

A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project

Annette Capel Cambridge University Press

Abstract

The English Profile Wordlists provide a web resource showing the most common words and phrases in use by learners of English. Designed for use by language professionals from teacher trainers to examination writers, and with input from contributors and reviewers from around the world, the Wordlists document learner vocabulary through corpus-informed research. Vocabulary is graded according to the four Basic and Independent User levels of the Common European Framework of Reference (levels A1–B2) and is selected following a 'can-do' rationale – focussing on what learners do know rather than on what they should know. Preview versions of the British and American English Wordlists are available through public preview, offering detailed search functionality. There are further development opportunities to extend the Wordlists for young learners and for Business English, as well as to cover vocabulary at the C levels.

Keywords: language learning, vocabulary, CEFR, corpus, wordlists, affixation

1. Introduction

In March 2007, Cambridge University Press approached me about working on a new project within the English Profile Programme which would investigate learner vocabulary, largely through corpus-informed research, in order to grade it at four different Common European Framework of Reference (CEFR) levels (A1–B2). Initially, the *English Profile Wordlists* were seen as a resource for authors and editors, syllabus designers, teacher trainers, researchers and examination writers. Cambridge ESOL, an English Profile partner, has actively encouraged the development of the Wordlists, viewing them as a useful adjunct to their own Vocabulary Lists.

Working with John Sinclair and Gwyneth Fox on the COBUILD project through the 1980s gave me a lot of experience of corpus-informed research, and my fifteen years as a Cambridge examiner have exposed me to tens of thousands of candidate answers from all over the world, at different CEFR levels. I have also written corpus-informed textbooks for learners at A2 to C2 levels and have worked on Cambridge ESOL examinations as an item writer since 1990, using their Vocabulary Lists in the construction of exam tasks. The Wordlists project has brought all these strands together in a unique way.

Originally conceived as a book and CD-ROM, the *English Profile Wordlists* project got underway in May 2007 and a first sample of approximately 2% of the alphabet – the letter G – was presented at the English Profile seminar meeting of 19 July 2007. Individuals attending that meeting and other reviewers around the world commented on this sample and made invaluable suggestions as to the coverage of the Wordlists and the organization of its entries. From this feedback and subsequent contributions from Cambridge University Press authors and editors, it soon became apparent that the confines of a book would restrict the scope of the Wordlists and it was decided that a web-delivered resource would be more appropriate.

In October 2009, some two and a half years since the inception, a crucial development stage in the project was reached, with the fully searchable electronic pilot version of the British English Wordlists available online to English Profile partners only, during an evaluation and validation phase. A public preview version of the letters D, J and K with both British English and American English Wordlists is available on the English Profile website, and feedback is also being collected from this version¹. The full American English version of the Wordlists is currently in preparation in New York.

2. The A and B Common European Framework of Reference levels

The macro CEFR levels A (Basic User) and B (Independent User) subdivide into the four levels A1, A2, B1 and B2. There seem to be key differences between a Basic User and an Independent User in terms of their vocabulary development. As the CEFR describes, the Basic User encounters familiar topics and through these, starts to acquire basic lexical sets, together with the top senses of high-frequency words. Many of these words are the grammatical building blocks that enable learners to structure their language at phrase and sentence level: *a, the, some, any, and, but, if, so* and so on. There is also some learning of formulaic phrases at A1 and A2 level, for example *See you soon, Excuse me, No thanks, Take care*.

The Independent User starts to deal with less familiar topics and thus more topic sets are added, along with additional senses of ‘known’ words. It should be noted that the B levels in particular are broad bands, usually covering more than a single year of language tuition. By the higher end of B1 and into B2, learners appear to be acquiring more ‘chunks’ at phrase level and there is gradual use of frequent collocations and certain phrasal verbs.

Various sources already exist that provide detailed information about these CEFR levels, although rather less documentation is available on the precise vocabulary that attaches to these levels. Nevertheless, the CEFR, although primarily a skills-based description, has been a good starting point for the Wordlists. The T-series titles by J. A. van Ek and John Trim *Waystage, Threshold, Vantage* and *Breakthrough* (now available on the English Profile website) contain much of relevance, including a word index at Waystage and Threshold levels, ‘lexical exponents of specific notions’ at Vantage level, and ‘theme-related can-do statements and vocabulary’ at Breakthrough level.

Cambridge ESOL examinations are reliably pegged to the CEFR and, at A1, A2 and B1 levels test writers consult Vocabulary Lists to check whether it is permissible to test a word

¹ Visit englishprofile.org and go to English Profile Wordlists

at a given level. These lists illustrate the span of learner understanding of vocabulary within the range of topics covered by the relevant examinations and have therefore been useful documents for the Wordlists project. The KET (A2) and PET (B1) Vocabulary Lists have been in use for more than fifteen years and are in part derived from the van Ek and Trim titles mentioned above. These lists are updated annually, with reference to native speaker and learner corpus evidence. Cambridge ESOL has also developed a wordlist at A1 level for young learners – the Starters and Movers lists in the Young Learners Examinations (YLE) handbook. However, the YLE lists have been of marginal use in the Wordlists compilation, due to their focus on a single, young age group.

3. Rationale

The core objective of the Wordlists project has been to establish which words are commonly known by learners around the world at the CEFR levels A1 to B2, and to assign these levels not merely to the words themselves but to their individual meanings. In other words, rather than providing a syllabus of the lexis that learners *should* know, the Wordlists project has concentrated on verifying what they *do* know. This ties in with the CEFR emphasis on ‘can-do’ statements and also reflects research being done by other English Profile partners on criterial features at each of the six levels.

What is meant by ‘know’ in this context? In early discussion at the sample stage, before the main compiling commenced, the question of possible disparity in level between receptive and productive language was raised, both internally and by some reviewers. This is an issue that has often been researched over the years, (e.g. Melka, 1997), and yet arguably remains intangible and unproven. Moreover, much will depend on learning styles and, indeed, classroom dynamics – the extent to which opportunities are given for productive use. In exam classes, balanced preparation of all four skills has to take place if candidates are to succeed and so it is advisable to encourage learners to actively use the words and meanings they are exposed to in textbooks and classroom teaching. In general, modern communicative classrooms provide more consistent opportunities for actively using new language than a generation ago and, especially in relation to vocabulary, the prevailing advice seems to be ‘use it or lose it’. So perhaps the gap between receptive understanding and productive use is not as wide as some people have claimed.² The evidence in the *Cambridge Learner Corpus* suggests this might be the case, and illustrations of relevant findings will be given later in this article.

4. Compiling methods

Most of the words and phrases covered in the Wordlists are derived in the first instance from lexicographic research into frequency carried out by Cambridge University Press, which has informed both the second and third editions of the *Cambridge Advanced Learner’s Dictionary*

² See Francine Melka in *Vocabulary: Description, Acquisition and Pedagogy* for a fuller discussion of this area.

(CALD). This research made use of the *Cambridge International Corpus*, a billion words of written and spoken English text taken from a very wide range of sources, in order to decide on the relative frequency of senses for words within the top 5,000–6,000 words in English. To this end, lexicographers manually counted concordance lines for these words and, according to the number of occurrences of a given sense, assigned one of three relative frequency levels to it: E, I and A (Essential, Improver and Advanced), where E represents the highest frequency of occurrence.

At the outset of the *English Profile Wordlists* project the dictionary entries for all words/senses tagged E, I or A were placed in a database, and formed the starting point for the compiling process. It was thought likely that anything tagged ‘E’ or ‘I’ would fall within the four CEFR levels A1–B2, but that the ‘A’ category might well contain many words above B2 level and would need careful scrutiny and cross-checking with other sources, including other learner dictionaries and classroom materials. Following this analysis, around 55% of the ‘A’ words have been included in the Wordlists data up to B2 level. The remaining ‘A’ words will shortly be revisited in the context of the C levels Wordlists development.³

In the event, a few ‘I’ words have also been omitted from the Wordlists due to lack of learner evidence worldwide. For example, although the ‘I’ verb *eliminate* is fairly frequent in native speaker corpora⁴, its use at FCE/B2 level is largely limited to L1 first language speakers. It also seems more appropriate to the C levels in terms of its register and use.

An analysis based solely on native speaker frequency does not capture certain words that are useful to learners and which have a high frequency in the language classroom. The *Cambridge Learner Corpus*, a collection of student writing which currently comprises over 40 million words, with more than 150 nationalities represented, provided substantial evidence of these words. Wordlists from course books and other materials for learners have also been scrutinised in order to support the inclusion of words or senses in the *English Profile Wordlists*. Some examples of these words are: *album*, *download*, *guidebook*, *haircut*, *questionnaire*, *skateboard*, *trainer*. Most of these additions are nouns and either represent lifestyle choices that are important to learners – downloading music or skateboarding, for example – or are words that come directly from the teaching and learning experience, as in *questionnaire*. All the examples above are listed in either the KET or PET Vocabulary List.

One further source that was consulted throughout was the *Cambridge English Lexicon* (Hindmarsh, 1980). This landmark title in English Language Teaching was compiled by Roland Hindmarsh and involved many years of painstaking and detailed work. First published in 1980 (and out of print for many years), it has proved to be a very helpful guide, as it too was organized at sense level. The *Lexicon* effectively spanned all six CEFR levels, although its primary purpose was to ‘produce a lexicon worth teaching and learning at the intermediate level of FCE’, where level 1 would now correspond to A1, level 2 to A2, levels 3 and 4 to B1, and level 5 to B2. Hindmarsh then assigned levels 6 and 7 to the harder senses of the words he included within FCE level and, broadly speaking, his level 6 corresponds to C1 and level 7 to C2. It should be noted that Hindmarsh had very little computational support

³ Cambridge University Press is intending to extend the scope of the Wordlists to cover C1 and C2 levels.

⁴ It is also given two stars in the *Macmillan English Dictionary*, although is not highlighted in the *Oxford Advanced Learners Dictionary*.

beyond access to basic word frequency lists and yet, where the language has not changed, his intuitions seem extraordinarily accurate and well-judged.

Interestingly, the number of entries in the Wordlists pilot version is quite close to Hindmarsh's final total of 4,500. There are currently 4,666 entries in all, and the addition of 'new' words at each level which add up to this total, is given below:

Words at A1	601
New words at A2	925
New words at B1	1,429
New words at B2	1,711

These totals will be subject to change over the following months, as the Wordlists are evaluated and validated.

5. Structure of a Wordlists entry

Although the Wordlists are primarily organized by conventional dictionary headword, many fixed and semi-fixed phrases, phrasal verbs and other multi-word 'chunks' have been dealt with prominently as separate senses, reflecting current approaches to vocabulary learning and teaching (see O'Keeffe *et al.*, 2007). In each headword entry senses are ordered by CEFR level and within that according to their relative frequency in native speaker use.

Each entry uses reliable information from Cambridge dictionaries⁵, including audio and phonetic pronunciations, grammar and usage information, guide words to senses, a short definition, and examples of typical use, which often highlight important collocations. These examples are seen as being within B2 level, but do not necessarily reflect the actual CEFR level assigned to the word or phrase they are exemplifying. Generally, the lower level examples come first.

The majority of senses also include a learner example taken from the *Cambridge Learner Corpus*, showing typical use of the word or phrase. Any errors made by the writer that are peripheral to the use of the target word or phrase are corrected within square brackets. The learner example is presented in a box which also provides information about the examination taken by the writer, the CEFR level of that examination, and the candidate's first language. Every effort has been made to spread the selection of learner examples across a wide range of first languages and to avoid examples of cognate use that might misrepresent the norm in terms of average ability around the world. As already mentioned, the intention is to make the Wordlists as widely applicable as possible, rather than viewing them as a Eurocentric tool. The CEFR itself is used in many different teaching and learning contexts all over the world, so any attempt to quantify vocabulary within its levels should mirror this reality.

Wherever possible learner examples have been selected at the same level as that assigned to the word or sense. This is very often the case within the B levels. However, because of the current lack of substantial amounts of data at A1 level and due to the constraints of the

⁵ Both the *Cambridge Advanced Learner's Dictionary* and the *Cambridge Learner's Dictionary* have been used, but examples come mainly from the latter, which targets learners at a lower level.

ESOL KET writing task at A2 level, it has sometimes been necessary to move up to the CEFR level above in order to find a suitable example. As the *Cambridge Learner Corpus* grows⁶ further searches will be made in an attempt to rectify this and to select learner examples for any senses where they are missing.

At the same time, it has to be remembered that the *Cambridge Learner Corpus* is a corpus of *written* learner English, so spoken learner uses of words and expressions cannot be exemplified in the pilot version of the *English Profile Wordlists*. However, extensive work is now ongoing under the direction of Mike McCarthy to collect spoken learner data, consisting of both exam and other non-exam sources (see McCarthy, this issue, for a discussion of spoken fluency in the CEFR context).

6. Using the *English Profile Wordlists*

As the Wordlists are delivered electronically, the data can be searched in any number of ways. It is recognized that different users will have different requirements and constraints. So, for example, culturally sensitive words which could give offence in materials or exam tasks – *alcohol, invade, murder* – or specific senses of a word, such as the use of the verb *drink* to mean ‘to drink alcohol’, have been highlighted in the Wordlists and can be hidden in the data.

It is possible to specify certain CEFR level(s) only (the default search runs on the entire A1–B2 data). The user can key in a particular word or phrase and check its level, or access the data alphabetically by browsing a particular letter or stretch of entries. ‘Wildcard’ searches can be conducted using the asterisk: for example, all words ending in *-ing* can be found by keying in ‘*ing’. In addition, the Advanced Search facility allows for restricted searches on a specific aspect or combination of aspects, including grammar, usage, topic or affixation. So, for example, the user might want to find out which nouns are known at B2 level within the topic of crime, or want to obtain a list of adjectives at B1 level containing both the prefix *un-* and the suffix *-able*.

7. Affixation

Affixation is an area that has been of special interest to me during the research and compilation of the Wordlists, and an early decision was taken to separate out all dictionary ‘run-ons’ and treat them as potential headwords in their own right provided they were sufficiently frequent in native speaker use. An experienced lexicographer, Elizabeth Walter, carried out some research into native speaker frequency and produced suggestions as to the level of individual prefixes and suffixes based on corpus evidence and in comparison with the Bauer and Nation (1993) word families table. The ‘transparentness’ of meaning was a deciding factor in determining inclusion or otherwise within the Wordlists. Two levels for affixes were arrived at within the scope of A1–B2 learners: a small group of the most key and common affixes, likely to be known from a relatively early stage of learning, such as *un-*, *-ed* to form adjectives, *-er* to

⁶ Around three million words of candidate writing at all levels are added each year.

refer to people; above that, a somewhat larger set of affixes likely to be known by students of intermediate level and above, for example *dis-*, *re-*, *-ful*, *-less*.

Further issues arise because the application of these affixes is not always straightforward. Take the frequent suffix *-able*, which has been given two levels in the Bauer and Nation table: 3 and 6. On some occasions, the addition of the *-able* suffix is transparent, in other words it is easy to work out the meaning of the newly formed word, which reflects the core meaning ‘able to be’, for example *downloadable* from *download*.⁷ On the other hand *changeable* doesn’t mean that something is ‘able to change’, it means that it frequently *does* change. Similarly, *knowledgeable* falls outside the core meaning since it means ‘knowing a lot’.

To return to the prefix *dis-*, all of the words formed from it that are included in the current Wordlists are at the B levels, the majority at B2. At B1, there are the verbs *disagree*, *disappear* (two of the three senses included in the Wordlists are at B1), *dislike*, the adjective *disabled* and the noun *disadvantage*. As mentioned earlier, the frequency of words in native speaker corpora has often influenced their inclusion or otherwise in the Wordlists. So, for example, the verb *disbelieve* has been omitted, as the norm seems to be to use ‘not believe’ rather than ‘disbelieve’; the noun *disapproval* has also been omitted, partly on grounds of low frequency, but also because ‘double’ affixation (*dis-* plus *-al* in this case) appears to be more challenging for learners.

There are a few interesting exceptions to this, though they usually consist of the combination of a common prefix and suffix, such as *un-* plus *-ed* or *-ly*: the adjective *unexpected*, more frequent in native speaker use than the adjective *expected*, has been given B1 level in the Wordlists due to evidence in the *Cambridge Learner Corpus* of its use worldwide, whereas *expected* barely makes B2, with few learner examples. This is possibly due to its more restricted usage, coupled with the constraint of having to come before a noun. Indeed, the *Cambridge Learner Corpus* shows that at CAE/C1, learners are producing this pattern much more frequently, in phrases such as *the expected amount/level/standard*. Both adjectives are tagged in CALD as ‘I’ words, though there seems to be evidence that *unexpected* is the more frequent of the two for native speakers.

An interesting parallel with native speaker frequency can be observed in relation to members of the word family for *fortunate*: there is good evidence at KET/A2 level for the adverb *unfortunately* so it is assigned to A2 in the Wordlists, whereas the adverb *fortunately* seems to belong to B1 level⁸; the two adjectives are much less used by learners, and then only from FCE/B2 level. This reflects the relative frequency of the words in native speaker corpora: *unfortunately* is a CALD ‘E’ word, and also gets three stars in the *Macmillan English Dictionary*; *fortunately* is CALD ‘I’, while the two adjectives are only ‘A’, indicating a somewhat lower frequency of use. From this example, it can be demonstrated that learners do not always meet the ‘root’ word first, and a flexible approach can be taken to word formation in the classroom, based on frequency and usefulness. Clearly, the word *unfortunately* is of practical use to learners, in a variety of spoken contexts and written genres.

⁷ *Downloadable* has not been included in the Wordlists as it has a very low native speaker frequency, relative to the verb form.

⁸ In 35 million words of CLC data, there is one cite for *Fortunately* at KET, and this by a Spanish-speaking learner, illustrating cognate use.

8. Word family panels

As the above discussion shows, affixation is far from straightforward for learners and arguably needs to be focused on systematically at the B levels and above. For this reason, the Wordlists feature word family panels, which appear at the head of any entry where two or more related forms have been found to be within B2 level. So, for example, at the entry for the adverb *separately*, the family displayed is as follows:

Word family:
Nouns: separation
Verbs: separate
Adjectives: separate
Adverbs: separately

This family seems to be confined to the B levels, with only two senses of the adjective known at B1 and the remaining family members belonging to B2. There are however several instances where members of word families begin to be acquired in the early stages of learning and yet related words and phrases of lower frequency are only acquired much later – the adjective *different*, for example, is an A1 word, and the noun *difference* comes in at A2 (both are CALD ‘E’ words), but the verb *differ* (a CALD ‘A’ word) is only used from B2, as are the phrases *make a (big) difference* and *make no difference/not make any difference*. Interestingly, *differ* is currently included in the PET Vocabulary List, but there is an argument for removing it.

9. Prioritizing senses: the relative difficulty of senses for the verb *keep*

The verb *keep* is a very frequent word in English, occurring within the top 300 words in the frequency listing of the *Cambridge International Corpus*. It has many senses, features in many verb-noun collocations, and combines with particles to form a number of phrasal verbs. For all these reasons, I was keen to investigate it further, so chose to tackle the letter **K** early on in the compiling. In the KET Vocabulary List, its use in testing is limited to two senses, as in the examples *May I keep this?* and *Keep right!* The former sense, meaning ‘to have something permanently’, is the first sense in the Cambridge dictionaries, illustrating that it is probably the most frequent for native speakers; the latter, meaning ‘to stay’ is also high up in terms of frequency (both senses are CALD E). These two senses have been assigned A2 in the Wordlists, along with the phrase *keep sth in/on, etc* which has been added from the *Cambridge Learners Dictionary* (here, the first CALD sense has been split to provide more support to lower level learners).

One further sense which is given ‘E’ status in CALD and is very frequent in native speaker use is ‘to continue doing something without stopping, or to do it repeatedly’. This appears in the Wordlists under the phrase *keep doing sth* and has been assigned B1 level. Interestingly, there is also evidence at PET/B1 for the closely related phrasal verb *keep on doing sth*. This phrasal verb is listed in the PET Vocabulary List, where *keep* itself is listed without any sense restrictions.

As already highlighted, *keep* appears in many useful phrases and collocations, and the entry in the Wordlists is quite lengthy for this reason. Users can choose to select ‘outline view’ rather than ‘full view’, to get a briefer listing of senses, with only guide words and definitions displayed. At B1 level, there is one further sense, with the guide word DELAY, as in *I’m sorry to keep you waiting*, and the phrases *keep in touch* and *keep a/ sth secret*. There are four more phrases listed at B2 level, together with two other less frequent senses: the intransitive use to do with food, as in *This product will keep for three days if refrigerated* and the sense ‘to have and look after animals’. This then is a further refining of the information contained in the PET Vocabulary List, and provides evidence of which senses B1 level learners can cope with.

Finally in the *keep* entry, there are fifteen senses of phrasal verbs, all but three at B2. Researching the level of phrasal verbs has been problematic: coursebooks feature them, especially from B1 level, and many are listed in other sources. Hindmarsh included a phrasal verbs appendix that runs to ten pages⁹. Furthermore, although they are often quite frequent in native speaker corpora, they occur relatively rarely in the *Cambridge Learner Corpus*. This could be down to ‘task effect’, in that some of the written genres at PET and FCE do not lend themselves to the appropriate use of phrasal verbs. They may also be more commonly spoken by learners rather than written, so it will be interesting to interrogate the spoken learner corpora in due course. For the moment, of the fourteen phrasal verbs at B1 and B2 listed in the *keep* entry, only four have learner examples accompanying them. A decision on the final inclusion policy for phrasal verbs will have to be taken towards the end of the pilot phase, and it is hoped that feedback will be forthcoming in this area from users and public viewers.

10. The senses of the word *case*

At the February 2009 English Profile seminar, when compiling had reached the halfway stage, an interim report was delivered about insights and issues, from which this article gets its title. The lack of production of phrasal verbs within B2 was one such observation, along with the fact that the most frequent senses of a word are not always the first taught. A good example of this is the word *case*, whose most frequent sense for native speakers is ‘a particular situation or example of something’, as in the following dictionary examples:

Over fifty people were injured, in several cases seriously.
 I wouldn’t normally agree but I’ll make an exception in this case.
 The number of new cases of the new illness has risen.
 We have lots of applications from people who want to study here and in each case we consider the candidate very carefully.
 She was suffering from an extreme case of sunburn.

However, the *Cambridge Learner Corpus* and other checking sources have shown that learners meet the lower frequency senses of ‘container’, as in *pencil case*, and ‘bag’ (the latter in British English only) earlier than other, more frequent senses. There are practical reasons in this

⁹ Appendix Q in the *Cambridge English Lexicon*.

case – pencil cases are found in most school classrooms around the world, after all. What is perhaps more surprising is that the meaning of ‘situation’ exemplified above often seems not to be explicitly taught in coursebooks, and the sense is not currently included in the PET Vocabulary List. At the same time, Hindmarsh clearly recognized its importance in learning by allocating it the same level (Level 2) as the container sense. The sense has been assigned B1 in the Wordlists for the moment, but there may be a good case for lowering it to A2.

Another insight, which arises from browsing the *Cambridge Learner Corpus*, is the influence of the learner’s first language, especially at the A levels. This can help or hinder the learner. Close cognates, as already noted, will mean earlier than average use of a word or phrase, whereas ‘false friends’ may lead to errors or inappropriate use. One interesting example in the compiling process was the phrase *in fact*. This has been assigned B1 in the Wordlists (the less frequent *in actual fact* is B2), in spite of there being evidence for the phrase at A2, i.e. being produced in the KET writing task. In consultation with reviewers, B1 was seen as more suitable, as many of the A2 cites were using the phrase wrongly. A large proportion of these learner examples were written by first language speakers of Italian, where the word *infatti* is frequently used; a common error in the *Cambridge Learner Corpus* is the writing of the phrase as a single word ‘infact’, demonstrating first language interference.

11. Issues arising from the compiling process

Although it has proved to be a very time-consuming project, I have become more and more convinced of the desirability of reporting CEFR level for individual senses, in order to provide more solid support for language teaching professionals such as authors, editors and exam writers. This is nothing new – I am merely following in the sturdy footsteps of Hindmarsh. However, combined with this is an increased attention to phrases and collocations, which corpus evidence can readily highlight. Further work could be done in the area of collocation and this will certainly be needed in any development of Wordlists for the C levels.

One factor of current concern is the very wide age range of learners around the world, and whether a single resource such as the Wordlists can actually report on a level for all ages. The Wordlists are recommended for anyone dealing with learners aged 11 and upwards; for young learners, a different grouping of words/senses seems inevitable and has already been developed in the ESOL Young Learners Vocabulary Lists at A1 and A2.

Coupled with this is the fact that the Wordlists focus on general English and therefore cannot really help those working in Business English, or teachers involved in CLIL projects. It is hoped that ‘add-on’ lists might be developed for these and other specific domains in due course.

Additionally, the potentially restricted nature of the *Cambridge Learner Corpus* means that other learner corpora will need to be accessed in the future to verify the findings included in the Wordlists. The current development of the *Cambridge English Profile Corpus* (CEPC) is therefore an exciting one, as eventually there will be ten million words of non-exam learner data (20% spoken and 80% written). The CEPC will also include English for Specific Purposes data.

It is obvious that a lot more work is waiting to be done, both within the current levels of the Wordlists and at the C levels. Quite apart from the inclusion of spoken learner evidence just referred to, much more research could be done on affixation and its challenges for learners; the inclusion of a selected group of headwords for certain word families based on frequency is a modest first step.

Thanks to the efforts of the highly talented and diligent electronic dictionaries team at Cambridge University Press, the *English Profile Wordlists* are an extremely attractive and powerful resource. The preview version for the letters D, J and K, together with selected entries from the A–Z pilot version known as the ‘Word of the Week’ feature can be viewed now. Please take the time to send in your feedback via the website, so that the Wordlists can be further improved.

References

- Bauer, L. & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, 6 (4): 253–279.
- Hindmarsh, R. (1980). *Cambridge English Lexicon*. Cambridge: Cambridge University Press.
- Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 84–102). Cambridge: Cambridge University Press.
- O’Keeffe, A., McCarthy, M. J. & Carter, R. (2007). *From Corpus to Classroom*. Cambridge: Cambridge University Press.