.

This paper details the development of a recent list: the Essential Pharmacology Word List (EPWL), which was introduced in Fraser (2012) in an examination of factors affecting the difficulty of specialized vocabulary. Pharmacology was initially chosen for investigation because I have knowledge of it at the tertiary level; also, pharmacology is an important university discipline in its own right. However, as we shall see from a comparison with the Medical Academic Word List (Wang et al., 2008), the lexical characteristics of pharmacology and medical texts have much in common. The Medical Academic Word List (MAWL) is an important new set of vocabulary items, developed from a corpus of medical English articles, and designed for learners of medical English. The paper concludes with a consideration of the implications of the findings for the design and development of future corpora, word lists, and materials for medical students.

CREATING AN "ESSENTIAL" WORD LIST FOR A MEDICAL DISCIPLINE

In Fraser (2009), investigating the lexis of pharmacology, a discipline at the core of medical science, a 2,000-item pharmacology word list (the Integrated Pharmacology Word List, or IPWL) was created which provides almost 90% coverage of a corpus of 100 pharmacology journal articles. This is a good result for a list of its size, and, like the similarly-sized General Service List (West, 1953) would not present students with an impossibly large number of words to learn. However, the findings suggested that still further refinements and improvements could be made,

and that it should be possible to create an even more specialized, focused, and efficient list. There are a number of reasons why this would be desirable:

1) Coverage of 73% is given by as few as 500 word families, while the next 1,500 words further increase coverage by only 19%.
2) The learning burden of a word list consisting of 2,000 word families may be greater than we realize; in the IPWL, this number of families equates to more than 7,000 individual words.
3) There are several words in the list which, although they occur with a relatively high frequency, are found in only a few of the articles in the corpus. For example, *renin* and *beverage* occur more than 100 times in the corpus, but they are each found in only three journal articles.
4) In addition to "pharmacology" words, the IPWL contains function words and other general words that will not (or, at least, should not) provide learners at the university level with problems.

Building a Pharmacology Corpus

Students of medical disciplines at university are going to need English primarily in order to read and write scholarly articles at the post-graduate level, and there is little necessity for them to be able to cope with English-language medical textbooks (although they may encounter some English material at the undergraduate level). It was decided, therefore, to create a corpus made up of research articles; these were taken from online pharmacology journals for convenience. Next, it was necessary to select which articles to use. Making this choice was not as easy as it might appear, especially with a subject like pharmacology, which is part of the much wider field of medicine and overlaps with many other closely related disciplines. Where, for instance, does pharmacology end, and toxicology or molecular biology begin? Should texts such as these be included in the corpus? Without having a knowledge of the field it is difficult to answer these questions. It is important to know that pharmacology is a subject that integrates knowledge of biochemistry, cell and molecular biology, physiology, and chemistry; that it is concerned with the therapeutic and toxicological actions of drugs on humans, animals, and microorganisms. Should toxicology texts, then, be included in the corpus? Yes: toxicology is very closely related to – indeed inseparable from – pharmacology.

The pharmacology corpus used in the present research consists of 100 recently published articles taken from a wide international selection of pharmacology journals available in electronic form on the Internet. The journals are published in several different countries, and use both British and American English. The areas of pharmacology represented include cardiovascular pharmacology, autonomic pharmacology, biochemical pharmacology, clinical pharmacology, alimentary pharmacology, and toxicology. In total, 41 different journals were sampled in compiling the corpus. Figures and tables were kept in, but abstracts and references removed. It was felt that a sufficiently representative selection was achieved; however, the choice was inevitably subjective, and it is possible that a different selection of articles, or a broader (or

indeed narrower) selection of areas of pharmacology, would have yielded different results. The corpus consists of a total of 369,000 words.

Table 1 shows the areas of pharmacology covered by the corpus, and the number of articles in the corpus representing each of these areas.

TABLE 1. Areas of Pharmacology Represented in the Pharmacology Corpus

Area	No. of articles
Alimentary/gastrointestinal pharmacology	4
Autonomic pharmacology	5
Biochemical/molecular pharmacology	10
Cardiovascular/vascular pharmacology	21
Clinical pharmacology	13
Endocrine pharmacology	2
Immunopharmacology	8
Neuromuscular pharmacology	3
Neuropharmacology/behavioural pharmacology	10
Pharmacokinetics	3
Renal pharmacology	2
Respiratory/pulmonary pharmacology	5
Therapeutics	8
Toxicology	5
Veterinary pharmacology	1

This is, of course, not the only way in which we can divide the field of pharmacology. Placing an article in a particular category is, in fact, a very difficult job given the many and complex ways in which the different areas interact. One might argue that cancer pharmacology or ocular pharmacology, for example, are not represented; however, articles dealing with cancer or diseases of the eye can be found in the "therapeutics" category. Also, "clinical pharmacology" is a very broad field indeed, with the articles falling under this label dealing with areas as diverse as cancer pharmacology and immunopharmacology. In some instances, a decision had to be made as to the particular area on which the findings of the study had a major impact; sometimes the journal in which the article appeared or even the title of the article would be a decisive factor when making this judgement.

Creating a Pharmacology Word List

To create the 2,000-word Integrated Pharmacology Pharmacology Word List (IPWL), frequency was the primary criterion, although the range of the articles in the corpus was also taken into account (see Fraser, 2009). This meant, for instance, that there were some words in the list which, although they occurred with high overall frequency, were found in only two or three articles; such words will obviously be less useful for learners to know than those occurring in more than half of the articles. The next step, therefore, was to narrow the focus of the list

still further by making wide range a requisite, and excluding function words and the general words that should not be problematic for university students.

Refining the List: Applying a Strict Range Criterion

Obviously, we would like the words in the list to occur in as many of the journal articles as possible. However, if the range requirement is too strict, the result will be a very short list indeed: a problem with a broad field such as pharmacology, which encompasses many different sub-disciplines. Even the most frequently occurring item in the IPWL —*et al*— is found in only 70 articles. In fact, the pharmacology word with the widest range is *activity*, which occurs in 91 articles; the most obviously technical word with the widest range (and also the highest frequency) is *receptor*, which, with a range of 66, is still found in only two out of three articles.

A decision, therefore, was made to set the minimum range at 15 articles. This figure may appear rather low for a corpus of 100 articles, but it seemed that a natural cut-off point existed at around this level: words with a lower range than this, with a few exceptions, did not intuitively appear to be fundamental pharmacology words. Words with a range of 15 included *catalytic, competition, degradation, macrophage, relaxation, syndrome,* and *symptom,* all of which were judged to be important in the field. There were some words with narrower range than this which I also felt would be useful for learners to know, but there were concerns about the size of the final list, and it would always be possible to go back later and adjust the range figure if necessary.

The IPWL was examined and any word family distributed across fewer than 15 articles was excluded. This resulted in a list of 982 words, about half the size of the original list. However, the aim was to create a list of "essential" pharmacology items, and to this end there were still a number of items which could be removed from the list.

Removing "Unproblematic" Words

It was obvious that the list contained a large number of words that are used in everyday language and are found in a list like West's (1953) General Service List (GSL). Careful examination revealed that there were two distinct types of general words, however: words with an additional, specialized meaning in pharmacology — words which can be labelled "cryptotechnical" — and those which are used in much the same way in both general language and medical science. The former group of words is obviously very important, as they are likely to be a source of confusion, and these were retained in the list. The latter category was deemed to consist of items that would not cause problems for learners at the university level in Japan, and words of this type were excluded. Admittedly, this inventory of "unproblematic" words is a highly subjective one, and is based on the researcher's knowledge of pharmacology and on the expectation that learners will be thoroughly familiar with the words in the GSL or equivalent list. It is possible, however, that some of the excluded words may actually have additional technical meanings, or be used in a special way in pharmacology texts that merits their inclusion in the list. Also, "easy" words may not be the same for students at different levels or in different

learning environments. It would, therefore, be of value to test students on their knowledge of such words, in order to investigate empirically the extent to which they are in fact unproblematic. The findings could inform future refinements of the list. Altogether, 411 words, including function words, were removed from the list (see Appendix 1 for the full list of words).

THE ESSENTIAL PHARMACOLOGY WORD LIST

After excluding all the word families found in fewer than 15 articles, and removing the "easy" general words from the list, we are left with a list containing 570 word families. This list, the Essential Pharmacology Word List (EPWL), is, coincidentally, exactly the same size as Coxhead's (2000) Academic Word List. This is fortuitous for two reasons: (1) the AWL is widely accepted by practitioners and learners as being a list of ideally manageable size; and (2) it enables direct comparisons with the AWL regarding coverage and the words contained in each list.

The frequency (with range, in brackets) of each of the headwords of the EPWL families is given in Appendix 2. The headword of each family is the most frequently occurring form in that family, which means that we can, at a glance, get a clearer idea of how each word is functioning in the corpus, and better determine what kind of word it is likely to be. For example, having *formation* as the headword makes it clear how *form* is most likely to be used (to describe the process of coming into being); *clinical* rather than *clinic* suggests that it is the diagnosis and treatment of a condition that is more important. If the *–ed* or *–ing* form of a verb is the most common, then that form is shown. Certain verbs are most frequently found in their past forms, either to describe experimental procedure (e.g., *conduct, demonstrate, monitor, reveal*) or a pharmacological process (*synthesize, induce*), and the headword list reflects that. Also, if an abbreviated form is the most frequently occurring form of a family, that was recognized by placing the abbreviation as the headword (e.g., *Cl* was more frequent than *chloride*).

Coverage Provided by the List

In order to get an initial idea of the usefulness of this 570-word, streamlined list, it is helpful to look at the coverage that it gives of the Pharmacology Corpus. The list's wider validity can also be investigated by testing its coverage of other corpora. We see from Table 2 that, for a list of its size, the EPWL gives extremely good coverage not only of the Pharmacology Corpus, but also of a 58,000-word pharmacology textbook (Neal, 2003) and of a 500,000-word corpus of online medical texts built using WebBootCaT software (Baroni & Bernadini, 2004).

TABLE 2. Coverage of Pharmacology Corpora by the Essential Pharmacology Word List

	Tokens (%)	Types (%)	No. of word families used
Pharmacology Corpus	26.73	13.18	570 (100%)
Pharmacology Textbook	24.28	19.51	486 (85.3%)
Pharmacology BootCaT	20.41	12.78	551 (96.7%)

Table 2 shows that 486 of the EPWL families (85.3%) are used in the textbook corpus, and as many as 551 families (96.7%) are used in the BootCaT corpus. The equivalent figures for AWL families used in the Pharmacology Corpus, Textbook Corpus, and BootCaT corpus are 532 (93.3%), 322 (56.5%), and 554 (97.2%) respectively (see Table 3).

TABLE 3. Coverage of Pharmacology Corpora by the Academic Word List

	Tokens (%)	Types (%)	No. of word families used
Pharmacology Corpus	9.43	8.61	532 (93.3%)
Pharmacology Textbook	6.58	9.81	322 (56.5%)
Pharmacology BootCaT	10.10	11.07	554 (97.2%)

The coverage given by the AWL is considerably less than that given by the EPWL, although the two lists are of comparable size. The AWL provides particularly poor coverage of the Textbook Corpus, and only a little over half of the AWL families are used. Interestingly, more AWL families than EPWL families are used in the BootCaT corpus (the reverse is true for the other two corpora), although the AWL provides only half as much coverage.

It could, of course, be argued that higher coverage would be expected from a list which includes General Service List words, which, by definition, are words that occur frequently. However, Table 4 shows that even when we exclude the GSL words from the EPWL, coverage of the Pharmacology Corpus is still over 19%, with coverage of the Textbook Corpus almost as good. The figure for the BootCaT Corpus is lower (13.23%), but the combined coverage of 80.66% given by the GSL and EPWL is actually better than that of the Pharmacology Corpus (80.45%).

TABLE 4. Coverage of Pharmacology Corpora by the GSL and EPWL Combined

	GSL (%)	EPWL (%)	Total (%)
Pharmacology Corpus	61.34	19.16	80.45
Pharmacology Textbook	66.04	16.98	83.02
Pharmacology BootCaT	67.44	13.23	80.66

Comparison with the Medical Academic Word List

At this point, in order to put these coverage figures into perspective, some comparisons will be drawn with the findings of Wang, Liang, & Ge (2008). Wang et al., by targeting medical science specifically, created a list for students of medical English which they hoped would address the shortcomings of the Academic Word List for this particular group of learners (medical texts were not included in Coxhead's corpus). Their word list, the 623-word family Medical Academic Word List (MAWL), was compiled from a corpus containing just over one million words of medical research articles from online resources. Wang et al.'s methodology has much in common with that used in the creation of the pharmacology lists, with wide range and high frequency being the most important criteria. However, there are two crucial differences between the EPWL and the MAWL: in the medical list, the word families all had to be outside the 2,000 words of the General Service List; also, believing that strictly technical terms are not a problem for learners of medical English, Wang et al. excluded any words they believed to fall into this category, such as *necrosis* or *hepatic*. It is hard to understand the justification for either of these decisions: current thinking does not support the view that vocabulary for specific purposes should (or, indeed, can) be separated into lists of general, academic, and technical words (see, e.g., Fraser, 2009). Learners need to be able to cope with all the words that they frequently encounter, and as Fraser (2012) shows, the different types of words found in a specialized text all pose their own particular challenges. These include the confusion caused by GSL words being used with a technical meaning, and difficulties with the pronunciation and orthography of fully technical words.

As Table 5 shows, the MAWL accounts for 12.24% of the tokens in the medical corpus. Coverage of a passage randomly selected from one of the articles in their corpus was a similar 12.13%.

TABLE 5. Coverage given by Wang et al.'s Medical Academic Word List

	Coverage (%)
Medical Corpus	12.24
Medical text passage	12.13

Even after we exclude GSL words (however unwise this may be), the EPWL still gives close to 20% coverage of the Pharmacology Corpus, and 17% coverage of the Pharmacology Textbook corpus. What is surprising, however, is that it gives coverage of 13.9% of the randomly selected passage from a medical research article in the Medical Corpus (see Table 6); this is even better than the 12.1% given by the MAWL itself. We see from Table 6 that this finding was not a singular occurrence, and it could be replicated with other medical articles. The "Medical Minicorpus" totalled 9,203 running words and was made up of the following three randomly chosen online medical articles:

Self-reported impressions of insulin detemir among patients with type 2 diabetes: insulin-naïve vs. prior insulin users (*Medicine On-Line*, 2008)

Fruit and vegetable intake and incidence of type 2 diabetes mellitus: systematic review and meta-analysis (*BMJ Online*, 2010)

Effect of unsupervised home based proprioceptive training on recurrences of ankle sprain: randomised controlled trial (*BMJ Online*, 2009)

TABLE 6. Coverage of Medical Texts by the GSL and EPWL Combined

	GSL (%)	EPWL (%)	Total (%)
Medical text passage	71.80	13.86	85.48
Medical Minicorpus	70.55	13.94	84.49

As we see from Table 7, the coverage of this medical corpus provided by the EPWL was 21.16%, and 13.94% when GSL words were excluded; both of these figures are consistent with the coverage given of the medical text passage. The corpora used are, of course, too small for us to be able to say with any certainty that the EPWL provides even better coverage of medical texts than the MAWL does, but the results certainly suggest that it is possible.

TABLE 7. Comparison of EPWL Coverage of Pharmacology and Medical Corpora

	Coverage (%)	Coverage (excl. GSL words) (%)
Pharmacology Corpus	26.73	19.11
Medical text passage	20.79	13.86
Medical Minicorpus	21.16	13.94

A CLOSER EXAMINATION OF THE WORDS IN THE LIST

Table 8 shows the 100 most frequent words in the EPWL, listed in order of frequency of occurrence. *Cryptotechnical* words (words with a hidden technical meaning), *lay-technical* words (technical words which will be familiar to the non-specialist), and *fully technical* words (those highly specific to the field) have all been identified (see Fraser, 2012 for further discussion of these categories of words). What is noticeable is how familiar almost all of these words will appear even to someone without specialist knowledge in the field. It should come as no surprise, then, that despite the fact that we have removed all function words and "easy" general words from the list, no fewer than 38 words in the top 100 are found in the GSL.

TABLE 8. The 100 Most Frequent Word Families in the Essential Pharmacology Word List
(Listed in Order of Frequency)

1. ET AL	35. BLOOD**	68. VARIABILITY*
2. EFFECT**	36. INDICATE*	69. PLASMA †
3. ACTIVITY*	37. TISSUE**	70. PERFORMED
4. CELL**	38. POTENTIAL*	71. RELAXATION*
5. RECEPTOR †	39. CHANNEL*	72. DEMONSTRATED
6. TREATMENT*	40. METABOLISM**	73. FORMATION*
7. INHIBITION*	41. ROLE	74. GENE**
8. CONCENTRATION*	42. CLINICAL**	75. SELECTIVE*
9. FIGURE*	43. TABLE*	76. MECHANISM
10. INDUCED*	44. INFLAMMATORY**	77. NORMAL*
11. PATIENT**	45. CURRENT*	78. OBTAINED
12. DRUG**	46. FACTOR*	79. COMBINATION*
13. DOSE**	47. SUBJECT*	80. METHOD
14. SIGNIFICANT*	48. BINDING*	81. SENSITIVE*
15. CONTROL*	49. INJECTION**	82. TRIAL*
16. EXPRESSION*	50. ACID**	83. CONTRACTION*
17. PRESENT*	51. AGONIST †	84. EXPOSURE*
18. DATA	52. ACTION*	85. LIVER**
19. PROTEIN**	53. NEURON †	86. OXIDATION †
20. CALCIUM**	54. RELEASE*	87. APPLICATION*
21. FUNCTION*	55. SYSTEM*	88. PERIOD
22. ANALYSIS	56. RISK*	89. DIABETES**
23. MODEL*	57. PREVIOUSLY	90. ENZYME †
24. EXPERIMENT	58. MEDIATED*	91. COMPOUND**
25. BASE	59. ANTAGONIST †	92. STATISTICAL
26. DEPENDENT*	60. CONDITION*	93. PATHWAY*
27. REDUCED*	61. RATE*	94. ENHANCED*
28. THERAPY**	62. REGULATION*	95. INVESTIGATION
29. TEST	63. SOLUTION*	96. ASSAY**
30. MEAN*	64. SAMPLE*	97. INDIVIDUAL
31. OBSERVE	65. INVOLVED	98. RANGE
32. ADMINISTRATION*	66. SIMILAR	99. ASSESSED
33. STIMULATION*	67. VALUE*	100. SITE*
34. ASSOCIATED		

*Cryptotechnical words
**Lay-technical words
† Fully technical words

Cryptotechnical Vocabulary

We should not, however, be deceived into thinking that learners will already know these "familiar" words; as can be seen from the table, many of them are cryptotechnical words, and will be "technicalized" — take on very specialized meanings — in a medical context. In Fraser (2009),

it was established that words with the potential to behave as technical terms overwhelmingly do so; cryptotechnical items like *control, expression, action, channel, sample*, and *trial* are almost always used with their technical meanings, and others such as *order* and *reduce* are used with both general and specialized senses. Although it was also ascertained that there are a few words, including *present* and *dependent*, which are only rarely used with a specialized sense, it is still necessary to know that they can indeed function in this way. In total, 51 of the words in Table 8 are cryptotechnical, which means that more than half of the 100 most frequent words fall into this category.

Lay-technical Vocabulary

Lay-technical words account for nineteen, or just under one in five of the words in the list: examples are *protein, blood, acid, liver, therapy, injection, clinical,* and *diabetes* (although it could be argued that the last of these is fully technical). Again, like cryptotechnical words, these may not be as straightforward as they first appear: the layperson's understanding of "acid", for instance, is rather different to that of a specialist (to whom it is a substance, with a pH of less than 7.0, that can neutralize an alkali).

Fully Technical Vocabulary

Although we are saying that the words in the list may present more of a challenge for learners than is initially apparent, it is surprising that there are so few overtly "difficult" words. In the list of the top 100 essential pharmacology words, it turns out that there are only six — or just over one in twenty — fully technical words. They are: *agonist, antagonist, enzyme, neuron, oxidation,* and *receptor*. Although these are words that are used almost exclusively in pharmacological or medical contexts, they do not perhaps meet our expectations of what a specialized medical word should look like — they will at least *seem* familiar to the layperson. In fact *enzyme, receptor* and *agonist*, as we shall see shortly, all find a place in Wang et al.'s list of academic/subtechnical medical items. While *enzyme* could perhaps be considered to be a lay-technical word, *receptor* and *agonist* certainly are not. We would not consider the latter two words to be cryptotechnical, however; although they do appear in contexts outside medical science, they are not used in general language, and in each case the biological meaning is the primary one.

Other than the words mentioned above, there are very few highly specialized words which occur at the highest frequencies. However, the further we go down the list (71 words in total), the more we find: *KCl, NaCl, epithelial, homeostasis,* $CaCl_2$*, intraperitoneal, biochemical, centrifuge, aliquot* and *phenotype* are the bottom ten entries. Still, words of this type will often be familiar, at least to some extent, to the non-specialist; they are, by no means, all the long, unpronounceable words that we might expect to find in a medical field, which should be of some reassurance to the teacher of English for specific purposes who lacks expert knowledge in the subject.

Academic Vocabulary

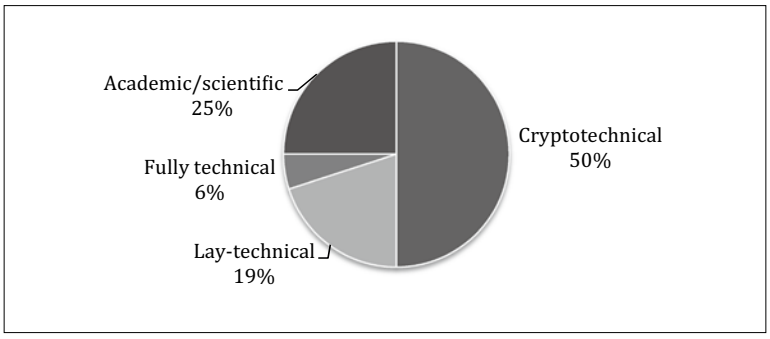
Let us now turn our attention to words that are neither fully technical, cryptotechnical, nor lay-technical. After accounting for the different types of technical words (75% of the total), the "non-technical" words that remain are important academic, or subtechnical, words which are necessary for reporting scientific investigation, analyzing results, and making comparisons. The following words fall into this category: *analysis, assay, assessed, associated, based, data, demonstrated, dependent, et al., experiment, individual, investigated, involved, mechanism, method, obtained, observed, performed, period, previously, range, role, similar, statistical,* and *test.*

This group of words contains items which are fundamental to any kind of empirical study: *data, experiment, statistical,* and *test,* for example. The fact that many of them are found most frequently in their past tense forms indicates that an important role of academic vocabulary is to describe experimental procedure (e.g., *assessed, investigated, performed*) or to analyze findings (*demonstrated, obtained, observed*). Words such as *analysis, data, experiment,* and *role* may also have an important discourse-structuring function by behaving anaphorically (referring back to a preceding stretch of discourse) or cataphorically (referring forwards).

Summary

Figure 1 shows at a glance the relative proportions of the different types of vocabulary making up the most frequent 100 words of the Essential Pharmacology Word List. It clearly shows the importance of both cryptotechnical and academic vocabulary — words that have traditionally been considered subtechnical — in pharmacology journal articles.

FIGURE 1. Breakdown of vocabulary types in the top 100 words of the Essential Pharmacology Word List



COMPARISON WITH THE MEDICAL ACADEMIC WORD LIST

In order to get some insight into the similarities between pharmacology and medical science as a whole, I would like now to compare the words in the Essential Pharmacology Word List with the equivalent items in Wang et al.'s Medical Academic Word List. This comparison will show how methodological differences in the creation of the two lists have important consequences for the final makeup of the lists.

Overlapping Items

Just a brief glance at both of the lists is sufficient to suggest that there will be considerable overlap between the EPWL and the MAWL, and this indeed turns out to be the case. Thirty-nine out of the 100 most frequently occurring words in the EPWL are also found in the MAWL top 100. They are: *cell, concentration, data, dose, induce, protein, function, analysis, therapy, indicate, tissue, potential, role, clinical, factor, acid, significant, stimulation, baseline, inflammatory, site, previously, mediate, involve, similar, variablility, demonstrate, gene, select, normal, obtain, method, exposure, liver, period, pathway, assay, individual*, and *range*.

Eighteen other words in the top 100 EPWL list are found lower down the MAWL list, although most of these still occur fairly frequently. These words are: *channel* (no. 108 in the MAWL list); *drug* (no. 122); *inhibit* (no. 102); *release* (no. 116); *calcium* (no. 147); *metabolism* (no. 202); *injection* (no. 207); *regulation* (no. 200); *plasma* (no. 169); *contraction* (no. 427); *enzyme* (no. 142); *statistical* (no. 362); *enhance* (no. 146); *investigation* (no. 133); *assess* (no. 119); *formation* (no. 195); *receptor* (no. 109); and *agonist* (no. 489).

In total, then, 57 of the 100 most frequently occurring words in the EPWL are also found in the medical list, which suggests that lexically, pharmacology and medicine in general have much in common. Most of these overlapping items are words that we might expect to be important in both fields, but there are some surprises. It is interesting, for example, that *previously* occurs with such high frequency in both lists; in medical and pharmacology research articles, this word must play a particularly important role in locating a piece of research in the context of work that has been done before.

Pharmacology Words Not Found in the MAWL

A sizeable minority of pharmacology words, however, are not found in the MAWL, and we will now turn our attention to these and see what they can tell us about the differences between the two disciplines. The following EPWL items do not occur in the MAWL:

et al, effect, activity, treatment, figure, patient, control, expression, present, model, experiment, dependent, reduce, test, mean, observe, administration, associate, blood, table, current, subject, binding, action, neuron, system, risk, antagonist, condition, rate, solution, sample, value, perform, relaxation, mechanism, combination, sensitive, trial, oxidation, application, diabetes, and *compound*.

The absence of *administration* is notable; it is found in the Academic Word List, and we certainly might expect it to be to be a very frequent word in medicine (e.g., *administration of drugs/medicine/treatment*). In the pharmacology corpus it is the 89th most frequent word, and occurs in as many as 65 articles, so it is surprising indeed that it does not register at all in the MAWL. *Compound*, too, is a word that we would predict to be in the MAWL: it is an AWL entry, and occurs frequently and with fairly wide range (44 articles) in the EPWL.

Et al is, by some distance, the most frequent item in the pharmacology corpus, but it does not appear in the MAWL. It, surely, is frequent in that corpus as well, but perhaps it was

excluded in the belief that it presents very little learning burden. It could be, of course, that Wang et al. simply did not include abbreviations in their list.

Oxidation is found in the top 200 items of the pharmacology list, and in 29 articles, and so is another unexpected omission from the MAWL. The term does, though, refer to an interaction at the molecular level, and perhaps that is why it is more likely to be found in pharmacology (which, we will recall, has very close links with biochemistry and molecular biology). The fact that *neuron* is not found in the MAWL could be for similar reasons: it may be particularly important in pharmacology when describing reactions at the cellular level, but there is perhaps less necessity for this type of description in general medicine.

Although the MAWL, as its name suggests, is intended to be a list of *academic* medical words, there are a number of words which could be considered to be strictly technical in the list. It would seem that decisions on whether to include a particular word in the list have been made on how difficult or technical a word *appears*, rather than on its actual degree of "technicalness". Perhaps this is why Wang et al., although they state that their list is an academic word list, apparently contradict themselves by including technical words such as *agonist*, *receptor*, and *plasma*. We do know, however, that it is often very difficult to draw clear boundaries between the different categories of words.

Wang et al.'s list, of course, excludes words from the General Service List, and most of the EPWL words that are not found in the medical list are GSL words. However, these missing words cannot be ignored; as we have seen, they are either cryptotechnical or important in some other way in medical science. The cryptotechnical words include *effect, activity, current, treatment, control, model, dependent, mean, administration, relaxation,* and *present*. There are words such as *figure, table,* and *subject*, which we have labelled cryptotechnical, but are perhaps more accurately "cryptoscientific", as they are essential in a wide variety of scientific texts. We also find lay-technical words: *patient, injection, blood*; and there are words which could be considered academic or subtechnical, although they are not in the Academic Word List: *experiment, test, associated, performed*, and *observed*.

There are, in fact, only five words in the top 100 of the EPWL which are not found in either the MAWL or the GSL, and they are *oxidation, diabetes, compound, mechanism, relaxation*, and *neuron*. It is surprising that *relaxation* is not found in the MAWL, as it is a very frequently occurring word with wide range in the pharmacology list. Oddly enough, *contraction* (which, in pharmacology, can be considered to form a pair of opposites with *relaxation*, as in *contraction/ relaxation of smooth muscle*) *is* found in the MAWL.

Specialized Pharmacology Words and the MAWL

Although *neuron* and *oxidation* may not be found, the MAWL unexpectedly contains several words that we might consider to be quite specialized pharmacology terms, even by comparison with the field of medicine in general. The following are all found in the MAWL: *receptor* (no. 109 ranked by frequency); *drug* (no. 122); *stimulate* (no. 154); *transport* (no. 296); *eliminate* (no. 395); *tolerance* (no. 450); *pharmacological* (no. 494); and *agonist* (no. 533).

It is hard to think of words which are more central to the field of pharmacology than *drug*, *pharmacological*, or *agonist*; the fact that these words, which are associated with fundamental pharmacological concepts, are found in a wide variety of medical journal articles is evidence that pharmacology occupies a position at the core of medical science. It is surprising, then, given that *agonist* is listed in the MAWL, that *antagonist* is nowhere to be found; these two words can be considered to be "of a pair" in pharmacology, and in the Pharmacology Corpus they occur with similar frequency and range.

What is interesting about these fundamental pharmacology terms which occur in the MAWL is that they include cryptotechnical words found in the Academic Word List (e.g., *stimulate, transport, eliminate*). In fact, the MAWL contains a large number of AWL cryptotechnical words, and these academic words will often be used with a medical meaning. GSL cryptotechnical words, though, are excluded completely from the list, although we have seen from the EPWL that they form a very important group of words. As a result of this, we find *affect* in the MAWL, for instance, but not *effect*; *effect*, though, is one of the most important words in the pharmacology list, and it surely also merits a place in the MAWL.

Cryptotechnical GSL Words

What we have been arguing, then, is that cryptotechnical GSL words have a very important role to play in medical science research articles, and they should be included in any word list that purports to be an "essential" or "academic" medical word list. The following list contains just some of the important EPWL words that can be found, used with their medical senses, in the Medical Minicorpus, but which are nowhere to be seen in the MAWL: *case, complications, composition, controlled, delivery, effect, fatty, history, markers, preparation, properties, risk, treatment*, and *trials*. On the other hand, we *do* find such apparently unproblematic words as *available, clinic, obtain, approach, goal*, and *seek*; they were, presumably, judged to be worthy of inclusion because they appear in the Academic Word List.

DISCUSSION

With the EPWL, which at 570 words is the same size as the AWL and smaller than the MAWL, the aim of creating a list of manageable size that will provide learners with the most important words in pharmacology has been achieved. It provides coverage of 27% of the Pharmacology Corpus, which is substantially better than the 9.5% coverage that the AWL gives of the same corpus, or the 12% that Wang et al.'s substantially larger Medical Academic Word List gives of medical corpora. Most interestingly, the EPWL performs almost as well on texts taken from the wider field of general medicine as it does on pharmacology corpora.

Importantly, we have shown that a word should not be included or excluded from a list of this kind simply because it does or does not appear in the GSL or AWL. Unlike the MAWL, the EPWL includes words that are found in the GSL. This is significant, because many of these words are cryptotechnical, and are often used with quite different meanings in pharmacology; even those that are not still have an important role to play as scientific or academic discourse-structuring

words. The inclusion of GSL cryptotechnical words is a major strength of the Essential Pharmacology Word List, and gives it an advantage over lists like the MAWL.

To further improve the list, the corpus could be expanded still more, and an effort made to achieve an even better balance of pharmacology sub-fields. It would then be possible to set a range criterion for a particular word to appear in a minimum number of areas rather than individual articles, and would probably mean that we could set a higher figure of, say, 50%. This might ensure that the words are even more representative of the general field than at present, and that they are well distributed throughout the corpus.

In the creation of the EPWL, words considered unlikely to cause the learners problems were removed. Admittedly, however, the list of removed words was not empirically derived; rather, judgements were made based on my knowledge of pharmacology and the expected level of learners in a Japanese university. For future research it would, therefore, be desirable to test students in order to investigate empirically the extent to which such words are in fact "unproblematic".

On the whole, the Essential Pharmacology Word List lives up to its name. However, a number of important words — words that we would expect to find in a dictionary or glossary of fundamental terms in pharmacology — are missing, primarily because their range did not meet the cut-off criterion. Often, words which fall into this category refer to the basic systems and transmitters in the body, and include *muscarinic, serotonin, dopamine, autonomic, noradrenaline, parasympathetic,* and *sympathetic*. It is possible that a larger, better-balanced corpus might ensure their inclusion in the word list. It may be, though that these words do not appear very often in journal articles precisely because they are so fundamental; knowledge of them is taken for granted, and they do not need to be explicitly mentioned. In any case, there are not very many words like these, and it should be relatively easy to add them to the list if necessary.

CONCLUSION

This paper has described the methodology involved in building a corpus of pharmacology journal articles, as well as the compilation of a manageable, highly efficient word list for pharmacology. Comparison with a similar list designed for medical students, the MAWL, has shown that there is a great deal of overlap between the lexis of pharmacology and that of medicine in general, and that a pharmacology list can provide good coverage of medical texts. This suggests that it might be possible to create a "one size fits all" list for all branches of medical science, with minor adjustments to final, more specialized, lists.

The efficiency of the Essential Pharmacology Word List with both pharmacology and medical corpora indicates that the methodology used in its compilation can be successfully applied in the construction of a corpus for university medical students. An important consideration was found to be the need for "expert" knowledge and intuition in both the selection of texts for the corpus and in refining the word lists. The challenge, therefore, will be how best to select the most relevant texts, and how to ensure that the lists truly reflect the English language needs of future medical practitioners. Davies (2013), investigating the use of English by a medical

practitioner in Japan, shows how interview-based research might be one way to address this. His exploratory study suggests that reading skills for case studies and clinical reports may be particularly important; these findings confirm that the work we have done so far, based on a corpus comprising research articles, has been on the right track. Interview research and surveys could also help to establish the key medical studies to be included in a corpus.

Future research will also require a consideration of important lexical units which have hitherto been afforded scant attention: multiword items. With the exception of a few very obvious units functioning as single items (e.g., *et al*, *in vivo*, and *in vitro*), these are absent from our lists. Fraser (2010), investigating a 5-million word corpus created with WebBootCaT, has shown that a large corpus can throw up many more combinations of words (terms and text-structuring sequences) which function in this way, and that lists which offer collocational information may be of more "productive" use for learners of specialized English than traditional single item word lists (i.e., of use in academic writing as well as reading classes). We can learn, too, from studies such as Gledhill (2000) and Marco (2000), which have moved away from lexical and grammatical analysis in their investigations into collocational patterning in medical research articles.

The obvious application of medical word lists is the compilation of glossaries (see, e.g., Davies et al., 2013), which can be improved by the provision of contextual information. Our ultimate concern is with syllabus and materials design, and word lists can help to ensure that materials writers create texts which include the most important medical words. Information on the collocational and colligational environments of these words can be used to devise exercises which check understanding of how the words are most commonly used. A carefully compiled list will also be extremely useful in determining the vocabulary learning goals of a specialized syllabus for medical students.

REFERENCES

Baker, M. (1988). Sub-technical vocabulary and the ESP teacher: an analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language*, *4*(2), 91–105.

Baroni, M. and Bernadini, S. (2004). Bootstrapping corpora and terms from the web. *Proceedings of LREC 2004*, 1313–1316. Lisbon: ELDA.

Chung, T. M. and Nation, I. S. P. (2003). Technical vocabulary in specialized texts. *Reading in a Foreign Language*, *15*(2), 102–116.

Chung, T. M. and Nation, I. S. P. (2004). Identifying technical vocabulary. *System*, *32*, 251–263.

Cowan, J. R. (1974). Lexical and syntactic research for the design of reading material. *TESOL Quarterly 8*(4), 389–399.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238.

Davies, W. (2013). Linking a medical practitioner's use of English for Medical Purposes in Japan to skills development at university. *Hiroshima Studies in Language and Language Education, 16*, 89–101.

Davies, W., Fraser, S., Lauer, J. and Howell, P. (2013). English for Medical Purposes: Teaching an intensive English course to third-year medical students. *Hiroshima Studies in Language*

and Language Education, 16, 49–64.

Fraser, S. (2007). Providing ESP learners with the vocabulary they need: Corpora and the creation of specialized word lists. *Hiroshima Studies in Language and Language Education, 10*, 127–143.

Fraser, S. (2009). Breaking down the divisions between general, academic, and technical vocabulary: The establishment of a single, discipline-based word list for ESP learners. *Hiroshima Studies in Language and Language Education, 12*, 151–167.

Fraser, S. (2010). Technical vocabulary and collocational behavior in a specialized corpus. *Language, Learning and Context: Proceedings of the 42nd Annual Meeting of the British Association for Applied Linguistics*, 43–48.

Fraser, S. (2012). Factors affecting the learning of technical vocabulary. *Hiroshima Studies in Language and Language Education, 15*, 123–142.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes, 19*(2), 115–135.

Marco, M. J. L. (2000). Collocational frameworks in medical research papers: A genre-based study. *English for Specific Purposes, 19*(1), 63–86.

Neal, M. J. (2003). *Medical Pharmacology at a Glance*. Oxford: Blackwell Science.

Salager, F. (1983). The lexis of fundamental medical English: classification framework and rhetorical function – a statistical approach. *Reading in a Foreign Language, 1*, 54–66.

Salager, F. (1985). Specialist medical English lexis: classificatory framework and rhetorical function – a statistical approach. *EMP Newsletter, 2*(2), 5–17.

Wang, J., Liang, S-l, and Ge, G-c. (2008). Establishment of a medical academic word list. *English for Specific Purposes, 27*(4), 442–458.

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

APPENDIX 1

Words Considered Unproblematic and Excluded from the Word List

A/AN	CALL	END	HOWEVER
ABSENCE	CAN	ENTIRELY	HUMAN
ACCORDING	CAUSE	ENTRY	IF
ACCOUNT	CENTRAL	EQUAL	IMMEDIATELY
ACHIEVE	CHANGE	ESPECIALLY	IMPORTANT
ACROSS	CHOOSE	ESSENTIAL	IMPROVE
ADDITION	CLASS	EVEN	IN
ADDRESS	CLOSE	EVERY	INCLUDE
ADULT	COLLECT	EXCEPT	INCREASE
ADVANCE	COMMITTEE	EXAMPLE	INDEED
AFTER	COMMON	EXIST	INFORMATION
AGAIN	COMPANY	EXPECT	INFLUENCE
AGAINST	COMPARE	EXPLANATION	INTERESTING
AGE	COMPLETE	EXPLORE	INTO
AGREEMENT	CONCERN	EXTENT	INTRODUCTION
AIM	CONNECT	FEATURE	IT
ALL	CONTENT	FEMALE	JUST
ALMOST	CONSIDER	FEW	KEEP
ALONE	CONTAIN	FIELD	KEY
ALONG	CONTINUOUS	FIND	KILOGRAM
ALREADY	CORRECT	FIRST	KNOW
ALSO	COULD	FIT	LACK
ALTHOUGH	COURSE	FIVE	LARGE
AMONG	CROSS	FOLLOW	LAST
AMOUNT	DAMAGE	FOOD	LATER
AND	DARK	FOR	LEAD
ANIMAL	DATE	FREE	LEFT
ANOTHER	DAY	FREQUENCY	LESS
ANY	DEAD	FRESH	LEVEL
APPEAR	DECREASE	FROM	LIFE
APPROVE	DEGREE	FULL	LIGHT
AROUND	DESCRIBE	FURTHER	LIKE
AS	DETAIL	FUTURE	LIKELY
AT	DEVELOPMENT	GAIN	LIMITED
ATTEMPT	DIFFERENT	GENERAL	LINE
AVERAGE	DIFFICULT	GIVE	LITTLE
AVOID	DIRECT	GOOD	LONG
BACKGROUND	DISCUSSION	GREAT	LOSS
BALANCE	DIVIDE	GROUP	LOW
BE	DO	HALF	MAIN
BECAUSE	DOUBLE	HAND	MAKE
BECOME	DOWN	HAVE	MALE
BEFORE	DRIVE	HEALTHY	MAN
BEGINNING	DRY	HEAT	MANAGEMENT
BELOW	DUE	HELP	MANNER
BETWEEN	DURING	HERE	MANY
BOTH	EACH	HIGH	MAY
BOTTOM	EIGHT	HOLD	MEASURE
BUT	EARLY	HOSPITAL	MEAN
BY	EITHER	HOUR	MIGHT
CALCULATE	EMPLOY	HOW	MILD

APPENDIX 1 (cont.)

MIXTURE	POINT	SCIENTIFIC	THREE
MODERATE	POSITION	SECOND	UNTIL
MODEST	POSSIBLE	SEE	UPON
MORE	POWER	SEPARATE	USUALLY
MOREOVER	PREFERENCE	SEEM	VIEW
MOST	PREVENT	SERVE	WASH
MOUSE	PREVIOUSLY	SET	WATER
MUCH	PROBLEM	SEVEN	WEEK
MUST	PRODUCE	SEVERAL	WEIGHT
MULTIPLE	PROGRAM	SHORT	WAY
NAME	PROMISING	SHOULD	WEAK
NATIONAL	PROPERTY	SHOW	WHAT
NATURE	PROPOSE	SIDE	THROUGH
NEAR	PROTECTION	SIMILAR	THROUGHOUT
NECESSARY	PROVIDE	SIMPLE	THUS
NEED	PROVE	SINCE	TIME
NEITHER	PURPOSE	SINGLE	TO
NEW	QUESTION	SITUATION	TOP
NO	RABBIT	SIX	TOTAL
NONE	RAISE	SIZE	TOWARDS
NOR	RAPIDLY	SKIN	TURN
NOT	RAT	SLIGHTLY	TWO
NOTE	RATHER	SLOW	TYPE
NOW	REACH	SMALL	TYPICAL
NUMBER	READ	SO	UNDER
OF	REASON	SOME	UNDERSTAND
OFTEN	RECEIVE	STAGE	UP
OLD	RECENT	START	UNIVERSITY
ON	RECOGNIZE	STEP	WHITE
ONCE	RECOMMEND	STILL	WELL
ONLY	RECORD	STORE	WHEN
ONTO	REFLECT	STRONG	WHERE
OPEN	REFER	STUDY	WHEREAS
OPPOSITE	REGARD	SUBSTANTIAL	WHETHER
OR	REGULAR	SUCCESS	WHICH
ORIGINAL	RELATE	SUCH	WHILE
OTHER	REMAIN	SUGGEST	WHO
OUR	REPEAT	SUPPLY	WHOLE
OUT	REPORT	SUPPORT	WIDELY
OVER	REPRESENTATIVE	SURFACE	WILL
PAIN	RESPECT	TAKE	WITH
PAIR	RESPECTIVELY	TEMPERATURE	WITHIN
PART	RESPONSIBLE	THAN	WITHOUT
PARTICULARLY	REST	THAT	WOMAN
PAST	RESULT	THE	WORK
PATTERN	RETURN	THEIR	WOULD
POOR	REVIEW	THEN	WRITE
PEOPLE	RIGHT	THERE	YEAR
PER	ROOM	THEREFORE	YET
PERHAPS	SAFETY	THINK	YOUNG
PLACE	SAME	THIS	
PLAY	SCALE	THOUGH	

APPENDIX 2

Frequency (and Range) of Headwords in the Essential Pharmacology Word List

ABBREVIATION	22 (21)	AVAILABLE	144 (61)	COMPLICATIONS	40 (21)
ABOLISHED	27 (17)	BACTERIAL	43 (15)	COMPONENT	80 (39)
ABSORPTION	117 (30)	BAR	106 (22)	COMPOSITION	55 (30)
ACCESS	45 (22)	BASED	563 (73)	COMPOUND	237 (44)
ACCUMULATION	52 (29)	BEHAVIOUR	170 (26)	COMPRISING	30 (20)
ACCURACY	46 (19)	BENEFIT	78 (28)	COMPUTER	29 (19)
ACETYLCHOLINE	145 (17)	BINDING	359 (54)	CONCENTRATION	1039 (79)
ACHIEVE	124 (48)	BIOCHEMICAL	31 (20)	CONCEPT	28 (17)
ACID	353 (63)	BIOLOGICAL	105 (32)	CONDITION	320 (68)
ACTION	350 (75)	BLOCK	305 (55)	CONDUCTED	100 (41)
ACTIVITY	1727 (91)	BLOOD	448 (61)	CONFIRMED	111 (54)
ACUTE	101 (38)	BODY	219 (49)	CONSEQUENTLY	60 (37)
ADJUSTMENT	77 (26)	BOVINE	21 (15)	CONSISTENT	205 (63)
ADMINISTRATION	472 (65)	BRAIN	303 (30)	CONSTANT	61 (35)
ADRENERGIC	47 (17)	BRIEFLY	28 (20)	CONSTITUENT	74 (18)
ADVERSE	130 (27)	BUFFER	130 (34)	CONSTRUCTION	49 (21)
AFFECT	160 (63)	CACL2	36 (23)	CONSUMPTION	53 (19)
AFFINITY	119 (22)	CALCIUM	650 (28)	CONTACT	27 (15)
AGENT	162 (44)	CANCER	182 (18)	CONTRACTION	254 (16)
AGONIST	353 (40)	CANDIDATE	26 (16)	CONTRAST	116 (54)
AIR	113 (21)	CAPACITY	41 (24)	CONTRIBUTE	174 (60)
ALCOHOLIC	161 (15)	CARDIAC	382 (35)	CONTROL	830 (85)
ALIQUOT	25 (19)	CARDIOVASCULAR	198 (23)	CONVERTED	52 (30)
ALTER	179 (57)	CARE	118 (36)	CORONARY	90 (19)
ALTERNATIVE	44 (27)	CARRIED	77 (32)	CORRELATION	123 (39)
AMINO	76 (22)	CASE	87 (44)	CORRESPONDING	65 (37)
ANAESTHETIC	103 (19)	CATALYTIC	56 (15)	COUPLE	90 (37)
ANALOGUE	44 (22)	CATEGORY	28 (15)	COURSE	53 (29)
ANALYSIS	624 (85)	CELL	1438 (78)	CRITERIA	36 (18)
ANOVA	81 (30)	CENTRIFUGE	27 (15)	CRUCIAL	28 (15)
ANTAGONIST	327 (47)	CEREBRAL	80 (17)	CULTURE	153 (30)
ANTIBODY	116 (23)	CHAIN	37 (16)	CUMULATIVE	50 (19)
AORTIC	104 (15)	CHALLENGE	95 (22)	CURRENT	371 (57)
APPARENT	75 (31)	CHAMBER	30 (15)	CURVE	182 (37)
APPLICATION	244 (50)	CHANNEL	417 (32)	CYCLE	101 (41)
APPROACH	94 (34)	CHARACTERISTIC	160 (65)	CYTOKINE	147 (19)
APPROPRIATE	49 (26)	CHEMICAL	156 (45)	DATA	697 (93)
APPROXIMATELY	151 (51)	CHRONIC	147 (41)	DECLINE	38 (18)
AREA	116 (44)	CIRCULATING	72 (23)	DEFICIENCY	43 (17)
ARTERY	155 (30)	CL	191 (25)	DEFINE	100 (47)
ASPECT	28 (20)	CLEARANCE	217 (73)	DEGRADATION	55 (15)
ASPIRIN	17 (16)	CLINICAL	397 (56)	DELIVERY	126 (30)
ASSAY	224 (45)	CLONED	57 (25)	DEMONSTRATED	290 (75)
ASSESSED	220 (57)	CO	48 (20)	DENSITY	90 (38)
ASSIGNED	21 (15)	CODE	49 (21)	DEPENDENT	561 (84)
ASSOCIATED	465 (77)	COEFFICIENT	49 (19)	DEPOLARIZATION	97 (16)
ASSUMED	57 (22)	COMBINATION	67 (57)	DEPRESSION	110 (23)
ATP	149 (19)	COMPARTMENT	78 (16)	DERIVATIVE	126 (43)
ATTENUATED	47 (29)	COMPENSATORY	25 (18)	DESIGN	90 (46)
ATTRIBUTE	21 (15)	COMPETITIVE	44 (15)	DESPITE	61 (43)
AUTHOR	48 (25)	COMPLEX	147 (44)	DETECTED	182 (55)

APPENDIX 2 (cont.)

DEVIATION	28 (15)	FACILITATION	70 (25)	INDICATE	449 (81)
DIABETES	239 (15)	FACTOR	360 (70)	INDIVIDUAL	223 (52)
DIAGNOSIS	69 (17)	FAILURE	125 (41)	INDUCED	1024 (80)
DIAMETER	40 (20)	FAST	83 (25)	INFECTION	73 (18)
DIET	66 (18)	FATTY	103 (24)	INFLAMMATORY	380 (40)
DIFFUSION	43 (17)	FED	58 (21)	INFUSION	141 (23)
DILUTED	59 (20)	FETAL	45 (17)	INHIBITION	1118 (79)
DISEASE	269 (51)	FIGURE	1024 (81)	INITIAL	191 (68)
DISPLAYED	55 (22)	FILTER	47 (21)	INJECTION	358 (44)
DISRUPTION	32 (18)	FINAL	116 (51)	INJURY	58 (19)
DISSOLVED	49 (28)	FIXED	28 (17)	INSERTED	37 (16)
DISTINCT	54 (22)	FLOW	166 (35)	INSTITUTE	23 (17)
DISTRIBUTION	158 (37)	FLUID	36 (18)	INSTRUMENT	42 (24)
DNA	108 (22)	FLUORESCENC	E60 (17)	INTACT	52 (16)
DOMAIN	52 (17)	FLUX	48 (21)	INTAKE	154 (15)
DOSE	948 (73)	FOCUSED	49 (27)	INTENSITY	36 (25)
DRUG	988 (77)	FOLD	128 (35)	INTERACTION	204 (64)
DURATION	69 (29)	FORCE	34 (15)	INTERNAL	53 (19)
EFFECT	1827 (98)	FORMATION	290 (63)	INTERPRETATION	25 (17)
EFFICACY	117 (35)	FORMULATION	53 (24)	INTERVAL	45 (30)
ELEVATED	135 (32)	FRACTION	84 (27)	INTESTINE	120 (16)
ELICITED	74 (21)	FUNCTION	628 (73)	INTRAPERITONEAL	33 (15)
ELIMINATION	50 (20)	FURTHERMORE	86 (45)	INTRAVENOUS	52 (21)
ENDOGENOUS	90 (30)	FUSION	36 (16)	INVESTIGATION	226 (70)
ENDOTHELIUM	151 (24)	GASTROINTESTINAL	46 (16)	INVITRO	151 (40)
ENHANCED	228 (51)	GENE	290 (48)	INVIVO	182 (46)
ENVIRONMENTAL	49 (21)	GENERATION	169 (45)	INVOLVED	299 (75)
ENZYME	238 (49)	GLUCOSE	305 (26)	ION	85 (22)
EPITHELIAL	42 (16)	GRADE	26 (16)	ISOLATED	161 (42)
EQUATION	36 (16)	GRAPH	59 (25)	ISSUE	35 (22)
EQUILIBRIUM	43 (21)	GROWTH	114 (32)	KCL	43 (24)
EQUIVALENT	37 (23)	GUIDELINE	37 (24)	KIDNEY	126 (19)
ERROR	56 (20)	HEART	188 (31)	KINASE	113 (25)
ESTABLISHED	95 (44)	HENCE	31 (21)	KINETIC	194 (34)
ESTIMATED	131 (29)	HEPATIC	100 (18)	KIT	36 (17)
ETAL	2087 (71)	HISTORY	40 (22)	LABEL	68 (20)
ETHANOL	89 (16)	HOMEOSTASIS	42 (22)	LABORATORY	95 (43)
ETHICS	27 (23)	HYPERTENSION	131 (22)	LAYER	47 (17)
EVALUATED	142 (57)	HYPOTHESIS	94 (48)	LIFE	77 (34)
EVENT	145 (36)	IC50	65 (16)	LIGAND	181 (18)
EVIDENCE	225 (69)	IDENTICAL	17 (16)	LINKED	39 (22)
EVOKED	116 (17)	IDENTIFIED	150 (55)	LIPID	74 (17)
EXAMINED	158 (53)	ILLUSTRATED	36 (19)	LIQUID	36 (23)
EXCITATION	74 (23)	IMAGE	66 (16)	LIVER	247 (25)
EXCLUDED	64 (25)	IMMUNE	42 (17)	LOADED	58 (22)
EXERT	50 (29)	IMPACT	45 (22)	LOCAL	114 (42)
EXHIBITED	111 (38)	IMPAIRED	107 (23)	LOCATED	60 (29)
EXOGENOUS	27 (15)	IMPLICATED	57 (36)	LOG	41 (17)
EXPERIMENT	595 (70)	INCIDENCE	68 (16)	LUNG	180 (18)
EXPOSURE	247 (54)	INCORPORATED	30 (16)	MACROPHAGE	140 (15)
EXPRESSION	738 (76)	INCUBATION	143 (36)	MAGNITUDE	44 (25)
EXTRACT	98 (29)	INDEX	50 (22)	MAINTAINED	130 (52)

APPENDIX 2 (cont.)

MAJOR	176 (68)	OXYGEN	202 (38)	PROMINENT	24 (15)
MAMMALIAN	37 (24)	PARALLEL	26 (21)	PROMOTER	85 (19)
MARKED	178 (48)	PARAMETER	189 (39)	PROPERTY	100 (37)
MASS	70 (24)	PARTICIPANT	57 (16)	PROPORTION	53 (18)
MATERIAL	67 (40)	PATHOLOGY	49 (22)	PROTEIN	651 (72)
MAXIMUM	221 (49)	PATHWAY	230 (56)	PROTOCOL	100 (40)
MEAN	497 (81)	PATIENT	987 (59)	PUBLISHED	74 (43)
MECHANISM	288 (75)	PEAK	152 (34)	PURCHASED	43 (24)
MEDIATED	328 (60)	PEPTIDE	192 (25)	PUTATIVE	25 (15)
MEDICATION	123 (20)	PERCENTAGE	125 (41)	QUANTIFIED	47 (27)
MEDIUM	105 (27)	PERFORMED	295 (73)	RANDOMIZE	117 (39)
MEMBRANE	210 (45)	PERFUSION	201 (26)	RANGE	221 (66)
METABOLISM	413 (45)	PERIOD	244 (55)	RATE	320 (65)
METHOD	261 (79)	PERIPHERAL	113 (32)	RATIO	120 (35)
METHYL	21 (15)	PERSISTENT	28 (15)	REACTION	264 (57)
MICROSCOPE	42 (20)	PH	173 (42)	RECEPTOR	1303 (66)
MINIMAL	76 (36)	PHARMACOLOGICAL	194 (51)	RECOVERY	34 (22)
MINOR	19 (16)	PHASE	148 (43)	REDUCED	533 (89)
MOBILE	42 (25)	PHENOTYPE	25 (15)	REGIMEN	71 (20)
MODE	27 (20)	PHOSPHORYLATION	58 (17)	REGION	151 (38)
MODEL	603 (72)	PHYSICAL	38 (23)	REGRESSION	38 (22)
MODIFICATION	146 (55)	PHYSIOLOGICAL	129 (47)	REGULATION	320 (68)
MODULATION	154 (42)	PLACEBO	211 (15)	RELAXATION	294 (15)
MOLAR	125 (15)	PLASMA	297 (45)	RELEASE	346 (54)
MOLECULAR	170 (56)	PLATE	63 (17)	RELEVANT	83 (36)
MONITORED	72 (34)	PLOT	34 (17)	REMOVED	97 (42)
MOUNTED	29 (19)	PLUS	80 (20)	RENAL	112 (22)
MRNA	95 (20)	POPULATION	177 (31)	REQUIRED	171 (60)
MUSCLE	243 (37)	POSITIVE	83 (36)	RESEARCH	80 (39)
MUTATION	80 (18)	POST	93 (30)	RESIDUAL	69 (27)
NACL	43 (27)	POTENT	177 (54)	RESISTANCE	146 (30)
NEGATIVE	62 (33)	POTENTIAL	428 (74)	RESPONSE	1037 (79)
NERVE	175 (35)	PRACTICE	41 (16)	RESTRICTED	27 (20)
NEURON	348 (36)	PREDICTED	116 (36)	RETENTION	41 (19)
NEVERTHELESS	32 (21)	PREDOMINANTLY	37 (23)	REVEALED	83 (42)
NITRIC	46 (18)	PRELIMINARY	24 (18)	REVERSED	105 (42)
NORMAL	308 (75)	PREPARATION	215 (52)	RICH	35 (17)
NOVEL	71 (31)	PRESENT	706 (92)	RISK	331 (33)
NUCLEUS	110 (28)	PRESSURE	219 (28)	RODENT	43 (21)
OBSERVED	482 (85)	PREVALENCE	45 (16)	ROLE	405 (80)
OBTAINED	271 (72)	PREVIOUSLY	330 (79)	ROUTINE	22 (15)
OCCURRED	179 (66)	PRIMARY	149 (53)	SALT	68 (24)
ONSET	51 (15)	PRINCIPLE	27 (17)	SAMPLE	302 (56)
OPTIMAL	31 (15)	PRIOR	99 (46)	SATURATION	44 (16)
ORAL	117 (29)	PROBABILITY	117 (43)	SCALE	46 (21)
ORDER	157 (64)	PROBE	48 (24)	SCORE	80 (18)
ORGAN	70 (21)	PROCEDURE	74 (36)	SECRETION	112 (25)
OUTCOME	77 (25)	PROCESS	176 (53)	SECTION	91 (22)
OUTPUT	40 (15)	PROFILE	98 (38)	SELECTIVE	289 (60)
OVERALL	58 (32)	PROGRESSION	86 (23)	SENSITIVE	261 (61)
OXIDATION	247 (29)	PROLIFERATION	54 (15)	SEQUENCE	91 (32)
OXIDE	160 (29)	PROLONGED	55 (22)	SERIES	71 (17)

APPENDIX 2 (cont.)

SERUM	156 (35)	SUBSTITUTED	26 (16)	TRANSDUCER	29 (22)
SHIFT	51 (22)	SUBSTRATE	120 (26)	TRANSFERED	38 (27)
SIGMA	82 (35)	SUBTYPE	53 (15)	TRANSIENT	72 (26)
SIGNAL	164 (41)	SUBUNIT	114 (15)	TRANSMISSION	89 (17)
SIGNIFICANT	915 (98)	SUFFICIENT	43 (28)	TRANSPORT	180 (28)
SIMILAR	299 (83)	SUMMARY	80 (41)	TREATMENT	1125 (89)
SIMULTANEOUS	37 (23)	SUPPLEMENT	33 (15)	TREND	33 (16)
SITE	220 (46)	SUPPRESSION	86 (20)	TRIAL	255 (35)
SMOOTH	137 (24)	SURGERY	62 (22)	TRIGGER	29 (15)
SODIUM	121 (37)	SURVIVAL	106 (15)	TUBE	46 (21)
SOFTWARE	54 (32)	SUSCEPTIBLE	28 (19)	TUMOUR	176 (20)
SOLUBLE	67 (17)	SUSPENDED	27 (17)	UNDERGOING	52 (31)
SOLUTION	317 (46)	SUSTAINED	34 (23)	UNDERLYING	45 (28)
SOURCE	52 (25)	SYMPTOM	36 (15)	UPTAKE	36 (15)
SPECIES	92 (34)	SYNAPTIC	199 (17)	URINARY	189 (18)
SPECIFIC	319 (78)	SYNDROME	33 (15)	UTILITY	48 (31)
SPONTANEOUS	74 (18)	SYNTHESIZED	130 (42)	VALUE	422 (72)
STABLE	143 (41)	SYSTEM	390 (69)	VARIABILITY	448 (84)
STAINING	53 (20)	TABLE	383 (63)	VARIANCE	104 (32)
STANDARD	172 (62)	TARGET	201 (51)	VASCULAR	152 (31)
STATE	86 (41)	TECHNIQUE	52 (26)	VEHICLE	133 (24)
STATISTICAL	236 (70)	TENSION	55 (15)	VEIN	49 (18)
STATUS	46 (18)	TERM	121 (46)	VERSUS	158 (48)
STEROID	65 (17)	TERMINAL	69 (27)	VESSEL	89 (15)
STIMULATION	472 (62)	TEST	509 (86)	VIA	174 (56)
STRATEGY	39 (27)	THERAPY	517 (64)	VISUAL	32 (19)
STRESS	166 (24)	THEREBY	39 (21)	VOLUME	148 (47)
STRUCTURE	155 (43)	TISSUE	432 (63)	WALL	31 (18)
SUBCUTANEOUS	45 (16)	TOLERANCE	128 (19)	WHEREAS	110 (46)
SUBJECT	360 (54)	TONE	49 (15)	WILD-TYPE	45 (17)
SUBSEQUENT	126 (62)	TOXICITY	185 (38)	WITHDRAWAL	61 (19)
SUBSTANCE	54 (21)	TRANSCRIPTION	111 (19)	YIELDED	24 (15)

要　約

医学専攻大学生のためのコーパスの構築と語彙学習リストの編纂

サイモン・フレイザー
広島大学外国語教育研究センター

外国語教育研究センターでは，医学専攻学生を対象とした語彙ベースのシラバス設計に供することを目的に，大規模な医学英語コーパスの構築を計画している。本論ではこれを念頭に置き，医学の一分野である薬理学の雑誌論文コーパスの構築，および，扱いやすい規模で効率性の高い語彙リスト「薬理学主要語彙リスト」（EPWL）の編纂について，これらの方法論を説明する。医学専攻学生のために作成された同様のリストである「医学語彙リスト」（MAWL）と比較すると，EPWL の効率性が際立っており，また薬理学の語彙と医学一般の語彙がかなりの部分で共通していることがわかる。EPWL による薬理学および医学コーパスのカバー率はいずれも高く，このことは同リストの編纂に用いられた方法論が，医学専攻大学生のためのコーパス構築および最重要語リストの作成において有効に適用されうることを示している。