The BNC/COCA word family lists
(17 September 2012)

The BNC/COCA word family lists consist of 29 word family lists. Twenty-five of the lists contain word families based on frequency and range data. The four 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, and (4) a list of abbreviations. 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. Basewrd26-30 just contain one nonsense word each. They were made to provide space for additional lists and to avoid having to keep changing the names of the proper nouns etc lists. Basewrd31 contains proper nouns, basewrd32 marginal words, basewrd33 transparent compounds and basewrd34 abbreviations. More detail on these additional lists can found in Nation and Webb (2011: Chapter 8).

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

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		**UK/NZ**	
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 were so strongly influenced by the written formal nature of the 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 and 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 several years and low frequency family members continue to be added to the existing families.

The nature of the families

The word lists were made to be used with the AntWordProfiler and Range computer programs and these program cannot distinguish between homonyms like *Smith* (the family name) and *smith* (blacksmith) and *March* (the month) and *march* (as soldiers do). Thus when the program runs, these uses are not distinguished and would be counted in the same family and as the same type. There was an attempt to deal with this wherever possible. *Marched, marching, marches, marcher, marchers* etc for example were put in one family and *March* into another. This does not completely distinguish the homonyms, but it is a step towards doing so.

The high frequency word families tend to be quite large as it appears that higher frequency stems generally can take a greater range of affixes than lower frequency words. For example, the high frequency word family *nation* has the following members *nations, national, nationally, nationwide, nationalism, nationalisms, internationalism, internationalisms, nationalisations, internationalisation, nationalist, nationalists, nationalistic, nationalistically, internationalist, internationalists, nationalise, nationalised, nationalising, nationalisation, nationalize, nationalized, nationalizing, nationalization, nationhood, nationhoods*.

The word family lists group items together that would be perceived as the same words for the receptive skills of listening and reading. If word lists were made for productive purposes, for speaking and writing, the lemma would be the largest sensible unit to use. Some researchers argues for the word type.

The word lists contain compound words but they do not contain phrases. *According to* or *au fait*, for example, might be best counted as a unit, but in the lists the unit is the single word.

The validity of the BNC word family lists

There are ways of checking whether the word family lists are properly ordered. From the 1st 1000 to the 25th 1000, the number of tokens, types, and families found in an independent corpus should decrease. That is, when the lists are run over a different corpus from the BNC or COCA, the 1st 1000 word family list should account for more tokens, types and families than the 2nd 1000 family list does. Similarly, the 2nd 1000 word family list should account for more tokens, types and families than the 3rd 1000 family list does and so on. While this does not show that each word family is in the right list, it does show that the lists are properly ordered. Table 3 presents such data using the Range output from the Wellington Written Corpus.

Table 3: Tokens, types and families in the Wellington Written Corpus

WORD LIST	TOKENS/%	TYPES/%	FAMILIES
one	772697/75.22	4762/11.74	999
two	91545/ 8.91	4299/10.60	999
three	53591/ 5.22	3903/ 9.62	999
four	17967/ 1.75	2853/ 7.03	995
five	10899/ 1.06	2336/ 5.76	981
six	7267/ 0.71	1986/ 4.90	950
seven	4513/ 0.44	1564/ 3.86	904
eight	4313/ 0.42	1336/ 3.29	853
nine	2592/ 0.25	1089/ 2.68	760
ten	2005/ 0.20	920/ 2.27	700
11	1533/ 0.15	721/ 1.78	585
12	1063/ 0.10	589/ 1.45	489
13	832/ 0.08	438/ 1.08	391
14	737/ 0.07	346/ 0.85	304
15	531/ 0.05	276/ 0.68	246
16	443/ 0.04	220/ 0.54	198
17	628/ 0.06	194/ 0.48	173
18	250/ 0.02	127/ 0.31	117
19	247/ 0.02	104/ 0.26	101
20	269/ 0.03	104/ 0.26	89
21	132/ 0.01	79/ 0.19	74
22	130/ 0.01	63/ 0.16	59
23	80/ 0.01	43/ 0.11	40
24	296/ 0.03	52/ 0.13	48
25	134/ 0.01	31/ 0.08	29
26	0/ 0.00	0/ 0.00	0
27	0/ 0.00	0/ 0.00	0
28	0/ 0.00	0/ 0.00	0
29	0/ 0.00	0/ 0.00	0
30	0/ 0.00	0/ 0.00	0
31	30991/ 3.02	3844/ 9.48	3691
32	3111/ 0.30	90/ 0.22	33
33	4203/ 0.41	1200/ 2.96	926
34	1380/ 0.13	191/ 0.47	188
not in the lists	12819/ 1.25	6803/16.77	?????
Total	1027198	40563	16921

A second way of checking the validity of the lists is to look at the total number of types in each list. Low frequency words tend to have less family members than high frequency words, so even though the number of families in each list is the same, one thousand, the number of types should be less. Table 4 contains this data.

Table 4: The number of types (family members) in each of the twenty-five 1000 word family lists

1	6857	6	4104	11	2941	16	2086	21	1651
2	6374	7	3679	12	2754	17	2076	22	1539
3	5880	8	3417	13	2415	18	1933	23	1394
4	4863	9	3196	14	2299	19	1872	24	1296
5	4294	10	2985	15	2283	20	1820	25	1675

The 1st 1000 word families contains 6,857 word types, an average of 6.857 per family as each list contains exactly 1000 word families. There is decrease in word types from one list to the next. The families in the newly created 25th 1000, which was made from a dictionary, may have been more diligently made than the preceding lists.

A third way of checking the validity of the lists is to make sure that no wide range, high or mid-frequency words are missing from the lists. To check this, the lists were run over a wide range of different corpora, existing lists, and texts. No frequent, wide range word families were missing.

Words not in the lists

Table 5: The percentage amounts of different kinds of word types in the British National Corpus and not in the twenty 1000 word family British National Corpus word lists and additional lists

Kinds of words	**Recurring words**	**One-timers**	**Total %**	**Projected total**	**Examples**
New words	12	6	18	49,101	tucuxi, pericentric, escritoire, polyacrylonitrile, trochar, pancreata
Proper nouns	23	25	48	130,937	Southwick, Akrokorinth, Frakes, Aalberse, Stycar, Thucyd, Wellferon, Mlungisi
Foreign words	2	2	4	10,911	nationaux, panellinion
Low frequency family members	2	4	6	16,367	obeyance, velcros, realizational, ungrouped
Transparent compounds	2	2	4	10,911	lockgates, poolrooms, countertop, duststorm
Acronyms, abbrev	5	2	7	19,095	USYN, MLD, EMW,
Alternative spellings	0	3	3	8,183	velem, Hindostani, cronicles
Letters with numbers	1	3	4	10,911	AW17, MX300, PTFMA215
Exclamations	2	0	2	5,455	fattafattafatta, cheeeeee
Errors	0	4	4	10,911	approprite, gorups, dispoal
Total	49	51	100	272,782	

There are 272,782 word types in the British National Corpus that are not in the first 20 word lists used with the Range program, plus a list of proper nouns, a list of transparent compounds, and a list of exclamations, hesitations and other spoken marginal words.

Note in Table 5 that almost half of the different words are proper nouns. Four percent are foreign words, and 6% are low frequency members of word families already in the 20 one thousand word lists. Ideally, these family members should be added to the families in the existing lists.

The main point of the table is to show that the new words (49,101) plus the 20,000 in the word lists total around 70,000 word families which is a figure not too far from Nagy and Anderson’s (1984) estimates, and the number of words in most reasonably sized non-historical dictionaries. The reason for distinguishing recurring words (those occurring 2 times or more in the British National Corpus) from those occurring only once in the corpus (one-timers) is to show that the proportion of new words in the one-timers is half that in the recurring words.

References

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., & Webb, S. (2011). *Researching and Analyzing Vocabulary*. Boston: Heinle Cengage Learning.