To Wordlist or Not to Wordlist?
The dilemma and challenges for language learning and testing

Yu-Hua Chen
Pearson

Abstract

This article argues that wordlists which can prioritise the lexical items that learners are most likely to encounter often in real life have an indispensible role in language teaching and testing. Yet compilation of a pedagogical wordlist should be collaborative work between language teachers, learning material publishers and exam boards instead of establishing a wordlist purely on the basis of or word frequencies from authentic texts. This article will not only demonstrate how a gap between a set of language courses and a suite of language tests is identified by comparing the words used in these materials but also exemplify how corpus-driven wordlists derived from these materials can be used to facilitate or facilitate the approaches to teaching and testing lexis.

1. Introduction

A wordlist, in the broadest sense, often entails single words as well as multi-word units and thus may be called a vocabulary list, a lexical syllabus or a lexical framework. It has been an important tool for language learning and testing. The Academic Word List (Coxhead, 2000), which was compiled from academic texts with corpus approaches, for example, has been widely used for EAP teaching and learning. However, in the context of general language learning, how such a list if needed should be compiled, and what the components of such should be is still open to much debate. This is particularly disputed in the field of language testing, where the fear is that coaching schools would simply teach the students to remember the whole wordlist by heart in order to pass the test. However, a good lexical syllabus can also include contextual information about language use, such as frequent word combinations, parts of speech, core meanings of ambiguous lexical items, rather than a simple list of individual words.

In this paper, both theoretical and practical perspectives of wordlists for language teaching and testing purposes will be discussed. Particularly the potential applications of corpus approaches in pedagogical wordlist research will be introduced with some examples of preliminary investigation using empirical data from language courses and language tests. The challenges and a proposed process of pedagogical wordlist compilation will be discussed too.

2. Typology of Wordlists

First of all, in order to discuss the issues surrounding wordlists, we need to understand the fundamental question, i.e. what is a wordlist? Various types of wordlists can be distinguished

according to their methodology of compilation and purposes, each of which will be briefly introduced and exemplified in this section.

Methodology

In terms of methodology, distinction can be determined by who compiles the wordlists: expert-compiled wordlists or corpus-driven wordlists. Traditionally human experts such as experienced language teachers, editors, or lexicographers compile the wordlists for different purposes. A dictionary, for example, can be seen as a type of most comprehensive wordlist which covers a broad variety of useful information for language learners including pronunciation, part of speech, definition, sample sentences, phrases and idioms, and even common learner errors (e.g. Macmillan English Dictionary for Advanced Learners 2002, see Gilquin, Granger and Paquot 2007). For any textbooks used for language courses, a wordlist which highlights a selection of words used in the text (usually the perceptibly important words) is often given, and similarly additional contextual information such as part of speech or sample sentences may be provided. Wordlists of this kind are selected and edited by human experts to various extents. A corpus-driven wordlist, on the contrary, is the result of pure computation, which lists each word present in the text input without any human intervention. This type of computer generated wordlists can be easily sorted by alphabetical order, frequency, or any other statistics. A good example of corpus-driven wordlist would be the book *Word Frequencies in Written and Spoken English* (Leech, Rayson, and Wilson, 2001). In the Forewords of this book, language education is listed as the first and probably the most important discipline that calls for word frequency research. This book provides exhaustive statistical information of the words in actual use derived from the British National Corpus (BNC). However, a list of frequency-driven words extracted from a corpus, informative and comprehensive as it is, is not readily usable for language teaching and learning. Corpus-driven wordlists generally require human expert judgement or selection for any meaningful interpretation or practical application. The examples of an expert-compiled wordlist with a dictionary entry and a corpus-driven wordlist can be seen below:

Figure 1: A dictionary entry from Free Dictionary (Retrieved from http://www.thefreedictionary.com/)

Figure 2: Wordlists produced with the corpus tool *WordSmith*

Purposes

In terms of purposes, wordlists can serve for the following various functions: research, reference, assessment, and teaching/learning. Corpus-driven wordlists are commonly used in corpus research, for instance, to determine whether the texts are written by the same author or whether the words with significantly high frequency are suggestive of any patterns that might not emerge if the text is read manually (Archer, 2009). In terms of reference, dictionaries, encyclopaedia, or the aforementioned BNC word frequencies can be seen as the most comprehensive wordlists that provide a wealth of information for reference, which may be used for any language users but particularly useful for writers, content developers, or editors of course materials or language tests.

For assessment purposes, some exam boards publish wordlists (particularly for the lower proficiency levels) from which the test papers are based on - i.e. words outside the published wordlists would not be tested. This type of assessment wordlists is often popular among teachers and students who are preparing for the tests because it provides a set of explicit lexical items for test preparation. However, assessment wordlists often arouses the criticism that such wordlists tend to be over-simplified and that students very likely would just try to remember all the words from the list by heart without fully understanding the usage and context of vocabulary. Moreover, to what extent these assessment wordlists are consistent with the content of language courses is still very much a question, which will be further discussed in the next section.

Two examples of assessment wordlists from General English Proficiency Test (GEPT, developed and administrated in Taiwan) and Cambridge Main Suite are provided below:

abandon	verb/noun		中級	
abandoned	adj.			中高級
abbey	noun			中高級
abbreviate	verb			中高級
abbreviation	noun			中高級
abdomen	noun			中高級
abide	verb			中高級
ability	noun	初級		
able	adj.	初級		
abnormal	adj.		中級	
aboard	prep./adv.		中級	
abolish	verb			中高級
aboriginal	adj./noun			中高級
aborigine	noun			中高級
abortion	noun			中高級
abound	verb			中高級

Figure 3: GEPT Word List (retrieved from http://www.lttc.ntu.edu.tw/academics/wordlist.htm) – the three columns on the right stand for proficiency levels: Elementary, Intermediate, and High-intermediate

cafe/café **(n)**
cake **(n)**
call **(v)**
 • I'll call (phone) again later this afternoon.
 • He's called John.
camel **(n)**
camera **(n)**
camp **(n & v)**
campsite **(n)**
can **(n & v)**
canal **(n)**
car **(n)**
card **(n)**

Figure 4: Cambridge Main Suite – Key English Test (KET) Vocabulary List (retrieved from http://www.cambridgeesol.org/assets/pdf/exams/ket/ket-vocab-list.pdf)

For the purpose of language teaching and learning, the Dolch Word List (Dolch, 1948) is probably one of the pioneering pedagogical wordlists developed with the concept of corpus-driven analysis. The Dolch Words (see Figure 5) were compiled from children's reading books from 1930s and 1940s with the aim to provide the most frequent vocabulary for children to quickly achieve literacy. The main wordlist is composed of 220 'service words' graded into five levels with 95 additional common nouns, and in total they are estimated to make up 50-75% of the words in any children's books. Another equally popular wordlist is the General Service List of English Words (GSL, see West 1953), which was designed for L2 English learners and listed the most frequent words for general service purposes. Despite being compiled around six decades ago, both the Dolch Words and GSL are still commonly used for developing educational materials nowadays. Whether these wordlists have been ageing with time still requires further research, but their long-standing popularity might also have the implication that no pedagogical wordlist with updated content compiled with modern methodology is available yet to replace their status.

Dolch Word List – by Frequency, by Grade

Pre-primer	Primer	First	Second	Third
the	he	of	would	if
to	was	his	very	long
and	that	had	your	about
a	she	him	its	got
I	on	her	around	six
you	they	some	don't	never
it	but	as	right	seven
in	at	then	green	eight
said	with	could	their	today
for	all	when	call	myself
up	there	were	sleep	much
look	out	them	five	keep
is	be	ask	wash	try
go	have	an	or	start
we	am	over	before	ten
little	do	just	been	bring
down	did	from	off	drink
can	what	any	cold	only

Figure 5. Cambridge Main Suite – Key English Test (KET) Vocabulary List (retrieved from www.dolchword.net/)

Developed in more recent years, the Academic Word List (Coxhead, 2000) is probably one of the most well known pedagogical wordlists. This corpus-informed wordlist (see Figure 6) consists of most frequently used 570 headwords and word families with 10 sublists covering various forms and parts of speech from the headwords in the academic discourse. This wordlsit was designed to facilitate English language learning for Academic Purposes. One of the downsides of this wordlist, just as many other similar pedagogical wordlists, is that it is still a 'single-word' list as opposed of taking into account of multi-word units such as phrases or idioms. Although phraseology has increasingly received more attention during the past decade, for example the development of the Academic Formula List (Simpson-Vlach & Ellis, 2010), a pedagogical lexical list integrated with single words and multi-word units compiled with the use of corpus-driven approaches are still not commonly seen.

analyse
    analysed
    analyser
    analysers
    analyses
    analysing
    analysis
    analyst
    analysts
    analytic
    analytical
    analytically
    analyze
    analyzed
    analyzes
    analyzing

Figure 6. The first item in Sublist 1 of the Academic Word List (retrieved from http://www.victoria.ac.nz/lals/resources/academicwordlist/awl-headwords.aspx)

One current ongoing project is the English Vocabulary Profile funded by Cambridge University Press and Cambridge ESOL Examinations collaborating with academics, government advisors, and expert in related areas. This project aims to provide graded vocabulary list in relation to the Common European Framework of Reference (CEF or CEFR) by using learners' written production from the Cambridge English Profile Corpus

derived from Cambridge Learner Corpus as well as learner output collected from classrooms and other teaching and learning scenarios. One feature distinctive from other pedagogical wordlists is that the English Vocabulary Profile includes not only single words but also multi-word expressions and hence called the English 'Vocabulary' Profile as opposed to the English Profile 'Wordlist'). However, as one potential caveat, learners' language production appears to be the major source of data collection, including data from Cambridge Learner Corpus, which largely consist of candidates' written scripts from Cambridge ESOL tests. This project thus focuses on 'what learners *do* know rather than what they *should* know' (Capel, 2010). Apparently the distinction between receptive language (reading and listening) and productive language (writing and speaking) is not acknowledged, which might cast some doubt on the generalisability of its CEF grading at least for some of the entries; therefore, to what extent the sole focus of learners' productive language is of much use to practitioners such as material developers or language testers which cater for all four skills is still to be observed. Additionally, we probably cannot expect that the graded lexical items in this project can accommodate various conditions in a wider range of geographic areas as some lexical items may be rather cultural specific. For instance, the reference noun 'dragon' might be a B1 or B2 item for Chinese L1 speakers as 'dragon' is a common symbol generally associated with auspicious powers and good luck in the Chinese culture but is not included in the CEFR levels A1-B2. By the same token, another item 'olive' is labelled as an B1 item but could be B2 or higher in some areas where olives are not grown. The risk of over-generalisation is therefore particularly true in the case of any noun entries which fall into the category of discourse specific references (e.g. culturally or geographically), although such danger is likely to be less severe with the majority of verbs and possibly quite minimal in function words (e.g. preposition or conjunctions).

While one might argue that CEFR was originally designed for European languages and nationals only and thus learners from other geographic areas might not or should not be considered in this framework, it is undeniable that nowadays the influence of CEFR is overwhelming worldwide. As the English Vocabulary Profile is still an ongoing project at the time of writing, it is probably still at too early a stage to evaluate to what extent the potential downsides resulted from design of data collection – focusing on learners' productive language only and possible over-generalisation of discourse sensitive entries – can possibly impact the practice of material development and language teaching/learning/testing. Nevertheless, the English Vocabulary Project sets off from aiming to develop a lexical framework to complement CEFR (a function and Can-do based framework), which can serve as exceptionally rich resources for researchers as well as practitioners. For future research, it would be interesting to compare this CEFR level defined L2 wordlists with its L1 English counterpart. In addition, mapping the graded vocabulary items against the CEFR to see to what degree the vocabulary items reflect the Can-do statements and level descriptors in CEFR might also provide some insights for the empirical underpinnings of rationale.

3. Mapping of Wordlists between Language Courses and Language Tests

As can be seen, the purposes of wordlists and ways of compilation vary from case to case. For educational purposes, expert-compiled wordlists can foreground the lexical items that may be new or most important for students and thus can function as a valuable tool. On the other hand, corpus-driven wordlists can provide reference for material writers and can also be used to validate the content. In this section, an empirical investigation aiming to identify the gap of wordlists between language courses and language tests will be illustrated, and some

preliminary exploration of corpus-driven wordlists in the validation of material development will be addressed in the next section.

The mapping of wordlists in this article originated from a request from a famous chain children's ESL school. The concern from the language school is that the teachers needed to evaluate the suitability of Pearson Test of English Young Learners (PTE Young Learners) for their courses before they could decide whether to adopt Pearson's suite of tests as an achievement test for their students. The PTE Young Learners suite of tests covers four levels ranging from Firstwords through Springboard, Quickmarch to Breakthrough. There is a syllabus including a wordlist and a description of the language expected to be known by test takers for each of the levels. The four levels of published wordlists, hereafter called the PTE YL assessment wordlists, were developed in 2007. Since then the test papers have been written on the basis of the PTE YL assessment wordlists, and the team of test writers have been instructed to use only the words in the assessment wordlist at the required level. As the teachers from this children's language school were keen to know the suitability of the PTE Young Learners for their courses, the first step would be to check the vocabulary coverage of PTE Young Learners in their courses materials. The course materials were also divided into four levels in order to map against the four levels in the Pearson tests. The words used in the four course levels were then compared with the four PTE YL assessment wordlists, and the results are presented in Figure 7.
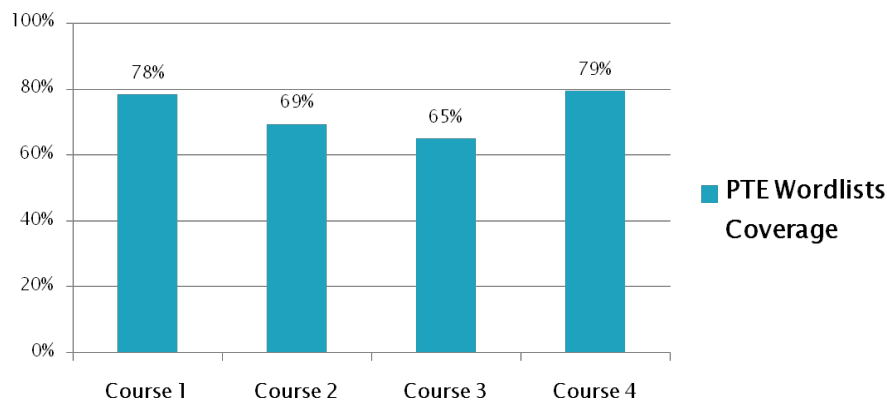

Figure 7. Mapping of Wordlists between ESL Young Learner Courses vs. Pearson Test of English Young
        Learners (published assessment wordlists)

As can be seen, the vocabulary coverage rate ranges from 65% to 79%, which does not look promising. It means that in the worst case scenario, students who finish Course 3 and sit the PTE Young Learners test would only be able to recognise around 65% of the vocabulary in the test paper, and surely this sends a warning signal to the language school. Yet one issue for this mapping exercise is that not all the words from the wordlists are used in the test papers. It is very likely that the core vocabulary in the PTE YL assessment wordlists which would be repeatedly tested in different test sessions is overlapped with that in the course materials. That is to say, the peripheral vocabulary (lexis outside the core vocabulary) in the assessment wordlists might be very different from that in the course materials, and hence the part that contributes to the difference here. In order to test this hypothesis, the next step is to compare the words used in a set of test papers (the most recent ones in 2011 for this investigation) and those in the course materials again. The results of new mapping are presented in red columns along with the original wordlist coverage rates presented in blue columns in Figure 8.
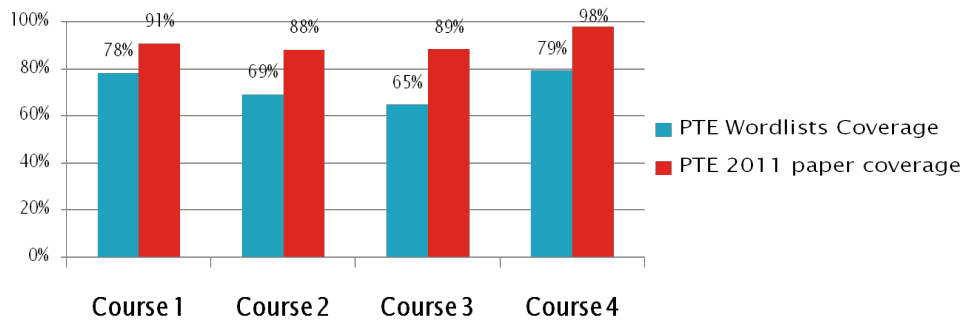
Figure 8. Mapping of Two Wordlists: ESL Young Learner Courses vs. Pearson Test of English Young Learner Wordlist (from the live 2011 test session)

The vocabulary coverage in the test papers, as expected, has arisen significantly by 13%-24%. For Course 4, the coverage rate even reaches as high as 98%. The inference is that if the students finish any course in this language school and sit the Pearson Young Learners test with a corresponding level, they should be able to recognise 91.5% of the words in the test papers on average.

The results of this mapping exercise are inspiring, yet it also leaves some unresolved questions. If the assumption of distinction between core and peripheral vocabulary holds true, how do we define this distinction? And to what extent is this distinction defined differently by course material developers and test developers? In the following section, the focus will move to the side of test development as an attempt to better understand the gap discovered in the assessment wordlists and the actual word usage in test papers.

4. Exploration of Corpus-driven Wordlists in Live Test Papers

As mentioned earlier, corpus-driven wordlists can be used to validate the content of material development including language tests as well as course books. A corpus of test papers from the Pearson Test of English Young Learners was built to explore the application of this approach. Here only the lowest level Firstwords is used for preliminary investigation. This pilot corpus consists of twelve test papers collected from three test sessions a year during 2008-2011. As this is a test designed for younger children, the test itself is rather short and hence a small corpus with 12,413 tokens only (see Table 1). Although each paper is supposed to be of similar length as the general expectation is that the input for test takers should be as close as possible, a significant gap of nearly 200 words' difference is discovered between Paper 10 (923 words) and Paper 11 (1,122 words). It turns out that Paper 8 was written by a new test writer whereas Paper 11 by an experienced writer who had written the test papers at this specific level for more than two years. It is noteworthy that even the basic statistical information regarding the text length of each paper which can be easily computed by any corpus tool already reveals some interesting information like this.

Table 1 Composition of the test paper corpus

Text	Tokens used for word list
Paper 1	1,028
Paper 2	1,031
Paper 3	1,026
Paper 4	1,010
Paper 5	1,075
Paper 6	1,086
Paper 7	1,095
Paper 8	992
Paper 9	1,026
Paper 10	923
Paper 11	1,122
Paper 12	999
Total	12,413

The following step is to compare the words used in the test papers and the published assessment wordlist for this level. The words which appear in the published list but are never present in the three years' papers are identified and grouped according to topics. The results are presented in Table 2. As can be seen, a number of very common nouns, such as body parts ('ear', 'mouth', 'nose', etc.) or food ('carrot', 'egg', 'vegetables', etc.) have never been tested during any of the 12 sessions. As the team of test writers is rather small, it is likely that writer idiosyncrasies lead to these 'forgotten' items, and this tendency should be rectified in future writer training to ensure that all the words in the published assessment wordlist are catered for. It also needs to be mentioned that the reason why colour words (e.g. 'blue' or 'red') have never been tested is due to the constraint of a practical issue that the test papers are printed in greyscale only except for the cover.

Table 2 Words in the published assessment wordlist but not in the Corpus of Test Papers

Body part	Colour	Food	School	Position	Life	Time
arm	blue	carrot	alphabet	above	breakfast	night
body	pink	egg	bell	below	song	
ear	purple	ham	board	close		
eye	red	meat	crayon			
face		potato	eraser			
foot		rice	pen			
mouth		tomato	pencil			
nose		vegetable				

Interestingly, although the test writers are specifically instructed to use only the words from the assessment wordlists and this has always been carefully checked during the review stages, some words in the corpus are still found to have deviated from the published assessment wordlist. These words are again categorised on the basis of topics and presented in Table 3. The results indicate that some lexical items that the writers should not have used are actually very basic and should have been included in the assessment wordlist. For example, the question words 'how' and 'who' with high frequency (41 and 42 times respectively), are not present in the assessment wordlist until the next higher level, yet they have always been used in the test papers forming the questions such as 'Who's in the living room?' 'Who is riding a bike?' 'How old is he?' or 'How much is this?' It also seems a bit odd that a number of very

common adjectives such as 'fine', 'nice', 'slow', 'tall' or the object reference 'tea', are not included in the published assessment wordlists for this lowest level, and these words all appear to be CEFR A1 items. As it is probably true that some words in Table 3 could have been easily replaced by other words in the test papers (e.g. some food reference entries such as 'biscuit' or 'hamburger'), this preliminary analysis also points to the gap in the assessment wordlist by identifying the words which were used by the writers but not in the assessment wordlist. From the perspective of test writers, it is possible that the writers just mistook these lexical items and thought they were in the lists. The information revealed from such a mapping exercise therefore can be used to amend and improve the existing wordlists to compensate for what might have been missed out at the stage of wordlist compilation.

Table 3 Words in the Corpus of Test Papers but not in the published assessment wordlist*

Common adjective	Object	Food	Position	Question word
blonde (2)	cup (1)	biscuit (2)	centre (2)	how (41)
dark (4)	newspaper (1)	coffee (1)	front (4)	who (42)
fine (1)	tea (1)	fruit (2)	near (11)	why (1)
hungry (6)		hamburger (1)	up (2)	
lovely (3)		sandwich (7)		
lucky (1)				
nice (20)				
silly (1)				
slow (1)				
tall (1)				
thirsty (1)				

*Frequency of words used in the test papers is indicated in brackets.

As this is simply a preliminary investigation, the analysis reveals only a limited part of the picture, but it also throws some light on how the corpus-driven approach can be employed to gauge the actual use of expected language in material development including language tests.

5. Discussion & Conclusion

Although the above results are encouraging in terms of the effectiveness of methodology, there are various aspects that this pilot investigation has not tackled. To start with, it would be interesting to compare the frequency bands of the words in the course materials and the words tested in the test papers in order to indentify the common ground and to define the core and the peripheral vocabulary from both sides. This information not only can be used for future design and development of curriculum and syllabus but also can be used to ensure that the content matches the specifications. Statistical information can also be extracted and sorted by writers to investigate whether writer idiosyncrasies play a role in the gap between the syllabus and the actual content in the materials. Last but not least, word frequencies can be used to investigate whether the levels progress as expected under the assumption that the higher proficient the materials target at, the wider range of vocabulary should be covered. It also has to be stressed that the proposed set of analysis can be applied in the development or validation of course materials as well as language tests.

However, there are also challenges for this explorative study. Similar with many other wordlist projects, multi-word expressions were not included in the investigation because of the technical limitation – so far there has been no automated technique which can precisely

identify phrases, idioms, or any other phraseological units. One possible solution might be to retrieve all recurrent word combinations of different lengths (a.k.a. n-grams or lexical bundles, see e.g. Chen & Baker 2010, De Cock 1998) and map this computer generated list of recurrent clusters against a pre-existing list of phraseology.

For most of the curriculum or test developers, the most fundamental question posed is – what is a good wordlists? On the one hand, those who oppose the provision of pedagogical wordlists believe that wordlist would only lead to spoon-fed education because students would not be able to learn words in natural context. On the other hand, we have seen that phrases, idioms or typical usage can be presented in the wordlists with sample sentences. Frequency information can also be used to prioritise the order of acquisition effectively. The key issue here is how we can ensure a coherent lexical framework from one end of language teaching towards the other end of assessment to measure the progress of learning. This lexical syllabus would require collaboration of all stakeholders including language course providers, publishers of teaching materials, and exam boards. The compilation process of such a lexical syllabus is proposed to comprise six phases (see Figure 9). It starts with defining the purpose and identifying the target source materials (phase 1). The source materials are used to establish a corpus to inform the construction of a wordlist (phase 2). Experts from experienced practitioners such as teachers or content developers then judge the quality of wordlist (phase 3), and the wordlist can be improved accordingly (phase 4). What follows is to empirically measure the vocabulary coverage using the amended wordlist in another independent set of target text (which is not used to build the wordlist) (phase 5). If this exercise supports the validity of wordlist, then the wordlist can be published with regular review and update (phase 6).
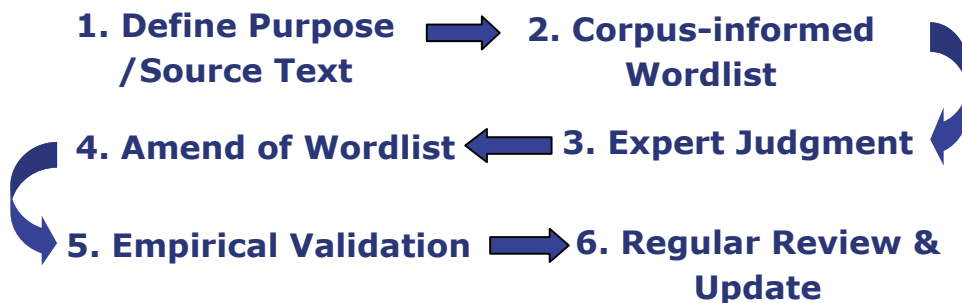

Figure 9. Proposed Process of Pedagogical Wordlist Compilation

Wordlists have been an important tool for the acquisition of vocabulary, and yet how to apply corpus approaches to compile or validate a pedagogically useful and informative lexical syllabus is still an area worth much research. A good lexical syllabus can highlight the words or phrases that learners most likely would encounter in real life and prioritise the order of acquisition to make learning more effective and possibly also more explicit. Additional information such as representative usage or context can be provided in the syllabus too. Such a lexical framework, however, would require collaborative work from all parties involved.

The functions of wordlists are versatile. At the stage of material development, lexical resources from native corpora such as BNC word frequencies can provide reference for syllabus developers as well as content writers to ensure that the content is at the expected level. At the stage of post-hoc validation, frequency-driven wordlists extracted from the materials can be used to gauge the degree of lexical coverage as demonstrated. As we have seen, a significant gap between different wordlists was found in the materials this way. A

further investigation of the extracted wordlists can also reveal whether some words are over-represented or under-represented. This information can not only provide feedback to the material writers but also be incorporated into future training. Wordlists, as discussed, have an indispensible role in language learning and testing. Compiling a good pedagogical wordlist is not easy, yet it is hoped that this article has shed some light on this tremendous task.

References

Archer, Dawn. (ed.) (2009). *What's in a Word-list? Investigating Word Frequency and Keyword Extraction.* Farnham: Ashgate.

Capel, Annette. (2010). A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1, e3. (retrieved from http://www.englishprofile.org/index.php?option=com_content&view=section&layout=blog&id=8&Itemid=99).

Chen, Y. and Baker, P. (2010). Lexical Bundles in L1 and L2 Academic Writing. *Language Learning and Technology*. 14 (2): 30-49.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34 (2): 213-238.

De Cock, Sylvie. (1998). A Recurrent Word Combination Approach to the Study of Formulae in the Speech of Native and Non-Native Speakers of English. *International Journal of Corpus Linguistic*s, 3 (1): 59–80.

Dolch, Edward William. (1948). Problems in Reading. Garrard Press.

Leech, Geoffrey, Paul Rayson, and Andrew Wilson. (2001). *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London: Longman.

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

Simpson-Vlach, Rita and Ellis, Nick C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31(4): 487-512.

West, Michael. (1953). *A General Service List of English Words.* London: Longman.