The Asian Journal of Applied Linguistics
Vol. 4 No. 1, 2017, pp. 3-15

AJAL

A corpus-based lexical analysis of Chinese medicine research articles

Cailing Lu
Victoria University of Wellington

Philip Durrant
University of Exeter

This study investigates the usefulness of two academic word lists - Coxhead's (2000) Academic Word List (AWL) and Gardner & Davies' (2014) Academic Vocabulary List (AVL) - for students of English for Chinese Medical Purposes. The two academic word lists were evaluated in terms of the coverage they achieved in a corpus of Chinese medical research articles (CMRAs) written in English. The AWL was found to cover 10.64% of tokens in the corpus, while the AVL was found to cover 21.17% overall. In both cases, the majority of the coverage was achieved by a relatively small subset of the lexical items on the lists. Analysis of the most frequently used words that are not included in the General Service List, Academic Word List and Academic Vocabulary List in the CMRAs shows that a small number of such words achieve a high level of coverage, suggesting that they should be given a great deal of attention by learners in this discipline. This suggests that a discipline-specific listing would be of great benefit to learners in this discipline. A list of the most prominent 100 off-list lexical items is provided.

Key words: word list; English for Chinese medical purposes; lexical analysis; academic vocabulary

Introduction

Due to the internationalisation of traditional Chinese medicine (TCM), a growing body of research in TCM is being carried out around the world and reported in English, especially in the form of English-language research articles. This trend has been further enhanced by the Nobel Prize awarded to Professor Youyou Tu, a traditional Chinese medicine scientist.

As a result of this tendency, and to further facilitate the development of TCM, academic courses in English for Chinese Medical Purposes (ECMP) have been established. These are interdisciplinary courses which bridge the gap between content knowledge and English-language knowledge, and have become compulsory for undergraduates and postgraduates in Chinese medicine. These TCM learners are a homogenous group of students with similar educational backgrounds who function in an English as a foreign language context. Though they have learned general English for at least 6 years, they still have a great deal of difficulty in reading English-medium academic texts. The fundamental aim of ECMP courses is to enable learners to read and understand international research articles in the field of TCM.

A key challenge facing students of ECMP is that of mastering sufficient vocabulary for their studies. Lack of vocabulary knowledge has been reported as one of the main

ISSN 2308-6262
http://caes.hku.hk/ajal

difficulties facing students of English for Academic Purposes (EAP) in a wide range of contexts (Berman & Cheng, 2001; S. Evans & Green, 2007; Stephen Evans & Morrison, 2011; Wu & Hammond, 2011). In response, researchers have produced word lists they claim meet those needs (Cowan, 1974; Coxhead, 2000; Farrell, 1990; Gardner & Davies, 2014; Ghadessy, 1979; Wang, Liang, & Ge, 2008; Xue & Nation, 1984; Yang, 2015). While the full breadth of vocabulary encountered during university-level programmes is immense, academic word lists are based on the insight that the vast majority of text is made up of a relatively small number of frequently-used words. As Nation and Waring (1997) describe, the most frequent 2,000 words in English account for around 80% of the tokens found in written texts. This suggests that targeted learning of high-frequency words will pay disproportionate dividends in helping learners perform in English. Nation and Waring (1997) point out that the added proportion of texts covered by words beyond the first 2,000 drops dramatically. They point to figures from Francis and Kucera (1982) which show that the third most frequent thousand words adds only an additional 4% coverage, while the fourth adds under 3%. Nation and Waring (1997) therefore suggest that, beyond the most frequent 2,000 words, students' attention is better directed to lexical items related to more specific needs.

The best-known attempt to put this recommendation into practice is Coxhead's (2000) *Academic Word List* (AWL). This is a list of 570 word families which are not amongst the 2,000 most frequent words of English set out in West's (1953) *General Service List* (GSL) but which were found to occur frequently across disciplines in a 3.5-million-word corpus of academic writing. Coxhead (2000) showed that these 570 families accounted for around 10% of tokens in academic writing, an impressive improvement on the 4% achieved by the third most frequent thousand words of general interest.

A more recent potentially influential word list is Gardner & Davies's (2014) *Academic Vocabulary List* (AVL) which is different in three key ways. First, whereas the AWL is a list of word families, the AVL is based on lemmas (i.e., only inflectionally-related forms are combined). Thus, while the AWL treats all inflectionally- and derivationally-related forms of a word (e.g. *constitute, constitutes, constituency, unconstitutional*) as a single item, the AVL only combines inflectionally-related forms in this way (*constitute* and *constitutes* are considered a single item, but *constituency* and *unconstitutional* are counted separately). Individual items on the AVL therefore achieve less coverage than items on the AWL, but they provide much more specific guidance and are less likely to conflate items with different meanings. Second, unlike the AWL, the AVL does not build onto an existing list of high-frequency words. Gardner and Davies (2014) point out that the age of the GSL has rendered it inaccurate for today's English, such that many items in the AWL could today be counted as general high-frequency vocabulary. They also note that, by excluding high-frequency words, the AWL excludes many words which, though frequent in general English, are particularly relevant to academic texts (e.g. *company, interest, market*). As Gardner and Davies (2014) point out, Coxhead's approach does not allow learners to distinguish such items from high-frequency words which are less relevant to their needs (e.g. *bed, pretty, fun*). Rather than excluding a set of word from the start, Gardner and Davies (2014) take a statistical approach to distinguishing academic from more general words, including in their list only lemmas which are at least 50% more frequent in their academic corpus than in a corpus of non-academic English. Third, the AVL is based on a much larger and more balanced corpus of academic writing than the AWL. As a number of authors have noted (Durrant, 2014; Hyland & Tse, 2007), Coxhead's corpus was strongly biased towards two particular disciplinary areas (Commerce and Law),

which led to a corresponding bias in the words included. There is therefore good reason to believe that the AVL will provide a more balanced list of generic academic vocabulary, and thus be relevant to a broader range of disciplines, than the AWL.

On the other hand, the AVL also has disadvantages. First, it is a relatively long listing, and therefore is less pedagogically friendly than the AWL. Evaluating the use of the AVL by successful student writers across 32 disciplines based on the British Academic Written English (BAWE) corpus, Durrant (2016) found that around half the items in the AVL make negligible contribution to the coverage, and thus may not be useful to the majority of learners (at least in terms of written production). Second, it contains a considerable amount of high-frequency vocabulary, of which students might have already gained control. The relative utility of the two lists therefore remains an open question.

A number of researchers have questioned the relevance of both lists to the needs of learners in particular disciplines. Most discussion has centred around the utility of the AWL. Hyland and Tse (2007) found that the GSL and AWL covered only 78.3% (rather than the claimed 90%) of tokens in their corpus of science writing. Similarly, Martinez, Beck, and Panza (2009) found coverage of only 76.6% in their corpus of agriculture research articles. Unpacking the 570 word families in the AWL into their 3,107 constituent lemmas, Martinez et al. (2009) also found that 37.5% of these did not appear in their corpus at all, suggesting that the AWL includes a large number of items which are not of great utility for students in this area. Similarly, Chen and Ge (2007) demonstrated that half of AWL families were infrequent in a corpus of medical research articles. While most criticism has focused on the AWL, problems of this sort may exist for any list of general academic vocabulary. Durrant (2014) found a high level of variation in the vocabulary used by student writers across different disciplines and argued that any cross-disciplinary list of vocabulary is likely to fall short of meeting the needs of any particular group of learners.

These findings suggest that further studies of the use of academic words in specific disciplines are needed, both to evaluate the utility of the AWL and AVL in particular areas and take a close look at the high frequency words that are not covered by any vocabulary lists used in the present study. TCM offers an interesting test case in this context, as a significant and growing field which was not included in the corpora on which these lists were based.

The present study therefore aims to evaluate the utility of the AWL and AVL for learners of ECMP and to provide a list of words which are likely to be important for these learners but are not included in either list. It examines the coverage and frequency of words from each list in a corpus of 309 Chinese medicine research articles (CMRAs) and provides a listing and evaluation of most frequent words in these articles which are off the GSL, AWL, and AVL used in this study.

Methodology

This study used a corpus-based approach in which the coverage and frequency of AWL and AVL items in a custom-built corpus (the TCM Corpus) were calculated for the corpus as a whole. To understand the nature and importance of other vocabulary in this corpus, a list of high-frequency words which were not found on the GSL, AWL or AVL was also created and its coverage evaluated.

The lists are evaluated here purely in terms of frequency: i.e. of their coverage of the CMRA corpus. As Durrant (2016) has pointed out, decisions on which words to teach need ultimately to be based on a broader range of factors, including the words'

relationship to other items being taught (e.g. the word *Wednesday* should be taught together with *Friday* and *Saturday*, despite being significantly less frequency); their role in particular texts students need to work with (less frequent words are often key to the meaning of a text); and learners' own learning histories and first languages. However, the general principle that learners should focus on the words they are most likely to meet (Mackey, 1965) is one that has stood the test of time. It is also important to note that, the focus on frequency and coverage in this study evaluates the lists in their own terms, as these were the principles on which they were created.

The TCM corpus

The corpus used in this study comprised 309 CMRAs, downloaded from four scholarly journals: Chinese Medicine[1], Journal of Traditional Chinese Medicine[2], BMC Complementary and Alternative Medicine[3], and Complementary Therapies in Medicine[4]. All are international quarterlies reporting clinical and theoretical TCM research. All articles included in the corpus follow an IMRD (Introduction–Method–Result–Discussion) structure with an abstract, and were published between the years 2008 and 2015. Journal articles were chosen as the target genre for the TCM Corpus because it is one of the most prominent genres that ECMP students are likely to read in their current study and future professional settings. The division of TCM into specialisms is not as clear-cut as in western medicine but it can be roughly divided into: Chinese herbal medicine including component analysis, pharmacology and toxicology; alternative medicine, including acupuncture, moxibustion, auricular acupuncture and scraping therapy; and basic research into theories, diagnoses, and other general aspects of TCM. In order to ensure balance, articles were sampled in equal numbers from each of these three categories.

Charts, diagrams, images, Chinese characters, and other components which were deemed irrelevant to the articles' vocabulary (e.g., statistical and chemical symbols and website addresses) were manually removed from the articles. Though charts and images were excluded, the key sentences and vocabulary related to them were kept. Bibliographies, index numbers, acknowledgements, and appendices were also excluded. The final TCM Corpus included 1,045,969 tokens (see Table 1).

Table 1. Corpus sample by subject areas

Subject areas	Number of texts	Tokens
Chinese herbal medicine	103	347,701
Alternative medicine	103	348,759
Other general aspects of TCM	103	349,509
Total	309	1,045,969

Procedures

Vocabprofile (Cobb, 2002; Heatley, Nation, & Coxhead, 2002), Frequency, and Range programmes (Cobb, 2002; Heatley et al., 2002; Nation, 2001) were employed to analyse the AWL's coverage, frequency, and distribution in the corpus. These programmes are

preloaded with GSL words and AWL items, and can be used to calculate the proportion of the corpus each list covers and determine the frequency and range of items from each list in any given corpus (Hyland & Tse, 2007). Since the Range programme has problems recognising sentences or words without punctuation, such as headings and subheadings of each article, the corpus was modified by adding full stops after each heading/subheading.

The AVL is not included in these programmes, so customised means were created to determine the frequency and coverage of this list. Specifically, the TCM Corpus was tagged for part-of-speech using the online CLAWS tagging service provided by Lancaster University's University Centre for Computer Corpus Research on Language (UCREL). Since lemmatization of the AVL was itself based on CLAWS tagging (Gardner & Davies, 2014), this enabled direct comparability of the TCM with the AVL. The open-source programming language *R* (R Core Team, 2014) was then used to create a lemma-based wordlist (the TCM wordlist) for the tagged corpus and to make comparisons with the AVL. Because the CLAWS tagger has a small but important error rate estimated at around 3-4% (UCREL, n.d.), the part-of-speech-tagged TCM wordlists were manually corrected by the first researcher to reclassify incorrectly tagged words.

Results and discussion

Overall coverage of the GSL and AWL in the TCM corpus

As shown in Table 2, the cumulative text coverage of the GSL (65.72%) and the AWL (10.64%) for the CMRAs as a whole is 76.36%, rather lower than that the 90% reported by Coxhead and Nation (2001). It is also lower than the 86% found for multidisciplinary corpora by Coxhead (2000) and the 85% by Hyland and Tse (2007). This low coverage is mainly due to the low coverage of the GSL (65.72%) in the TCM Corpus, 14.28% lower than that the 80% proposed by previous researchers (e.g. Coxhead & Nation, 2001). However, these figures are consistent with those reported for other discipline-specific corpora, as shown in Table 3. The discrepancy between these and the multi-disciplinary corpora is most likely due to the greater prominence of discipline-specific terminology in more specific corpora. In corpora which include a wide range of texts, the features that are distinctive of particular text types or topic areas tend to have relatively low overall frequencies, such that technical vocabulary becomes less prominent (Dakin, Tiffen, & Widdowson, 1968). Since it is the discipline-specific corpora which are likely to correspond more closely to actual learners' experience of the language, this suggests that the GSL is somewhat less important, and discipline-specific vocabulary somewhat more important, than analyses based on multi-disciplinary corpora suggest.

Table 2. Coverage of GSL and AWL in the TCM corpus

	Tokens	Coverage
GSL	687,374	65.72%
AWL	111,335	10.64%
GSL+AWL	798,709	76.36%
Off List	247,260	23.64%
Total	1,045,969	100%

Table 3. Comparison of AWL and GSL coverage in the TCM corpus and in three other discipline-specific corpora

	TCM Corpus	Hyland & Tse's science corpus (2007)	Martinez's AgroCorpus (2009)	Valipouri & Nassaji's CRAC (2013)
GSL	65.72%	69%	67.53%	65.76%
AWL	10.64%	9.3%	9.06%	9.96%
GSL+AWL	76.36%	78.3%	76.59%	75.42%

The AWL accounts for 10.64% of tokens in TCM Corpus, a figure which is slightly higher than that of Coxhead's (2000) multidisciplinary corpus (10%) and that of Chen and Ge's (2007) medicine corpus (10.07%). It is also slightly higher than the corresponding figures in other discipline-specific corpora such as Martinez et al.'s agriculture corpus (see Table 3), indicating that they constitute a high percentage of the running words in the field of medicine in general. This suggests that the AWL is of considerable value to vocabulary learning in ECMP.

Frequency of the AWL word families in the TCM corpus

Table 4 illustrates the frequency and distribution of the AWL word families in the TCM Corpus according to their frequencies of occurrence, and coverage achieved by each set of word families. Of the total 570 word families in the AWL, 563 are found in the TCM Corpus. Coxhead (2000, p. 221) classifies word families with a minimum of 100 occurrences in her 3.5 million words Academic Corpus (approximately 30 times per million words) as frequent. Using the same normalised frequency, word families that occur at least 30 times in the 1,045,969-word TCM Corpus are regarded here as frequent. As Table 4 shows, 405 (71.05%) AWL word families are frequently used in the TCM Corpus. The most frequent academic word is *significant*, followed by *analyse*, *participate*, *data*, *method*, *function*, *outcome*, *research*, *assess* and *indicate* (see Table 5). Apart from these 405 word families, the other 165 word families can be regarded as infrequent, with 7 (1.2%) word families never appearing, indicating that not all the AWL items are equally useful to ECMP learners.

Table 4. Frequency of the AWL words in the TCM corpus

Occurrences	Number of AWL words	Cumulative number of AWL words	Coverage (%)	Cumulative coverage (%)
>500	59	59	5.28	5.28
499-100	207	266	4.39	9.67
99-30	139	405	0.78	10.45
29-1	158	563	0.19	10.64
0	7	570		

Table 4 also shows that the most frequently used 59 AWL word families in the TCM Corpus account for 5.28% of the coverage, which is about half of the overall AWL coverage. The 158 word families that can be regarded as infrequent only account for 0.19% of the total tokens, indicating that these families are of relatively little utility for ECMP learners. It seems there is a rather restricted number of AWL word families which are truly useful in the CMRAs. A similar conclusion was also reached by Chen and Ge (2007) and Martinez et al. (2009) in their discipline-specific corpora, and by Hyland and Tse (2007) in their multidisciplinary corpus. Rather than focusing on the AWL as a whole, therefore, learners of ECMP may be best advised to focus on only the top 100 word families from the AWL (as listed in Table 5) which have a cumulative coverage of 6.74%.

Table 5 Top 100 AWL word families in the TCM corpus (arranged according to frequency)

1. significant	26. regulate	51. demonstrate	76. allocate
2. analyse	27. induce	52. major	77. consent
3. participate	28. investigate	53. detect	78. require
4. data	29. similar	54. item	79. protocol
5. method	30. conduct	55. compound	80. injure
6. function	31. consist	56. affect	81. volume
7. outcome	32. normal	57. random	82. final
8. research	33. specific	58. depress	83. distribute
9. assess	34. select	59. positive	84. ratio
10. indicate	35. bias	60. benefit	85. overall
11. intervene	36. process	61. convene	86. available
12. evaluate	37. evident	62. individual	87. stress
13. statistic	38. involve	63. promote	88. theory
14. respond	39. inhibit	64. component	89. recover
15. vary	40. primary	65. proceed	90. abstract
16. factor	41. design	66. index	91. culture
17. tradition	42. mechanism	67. locate	92. region
18. criteria	43. area	68. intense	93. survey
19. extract	44. range	69. duration	94. negate
20. previous	45. obtain	70. administrate	95. confirm
21. practitioner	46. differentiate	71. define	96. chemical
22. identify	47. potential	72. role	97. publish
23. formula	48. exclude	73. occur	98. target
24. medical	49. period	74. valid	99. plus
25. conclude	50. physical	75. react	100. structure

Coverage and frequency of AVL in the TCM corpus and comparison with that of AWL

Having examined the coverage and frequency of the AWL in the TCM Corpus, this section examines the coverage and frequency of the AVL, and makes an indirect comparison with that of the AWL to explore which list can better serve ECMP students' lexical needs.

The AVL was found to account for 21.17% of the total tokens in the TCM Corpus (Table 6). In addition to the token coverage, the lexical coverage (i.e. the coverage of nouns, verbs, adjectives and adverbs) is also provided to enable comparison with the findings of Durrant (2016). This increased the coverage to 35.62%. These figures suggest that the AVL can be regarded as an important learning goal for ECMP students. As can be seen in Table 5, both the overall and lexical coverage of the AVL in the TCM

Corpus is slightly higher than that of the BAWE corpus which were found by Durrant (2016). However, it is interesting that among the 13 text types used in Durrant's study, research report has the highest lexical token coverage, and this figure (35.62%) is very close to the lexical token coverage reported in the current study (35.55%). Therefore, the AVL is especially valuable to students who aim to conduct academic research.

Table 6. Coverage of the AVL in the TCM corpus and in the BAWE corpus

Corpus	Overall token coverage	Lexical token coverage
The TCM Corpus	21.17%	35.62%
The BAWE Corpus	16.82%	33.82%

Table 7 shows the number of lemmas meeting the frequency thresholds used above in the analysis of the AWL, along with the cumulative coverage of the corpus achieved by words at each frequency level. As can be seen, of the 3,014 lemmas in the AVL, 2,282 lemmas (75.69%) occur in the TCM Corpus. Nonetheless, these 2,282 lemmas are not of equal use for CMRAs readers. The most frequent 87 lemmas from the AVL covered 10.19% of the corpus, a figure that is more than half of the overall coverage (21.17%). *Group, study, effect, compare, report, include, result, analysis, data,* and *control* are the most frequently used items (Table 8). The least frequently used 1,458 items, with occurrences under 30, only cover 1.17% of the total tokens. The fact that 732 (24.31%) lemmas never appear suggests that the AVL is a less efficient list than the AWL.

Table 7. Frequency of the AVL lemmas in the TCM corpus

Occurrences	Number of AVL lemmas	Cumulative number of AVL lemmas	Coverage (%)	Cumulative coverage (%)
>500	87	87	10.19	10.19
499-100	330	417	7.37	17.56
99-30	407	824	2.44	20
29-1	1,458	2,282	1.17	21.17
0	732	3,014		

A good overall sense of the coverage achieved by different numbers of items in the two lists can be gained from Figure 1, which shows cumulative token coverage achieved by AVL lemmas and by AWL word families. The AVL achieves rather higher coverage of the corpus than the AWL with lower numbers of items, as well as its overall coverage being higher. In both cases, however, the additional coverage achieved by new items flattens out quickly, indicating again that only a limited number of words on the two lists are frequent in CMRAs. This is consistent with Durrant's (2016) finding in a

multi-discipline corpus where a small number of items accounted for a majority of the coverage. It seems that CMRAs use academic words frequently, but they do not use them diversely.

Table 8. TOP 100 AVL lemmas in the TCM corpus (arranged according to frequency)

1. group	26. change	51. conduct	76. describe
2. study	27. indicate	52. information	77. relate
3. effect	28. rate	53. western	78. improvement
4. result	29. perform	54. conclusion	79. finding
5. analysis	30. activity	55. subject	80. scale
6. data	31. system	56. identify	81. demonstrate
7. control	32. factor	57. statistical	82. both
8. significant	33. function	58. respectively	83. period
9. compare	34. value	59. total	84. process
10. difference	35. base	60. associate	85. thus
11. method	36. increase	61. therefore	86. present
12. level	37. use	62. measure	87. primary
13. report	38. traditional	63. observe	88. combination
14. significantly	39. evaluate	64. bias	89. depression
15. table	40. practitioner	65. previous	90. comparison
16. include	41. improve	66. pathway	91. differentiation
17. outcome	42. effective	67. determine	92. index
18. however	43. review	68. induce	93. measure
19. model	44. condition	69. important	94. apply
20. pattern	45. provide	70. mechanism	95. conventional
21. low	46. assess	71. analyze	96. specific
22. figure	47. suggest	72. common	97. positive
23. reduce	48. type	73. database	98. similar
24. research	49. response	74. characteristic	99. component
25. quality	50. practice	75. obtain	100. university

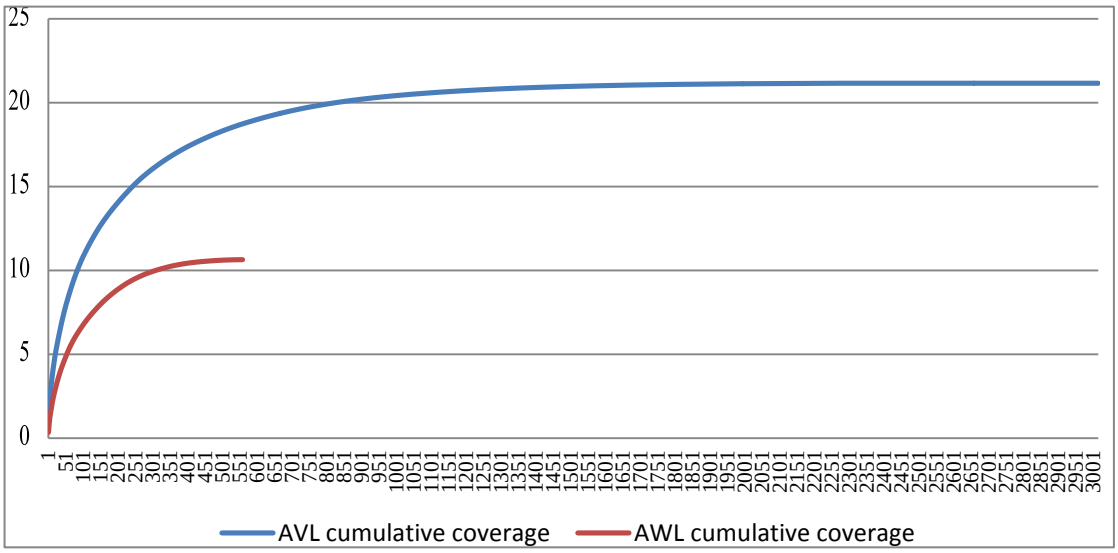

Figure 1. Comparison of cumulative token coverage by AVL lemmas and by AWL word families

The fundamental differences between the AWL and AVL, described above, prevent a meaningful direct comparison of their coverage. However, the AVL has two important

advantages. Firstly, it is not dependent on the GSL and so can be used for learners who have not gained control of the GSL, especially the second thousand word families. Secondly, the AVL consists of lemmas rather than word families which is less challenging than the AWL that requires knowledge of derivational words and inflectional words. Knowledge of derivational words normally develops later than for inflectional words (Gardner & Davies, 2014) and requires a high level of linguistic proficiency (Nation, 2001). Thus, the AVL might be a more realistic goal for learners without control of the GSL or who are at a low level of language proficiency; although for learners are at a high linguistic proficiency, the AWL is a relatively smaller set of vocabulary to master.

The most frequently used off-list words in the TCM corpus

The above analyses show that generic academic vocabulary has an important role in TCM. However, it has also shown that a substantial amount of the corpus is NOT covered by general academic words. To get a more comprehensive understanding of the vocabulary use in this discipline, this section examines the usefulness of such vocabulary by looking at the top 100 lexical word (verbs, nouns, adjectives and adverbs) that do not appear in the GSL, AWL and AVL (see Table 9 for details).

Table 9. TOP 100 off-list words in the TCM corpus (arranged according to frequency)

1. patient	26. acupressure	51. meta-analysis	76. administered
2. acupuncture	27. tissue	52. decoction	77. gastrointestinal
3. cell	28. baseline	53. software	78. renal
4. Chinese	29. therapeutic	54. glucose	79. surgery
5. clinical	30. qi	55. sensation	80. respiratory
6. symptom	31. placebo	56. granule	81. assay
7. score	32. hypertension	57. randomized	82. pregnancy
8. acupoint	33. questionnaire	58. inflammation	83. massage
9. therapy	34. serum	59. abdominal	84. correlation
10. liver	35. acid	60. diabetes	85. psoriasis
11. syndrome	36. diagnostic	61. oral	86. cognitive
12. drug	37. muscle	62. physician	87. headache
13. stimulation	38. pulse	63. inflammatory	88. ischemic
14. cancer	39. disorder	64. tumor	89. unclear
15. protein	40. injection	65. cerebral	90. constipation
16. deficiency	41. kidney	66. dose	91. pulmonary
17. moxibustion	42. activation	67. signalling	92. fatigue
18. efficacy	43. receptor	68. neuron	93. prescription
19. herbal	44. medication	69. stasis	94. metabolic
20. herb	45. session	70. versus	95. zheng
21. adverse	46. chemotherapy	71. nerve	96. hepatic
22. gene	47. apoptosis	72. moxa	97. randomization
23. diagnosis	48. acupuncturist	73. spleen	98. pharmacological
24. chronic	49. fibrosis	74. auricular	99. asthma
25. sham	50. acute	75. coating	100. metabolism

Note: the off-list words have been grouped into lemmas

These top 100 (lemmatised) words cover an impressive 5.99% of the total tokens, suggesting that they play a significant role in this discipline. This coverage compares well with the top 100 words from the AWL word families (6.74%) and the top 100 lemmas from the AVL (10.82%). This suggests that the off-list vocabulary is of crucial importance in achieving academic literacy in this discipline. As can been seen in Table

9, a vast majority of the off-list vocabulary, such as *acupuncture*, *syndrome, herb,* are discipline-specific (see the Appendix for an example of its impact). The compilation of a discipline-specific academic word list is thus suggested to facilitate the effectiveness of vocabulary instruction in this field.

Conclusions

The AWL and AVL are of great importance to ECMP teachers and learners alike as in other disciplines in that these lists achieve a high level of coverage of the TCM Corpus, even though CMRAs were not included in the corpora which were used in compiling the AWL and AVL. The high coverage of the AVL and AWL, and their wide spread use in the TCM texts, suggest that these academic word lists can provide valuable information for both course planning and individual lexical learning. The most frequently used lexical items as listed in Table 5 and Table 8 might be an attainable learning goal and, if possible, should be incorporated into the syllabus. Another important conclusion of the coverage analysis is that discipline-specific vocabulary has an essential role to play. A short list of the top 100 off-list items from the TCM Corpus demonstrated a high level of coverage. Moreover, words from this list appear to play crucial roles in establishing the core meanings of texts. This suggests the need for a combination of academic words and discipline-specific words to provide ECMP students with the maximum support in the generally limited course time available. This is consistent with the findings of Chen and Ge (2007) for mainstream medicine and implies the need for an ECMP discipline-specific word list to be used alongside the more general lists.

While the listing provided in Table 9 gives a sense of what this might look like, the present study has two key limitations which mean that this cannot be taken as authoritative. First, some words are not recognisable by the software used in this study due to the specificity of TCM terminology. Therefore, they were counted inconsistently by the software. For example, the names of herbs, such as Xixin, were sometimes calculated as multiple words, but other times were counted as one word only. The same is true for other names borrowed directly from Chinese. Second, the current research is based on a relatively small corpus and limited to research articles only. For the compilation of a discipline-specific word list, it would be desirable to compile a corpus including other published course and textbooks and the spoken form of language in this discipline. The creation of such a corpus, and of software capable of processing TCM terminology, is suggested as an important task for future research in this area.

Notes

1. Available online at http://www.cmjournal.org
2. Available online at http://www.sciencedirect.com/science/journal/02546272
3. Available online at http://bmccomplementalternmed.biomedcentral.com/articles
4. Available online at http://www.sciencedirect.com/science/journal/09652299

About the authors

Cailing Lu is a PhD candidate in applied linguistics at Victoria University of Wellington in New Zealand, and her research interests include vocabulary in English for Specific and Academic Purposes. The research reported in this paper was conducted when she was doing her Master's at the University of Exeter.

Philip Durrant is a senior lecturer in Language Education at the University of Exeter in the U.K. His main research interests are in corpus linguistics and academic writing.

References

Berman, R., & Cheng, L. (2001). English academic language skills: Perceived difficulties by undergraduate and graduate students, and their academic achievement. *Canadian Journal of Applied Linguistics, 4*(1-2), 25-40.

Chen, Q., & Ge, G. C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles. *English for Specific Purposes, 26*(4), 502-514.

Cobb, T. (2002). Compleat Lexical Tutor v.8.3 [computer program]. Retrieved 16 Feb, 2016, from http://www.lextutor.ca/cgi-bin/range/texts/index.pl

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

Coxhead, A. (2000). A new academic wordlist. *TESOL Quarterly, 34*, 213-238.

Coxhead, A., & Nation, P. (2001). The specialised vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes*. Cambridge: Cambridge University Press.

Dakin, J., Tiffen, B., & Widdowson, H. G. (1968). *Language in education: The problem in commonwealth Africa and the Indo-Pakistan sub-continent*. Oxford: Oxford University Press.

Durrant, P. (2014). Discipline- and level-specificity in university students' written vocabulary. *Applied Linguistics, 35*(3), 328-356.

Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? . *English for Specific Purposes, 43*(1), 49-61.

Evans, S., & Green, C. (2007). Why EAP is necessary: A survey of Hong Kong tertiary students. *Journal of English for Academic Purposes, 6*(1), 3-17. doi: 10.1016/j.jeap.2006.11.005

Evans, S., & Morrison, B. (2011). The first term at university: Implications for EAP. *ELT Journal, 65*(4), 387-397. doi: 10.1093/elt/ccq072

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

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305-327.

Ghadessy, P. (1979). Frequency counts, words lists, and materials preparation: A new approach. *English Teaching Forum, 17*, 24-27.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). RANGE and FREQUENCY programs. Retrieved 16 Feb, 2016, from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx.

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

Mackey, W. F. (1965). *Language teaching analysis*. London: Longman.

Martinez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes, 28*(3), 183-198.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.

R Core Team. (2014). R: A language and environment for statistical computing. from http://www.R-project.org/

University Centre for Computer Corpus Research on Language. (n.d.). CLAWS part-of-speech tagger for English. Retrieved 22nd March 2017, from http://ucrel.lancs.ac.uk/claws

Valipouri, L., & Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. *Journal of English for Academic Purposes, 12*(4), 248-263.

Wang, J., Liang, S.-l., & 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 and Co.

Wu, W., & Hammond, M. (2011). Challenges of university adjustment in the UK: a study of east Asian master's degree students. *Journal of Further and Higher Education, 35*(3), 423-438.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3*(2), 215–219.

Yang, M.-N. (2015). A nursing academic word list. *ESP English for Specific Purposes, 37*, 27-38.

Appendix: Example of use of TCM vocabulary

The following paragraph is an extract randomly selected from a Chinese medicine research article in the TCM Corpus:

> *A* RANDOMIZED *controlled trial on* ACUPUNTURE *treatment for* HYPERTENSION enrolled *192* patients *and the frequency of* Zangfu *patterns was recorded. However, no* **data** *related to observed* manifestations *were given and no association was* **investigated** *between* CLINICAL *findings* (e.g. *blood pressure*) *and patterns*. Flachskampf et al. RANDOMIZED *the* **allocation** *of 160* outpatients *with* uncomplicated HYPERTENSION *in a single-blind fashion to a 6-week course of* ACUPUNTURE **intervention**; *however, they did not report descriptive* **statistics** *on patterns or* manifestations *or association* **analysis**. Chu et al. *reported 59 cases of* HYPERTENSION *classified according to whether or not* abundant phlegm-*dampness was presented for* **analysis** *of* proteome. *Again, no* **analysis** *was* **conducted** *to explore the frequency* **distribution** *of patterns or its* manifestations. Gu et al. **investigated** *the frequency* **distributions** *of patterns in 477* untreated *subjects with* HYPTERTENSION *and did not find* **statistical** **significance** *in the frequency* **distributions** *of patterns within blood pressure levels, age or body mass* **index** (BMI). *This* heterogeneity *of* **analysis** *regarding patterns in subjects with* HYPERTENSION *led to the reports of opposite results of* ACUPUNCTURE *treatment for lowering mean 24-hour* ambulatory *blood pressures*.

Key: *italics* = GSL; **bold** = AWL; underlined = AVL; UPPERCASE = top 100 off-list TCM words.

Coverage:
GSL words = 71.59%
AWL words = 9.29%
AVL words = 22.95%
top 100 off-list TCM words = 6.01%