## The BNC/COCA word family lists

This description of the BNC/COCA lists has been kept as brief as possible. For more information on the lists and their use see Nation (2016) *Making and Using Word Lists for Language Learning and Teaching*. John Benjamins, Amsterdam.

## Versions of the lists and referencing the lists

Use the following method to cite/reference the BNC/COCA lists according to the APA style guide:

Nation, I.S.P. (2017). The BNC/COCA Level 6 word family lists (Version 1.0.0) [Data file]. Available from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Nation, I.S.P. (2017). The BNC/COCA Level 3 partial word family lists (Version 1.0.0) [Data file]. Available from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Major changes adding new 1000 word levels and reallocating substantial numbers of families to lists will be considered a change at the 1.0.0 version level. Making smaller reallocations of words to existing lists and families will be considered a change at the 0.1.0 version level. Minor correction of a few words or the addition of words to existing families will be considered at change at the 0.0.1 version level.

## A short history of the lists

The work on the lists began shortly after the early version of the *Range* program (*Vords*) was designed by Alex Heatley in the late 1980s (see Hwang & Nation, 1989). The *Vords* program counted the frequency of words in texts using word families. The lists used were the *General Service List of English Words* by Michael West divided into two lists, and the *University Word List* (Xue & Nation, 1984). *Vords* was later named *VocabProfile*. It was not until after the publication of the Bauer and Nation (1993) paper on word families, that a consistent definition of word families was used to make lists. In 1998, when Averil Coxhead was working on the building the *Academic Word List,* Alex Heatley redesigned *Vords* to count range (the number of different texts a word occurred in) as well as frequency. Alex called the program *FVords* (Family Vords), but having had enough of computer-programmer humour, we renamed it *Range*. Averil Coxhead's work (2000) highlighted deficiencies in the *General Service List* and work began on designing new word family lists. This was very time-consuming work, with each 1000 word family list taking well over a month to create. By 2003, there were three new1000 word lists drawing on Leech, Rayson and Wilson's (2001) list of lemmas in the British National Corpus (Nation, 2004). In 2004, Chris Andreae wrote the *AffixAppender* program that automated the building of word families. The creation of word families still required a lot of manual checking. By 2006, there were fourteen specially created word family lists (Nation, 2006), by 2008 twenty word family lists, and by 2018 twenty-eight word family lists.

## The purpose of the lists

The BNC/COCA lists are designed primarily for learners of English as a foreign language. Because of this, most of the 111 words of the survival vocabulary are in the first 1000, and the bulk of the AWL (319 families) is in the 3rd 1000 (473 of 570 families in the 1st to 3rd 1000 lists). In a set of lists designed to reflect native-speaker vocabulary size growth, most of the AWL would come around the 4th or 5th 1000. Lists designed to reflect native-speaker growth also need to reflect children's interests and learning opportunities, particularly the vocabulary of children's stories, children's movies and TV, and fascinating topics such as dinosaurs.

Lists designed for learners of English as a foreign language need to reflect opportunities for use (foreign travel, study in English, the internet) and opportunities for learning (graded readers and course books), making sure that words needed for these purposes occur early in the lists.

The first 9000 words in lists designed for learners of English as a foreign language and lists designed to reflect native-speaker vocabulary growth are likely to differ in content by only a small amount, although the sequencing of the words in those lists would differ more.

When making the Picture Vocabulary Size Test (PVST), the BNC/COCA lists were not used. Instead, a corpus of children's reading material and an adult spoken corpus were used to create new lists up to the 6000 word level, and these lists were used to sample from for the test. Using the BNC/COCA lists would have resulted in a test that did not reflect the vocabulary knowledge of young native-speakers.

## The contents of the lists

The BNC/COCA word family lists consist of 28 word family lists containing word families based on frequency and range data. Five additional lists are (1) an ever-growing list of proper names, (2) a list of marginal words including swear words, exclamations, and letters of the alphabet, (3) a list of transparent compounds, (4) a list of acronyms, and (5) a list of foreign words. In the lists for AntWordProfiler, each list has a name which describes its content. In the lists for Range, because of the requirements of the Range program, each list has a fixed name – basewrdx.txt, where x is a number. Basewrd1 is the first 1000 words, basewrd2 the second 1000 and so on. Basewrd31 contains proper nouns, basewrd32 marginal words, basewrd33 transparent compounds, basewrd34 acronyms, and basewrd35 foreign words. More detail on these additional lists and the word family lists can found in Nation (2016).

The lists are saved in UTF-8, without BOM (choose under Encoding in Notepad ++).

## Programs for using the lists

The best program for using the lists for the analysis of vocabulary in texts is AntWordProfiler which is available free from Laurence Anthony's web site (http://www.laurenceanthony.net/software/antwordprofiler/).

The lists were originally made to be used with the Range program, but the Range program has not been updated for many years and I now encourage the use of

AntWordProfiler. AntWordProfiler is easy to use, well supported and does everything that Range could do plus a lot more.

## The making of the lists

### *The 1st 1000 and 2nd 1000 word family lists*

The first two 1000 word family lists were made using a specially designed 10 million token corpus. Six million tokens of this corpus were spoken English from both British and American English (see Corpus/PN corpus for 2000) as well as movies and TV programs. The written sections included texts for young children and fiction (see Table 1).

Table 1: The corpus used for the first two 1000 word family lists

| **US** | **Tokens** | **UK/NZ** | **Tokens** |
|---|---|---|---|
| **Spoken** | | | |
| 1 AmNC spoken face to face, telephone 1 | 1,107,602 | 4 BNC 1 | 1,036,097 |
| 2 AmNC spoken face to face, telephone 2 | 1,029,831 | 5 BNC 2 | 1,125,523 |
| 3 Movies and TV | 1.000,000 | 6 BNC Plus half of WSC | 1,132,620 |
| **Written** | | | |
| 7 AmNC written fiction, letters 1 | 1,145,081 | 9 School journals | 1,028,842 |
| 8 AmNC written fiction, letters 2 | 939,407 | 10 BNC fiction | 1,040,204 |

This unusual step of creating a special corpus for the first 2000 word families was followed because the previous lists made from the British National Corpus (BNC) were so strongly influenced by the written formal nature of the BNC corpus that they were not suitable lists for creating language courses or graded reader lists (see Nation, 2004). Very common words in spoken English like *alright*, *pardon*, *hello, dad, bye* could then be included in the high frequency words. Other arbitrary adjustments included putting all the word forms of numbers (*one*, *two*, *hundred*) and weekdays in the 1st 1000, and the months of the year in the 2nd 1000, even though their frequency did not always justify this. The goal was to have a set of high frequency word lists that were suitable for teaching English as a foreign language and language course design.

### *The 3rd 1000 onwards*

The remaining 1000 lists were made by using COCA/BNC rankings in data kindly provided by Mark Davies (Davies COCA BNC.xls) after removing my specially created first 2000 word families.

## Word families

The criteria used to make word families were based on Bauer and Nation's (1993) level 6, which includes all the affixes from levels 2 to 6 (see Table 2).

Table 2: Word family levels

**Level 1**

A different form is a different word. Capitalization is ignored.

**Level 2**

Regularly inflected words are part of the same family. The inflectional categories are - plural; third person singular present tense; past tense; past participle; -ing; comparative; superlative; possessive.

**Level 3**

-able, -er, -ish, -less, -ly, -ness, -th, -y, non-, un-, all with restricted uses.

**Level 4**

-al, -ation, -ess, -ful, -ism, -ist, -ity, -ize, -ment, -ous, in-, all with restricted uses.

**Level 5**

-age (leakage), -al (arrival), -ally (idiotically), -an (American), -ance (clearance), -ant (consultant), -ary (revolutionary), -atory (confirmatory), -dom (kingdom; officialdom), -eer (black marketeer), -en (wooden), -en (widen), -ence (emergence), -ent (absorbent), -ery (bakery; trickery), -ese (Japanese; officialese), -esque (picturesque), -ette (usherette; roomette), -hood (childhood), -i (Israeli), -ian (phonetician; Johnsonian), -ite (Paisleyite; also chemical meaning), -let (coverlet), -ling (duckling), -ly (leisurely), -most (topmost), -ory (contradictory), -ship (studentship), -ward (homeward), -ways (crossways), -wise (endwise; discussion-wise), anti- (anti-inflation), ante- (anteroom), arch- (archbishop), bi- (biplane), circum- (circumnavigate), counter- (counter-attack), en- (encage; enslave), ex- (ex-president), fore- (forename), hyper- (hyperactive), inter- (inter-African, interweave), mid- (mid-week), mis- (misfit), neo- (neo-colonialism), post- (post-date), pro- (pro-British), semi- (semi-automatic), sub- (subclassify; subterranean), un- (untie; unburden).

**Level 6**

-able, -ee, -ic, -ify, -ion, -ist, -ition, -ive, -th, -y, pre-, re-.

The word families were developed over many years and low frequency family members continue to be added to the existing families.

Dang and Webb (2016) carried out a study of word lists which included the BNC/COCA lists. Their study showed that the BNC/COCA lists performed well on both spoken and written texts in comparison with other lists. See Nation (2016, Chapter 13) for an evaluation of the lists. Dang, Webb, & Coxhead, (2020) looked closely at the first 2000 words of the BNC/COCA lists compared with other lists and found that they performed well and were evaluated highly by teachers.

## The BNC/COCA Level 3 partial lists

There is a lot of useful debate and research about the appropriate level of word family to use for text analysis – lemmas, flemmas, level 3 word families, level 6 word families. To contribute to this debate I have made word lists, called the BNC/COCA Level 3 partial word lists which use the inflections of English and four derivational affixes from Level 3 of Bauer and Nation (1993) – *un* (not)-, *-ly* (making adverbs), *-er* (someone who …; does not include *–or* as in *actor*), *-th* (only for ordinal numbers) according to Bauer and Nation Level 3 restricted uses. These lists represent a next step from flemmas (lemmas where a family can contain different parts of speech). The lists are sorted on family frequency using a 14 million corpus made of 14 one million subcorpora including both spoken and written English. For learners who can handle inflections, these four derivational affixes should not be too big a step and could easily be the focus of a small amount of deliberate teaching and learning.

## Making your own lists

It is easy to make different lists and to add to the families on the existing lists. It is highly recommended that Notepad++ be used to do this and that the lists are saved in UTF-8, without BOM format (choose under Encoding in Notepad ++). Notepad++ is a very powerful, freely available text processing program.

When making lists, just use the same format as the existing lists. The same words should not appear in two or more different lists that are used at the same time.

## Adding to the lists

The BNC/COCA lists will always be a work in progress as the number of lists increases, as new word families are added to the lists, as new members are added to existing families, and as errors are corrected.

## References

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253-279.

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

Dang, T. N. Y., & Webb, S. (2016). Evaluating lists of high-frequency words. *ITL-International Journal of Applied Linguistics, 167*, 132–158.

Dang, T. N. Y., Webb, S., & Coxhead, A. (2020). Evaluating lists of high-frequency words: Teachers’ and learners’ perspectives. *Language Teaching Research*, 1-25.

Hwang, K., & Nation, P. (1989). Reducing the vocabulary load and encouraging vocabulary learning through reading newspapers. *Reading in a Foreign Language, 6*(1), 323-335.

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a Second Language: Selection, Acquisition, and Testing* (pp. 3-13). Amsterdam: John Benjamins.

Nation, I. S. P. (2016). *Making and Using Word Lists for Language Learning and Testing*. Amsterdam: John Benjamins.

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

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3*(2), 215-229.

(This document was revised on 18 May 2020)