Volume 2 | Issue 2 | 2022

http://langedutech.com

Language Education & Technology

Parts of Speech Distribution in the BNC-COCA Word Lists

Meral Öztürk[a] *

[a] Emerita, Bursa Uludağ University, Türkiye; http://orcid.org/0000-0002-5184-9851

Suggested citation: Öztürk, M. (2022). Parts of Speech Distribution in the BNC-COCA Word Lists. *Language, Education, & Technology (LET Journal), 2*(2), 128-140.

Article Info

Date submitted: 22.08.2022

Date accepted: 19.09.2022

Date published: 19.09.2022

Research Article

Abstract

The main strength of the BNC-COCA word frequency list is also its major weakness. The frequency-based organisation of the list is a strength as it allows a systematic and unbiased selection of target words for a vocabulary size test. Using frequency as the sole criterion for target word selection, however, is a weakness because lexicons are much more heterogeneous with a variety of factors affecting difficulty of words. The present paper is an attempt to augment the lists with parts of speech information. The words in the first fourteen baseword lists were tagged for parts of speech and counted. The results revealed 58% of the words in the list to be nouns, 21% verbs, 18% adjectives and only 3% function words. 1K level had a different distribution from other levels due to an uncharacteristically high proportion of function words (19%). It was also found that the relative distribution of the content word categories varied with frequency level. As such, the data did not support the use of a fixed ratio in size tests for all frequency levels. Item numbers for individual frequency levels were proposed for a 140-item vocabulary size test on the basis of the variable ratios obtained in the present data.

Keywords: BNC-COCA word lists, parts of speech distribution, vocabulary size tests, content validity

1. Introduction

The BNC-COCA word list (Nation, 2017) is one of the most comprehensive word lists of the English language based on modern corpora, and it has been used as a source list from which target words for several vocabulary size and vocabulary levels tests were drawn (Nation & Beglar, 2007; Coxhead, Nation & Sim, 2014; McLean & Kramer, 2015; McLean, Kramer & Beglar, 2015).

The content validity of these tests depends crucially on the validity of the source word list as well as the validity of the sample drawn from the list. The source list would be valid to the extent that it is representative of the target lexicon which the learners aim to acquire. The sample would be valid if it contains a sufficient number of words which are representative of the word list. The distinction between the validity of the source list and of the sample corresponds to Bachman's two aspects of content validity (1990, p. 244): content relevance and content coverage respectively. While it is safe to assume the BNC-

* Meral Öztürk. ELT Department, Bursa Uludağ University, Türkiye.
e-mail adress: mozturk@uludag.edu.tr

COCA list to have content relevance and thus, generally valid for most language learning purposes, it offers limited help in ensuring adequate content coverage and validity of the sample. The list is organised homogeneously in terms of a single criterion (i.e. frequency) although the natural language lexicons are known to be much more heterogeneous.

Frequency (alongside range and dispersion) has been the main criterion in the compilation as well as in the organisation of the BNC-COCA list. This focus on frequency is based on the idea of usefulness in using the language, which goes back to the works of Thorndike (1921) and West (1953). More frequent words are thought to be more useful because learners are more likely to encounter them when they read or listen to the target language, and more likely to need to produce them themselves. Therefore, it makes sense to prioritize vocabulary teaching by focusing on the higher frequency words and to check if they are learnt in vocabulary size and levels tests.

The frequency-based organisation of the BNC-COCA list indeed allows for valid sampling for a vocabulary proficiency test with respect to this variable. The list is divided into 1,000-word sublists (currently 28, Nation, 2016, p.132) ordered in frequency with the first sublist (1K list) containing the highest frequency words while the internal organisation of each sublist is alphabetical. From these lists, it is possible to draw an unbiased sample of target words for a vocabulary size or levels test through random sampling of a set number of items from each sublist. This sampling technique (stratified random sampling) results in a sample which is representative of the target word list with respect to frequency as there will be a similar number of items from all frequency levels, which, in turn, should allow a more accurate estimate of the learners' vocabulary size or current mastery level.

While frequency is important, its use as the sole criterion in item selection in current vocabulary size tests has recently been criticised. Hashimoto (2021) and Hashimoto & Egbert (2019) correlated scores on the Vocabulary of the American English Size Test and the rank frequency of the target words in the COCA corpus and found only a moderate correlation between the two (r=.499). This indicates there are other factors at play that determine whether a given test word will be known to the learners. We need to be aware of these factors and control for them in our tests if we want our test scores to be valid and reliable. Factors that arise from the learning context and particular characteristics of the learners may not be considered in a general vocabulary size test for practical reasons. Factors that relate to test words, however, can be controlled for.

For a valid sample of words which is representative of the target word list, we need to consider several word-related criteria in addition to frequency. The sample needs to be a small replica of the target word list in terms of factors that affect the difficulty of words such as cognateness, pronounceability, length, concreteness, parts of speech, etc. If these factors are not observed, the sample might end up being more difficult or easier than the target list. A size or levels test based on such a sample will, in turn, make an incorrect estimate of the learner's overall vocabulary size or their knowledge of individual levels. For example, a sample that contains many more L1 cognates than the target list will make an easy test and inflate the estimates. To ensure comparable proportions of words in the sample, we need to know the distribution of the target list with respect to these variables. The present study will provide this information for one factor, i.e. parts of speech, in the BNC-COCA list.

2. Literature Review

2.1. Difficulty of Words in Different Parts of Speech

English is said to have eight parts of speech: four in the major word classes which are also called content words or lexical words (nouns, verbs, adjectives, and adverbs) and four in the minor word classes alternatively called function words or grammatical words (determiners, modals, conjunctions, and

prepositions). These parts of speech distinctions will be relevant in target word selection for a vocabulary size test only if we can be sure that there are differences in learning between different parts of speech.

Parts of speech effects have been observed in a number of studies on L2 vocabulary learning: incidental learning of vocabulary from reading (Horst & Meara, 1999) and from listening (Van Zeeland & Schmitt, 2013), learning from vocabulary study using computer software (Tschichold, 2013; Ellis & Beaton, 1993) and learning from bilingual word lists (Rodgers, 1967; Glanzer, 1962; Morgan & Bohnam, 1944). In all of these studies, nouns were found to be generally easier than other major word classes. Adverbs were often the most difficult (Morgan & Bohnam, 1944; Horst & Meara, 1999) while verbs and adjectives were usually in between with no significant difference between them (Horst & Meara, 1999; Van Zeeland & Schmitt, 2013). In general, function words were more difficult than content words when presented in isolation, but they were easier when surrounded by context (Glanzer, 1962). Of the function words, conjunctions were among the hardest to learn (Glanzer, 1962; Morgan & Bohnam, 1944; Rodgers, 1967).

Several other studies failed to find a significant effect of parts of speech on the learning of L2 vocabulary: incidental learning from TV viewing (Puimege & Peters, 2019) or explicit vocabulary study using computers (Barclay & Pellicer-Sánchez, 2021; Ludington, 2015). Furthermore, some of the studies cited above in favour of parts of speech effects were confounded by a word length effect (e.g. Rodgers, 1967). Still, the differences reported among parts of speech categories can be considered sufficient to warrant control of this variable in vocabulary size tests. It's better to be safe than sorry.

Control in vocabulary experiments is usually exercised by selecting an equal number of items from each category. This, however, wouldn't work in a vocabulary size test. This is because words from different parts of speech are distributed unequally in the lexicon: we have very few items in the function word categories (a couple of hundred) in contrast to thousands in the content word categories. The number of nouns, verbs, and adjectives are very unlikely to be equal, either. Sampling equally from these categories would result in a very unrepresentative sample and a size test poor in content validity. Accordingly, the generalizability of scores obtained from such a test to the lexicon would be low. Therefore, vocabulary size tests need to sample unequally from different parts of speech and in proportion to their distribution in the lexicon. To do this, we need to know how different parts of speech are distributed in the target lexicon (e.g. English for EFL learners) in the first place. For instance, how many nouns, verbs, adjectives, adverbs, prepositions, etc. are there in English? The next section will review the literature on this question.

2.2. Distribution of Parts of Speech in the English Lexicon

Most parts of speech counts in English are restricted to function words probably because they are small in number and easy to count. A summary of these counts adapted from Öztürk (2020) is given in Table 1 below. To enable comparability, Öztürk (2020) converted each list to word families using Tom Cobb's Familizer software (https://www.lextutor.ca/familizer/). The number of word families in the converted lists ranged between 141 (Dang & Webb, 2016) and 336 (Öztürk, 2019). The higher figure in the latter count was obtained by scanning the 25 sub-lists of the BNC-COCA word frequency lists covering 25,000 word families (Nation & Davies, 2012). The number of function words in this count is only a fraction (.013%) of the entire list. If we included function words in a vocabulary size test of, say, 100 words, we would have to choose only one word from the list to maintain this proportion in the test. For a principled selection of the remaining 99 words, we would still need to know the distribution of content words in the language.

Table 1.

Function word counts in English**.**

Lists	Number of Items	N of function word families
Cook (1988)	225	146
Higgins & Higgins (1994)	321	171
O'Shea (2010)	277	201
Dang & Webb (2016)	176	141
Öztürk (2019)	336	336

(Adapted from Öztürk, 2020)

In spite of the latest advancements in corpus-based word lists, there is almost no formal counts of content words in the English lexicon. The only exception is Hudson (1994), who counted occurrences of various parts of speech in the Brown Corpus and the LOB Corpus. In his count, nouns had the highest frequency making about 37% of texts in each corpus. This figure was highly consistent between the two supergenres (informational vs imaginative) and across fifteen individual genres with only 9% difference between the most and the least nominal genres. Verbs were the second most frequent with 18%, prepositions 12%, adjectives 7%, adverbs 5%, and other categories 21%.

While these regularities are interesting from a linguistic stand point, they do not make a reliable guide in target word selection for a vocabulary size test, because what is counted are word tokens in texts rather than lexemes in a vocabulary list. The fact that 37% of texts are nouns does not automatically mean that nouns make 37% of words in the lexicon represented as a word list. This distinction between *corpus counts* and *list counts* will become clearer if we compare categories, such as prepositions, for which both types of counts are available. Hudson's corpus count has shown that prepositions cover 12% of texts in the corpora. However, there were only 107 prepositions in Öztürk's (2019) count of the BNC-COCA word list making only .004% of the list words. The higher percentages in texts result from repeated occurrences of words in texts whereas they appear only once in a word list. In a size test, we are primarily interested in the extent of the learner's knowledge of the target lexicon operationalized as a word list rather than how much of the words in target language texts they understand. Vocabulary size tests, therefore, need to be based on counts of parts of speech in word lists.

A study that seems promising in this regard is Brysbaert et. al. (2012). They augmented the SUBTLEX-US word frequency list with parts of speech frequencies to help word recognition researchers in psycholinguistics with target word selection. The list contains 74,286 words ordered alphabetically, and for each word information concerning the dominant part of speech out of fourteen categories, the frequency of the dominant part of speech, other parts of speech as well as their frequencies are provided in separate columns in an Excel worksheet. One of the disadvantages of the list is that parts of speech information is provided for individual words with no total counts for different parts of speech categories. For a vocabulary size test, this is not enough. We also need to know how the different parts of speech categories are distributed in the list so that we know how many nouns, verbs, etc. to choose from the list. If we want to design a test of 100 items, and if we know that half of the words in the target list is nouns, then 50 words in our test should be nouns. I reorganised the Excel worksheet they provide (http://expsy.ugent.be/subtlexus) to obtain this information. Here are the results:

Table 2.

Parts of speech distribution in Brysbaert et al. (2012).

Part of Speech	Number of Words	Percent of all words	Hudson (1994)
Noun	37,333	50.25%	37%
Verb	16,865	22.70%	18%
Adjective	10,325	13.89%	7%
Name	6,522	8.77%	-
Adverb	2,350	3.16%	5%
Other (incl. unclassified)	891	1.19%	33%
TOTAL	**74,286**	**100%**	**100%**

For the sake of comparison, Hudson's corpus counts are provided in a separate column. Clearly, content words have higher percentages in list counts, whereas function words (i.e. *Other* category) have higher percentages in corpus counts. For instance, half of the words in the word list are nouns, but only 37% of words in texts are nouns. On the other hand, function words are few in number, making up less than 1.9 % of the list words, although they cover around 33% of words in texts.

For several reasons, Brysbaert et al. (2012) list needs to be used with caution for the purpose of vocabulary size testing. For one thing, the list uses word forms as the unit of counting such that each inflectional and derivational form of a given base word is listed separately. For example, each of the forms *play, plays, played, playing, player* is a separate entry in the list. In a vocabulary size test, however, we hardly would want to see these as separate words. Vocabulary size tests generally use word families (e.g. Nation & Beglar's (2007) Vocabulary Size Test) as the unit of counting. Another limitation of the list is that it is based on a highly specialised corpus of subtitles from films and TV programmes in the US, which makes its generalizability to the whole English lexicon questionable.

EFL vocabulary size tests do not usually control for parts of speech in target word selection. The sole exception to this is Nation's Vocabulary Levels Test, which uses a fixed ratio for selecting items from the three content word categories of nouns, verbs and adjectives in all versions of the test (Nation, 1990; Schmitt, Schmitt and Clapham, 2001; Webb, Sasao and Ballance, 2017). While the ratio is 3 nouns, 2 verbs, and 1 adjective in the original 18-item test (i.e. 9 nouns, 6 verbs, and 3 adjectives in each frequency level), 30-item tests use a ratio of 5 nouns, 3 verbs, and 2 adjectives (i.e. 15 nouns, 9 verbs, and 6 adjectives in each frequency level). These correspond to 50% nouns, 33% verbs, and 17% adjectives vs 50% nouns, 30% verbs, and 20% adjectives, respectively. The same ratios were consistently applied in each frequency level. The origins of these ratios are, however, not clear and their validity is yet to be shown.

The present study aims to add parts of speech information to the first fourteen lists of the BNC-COCA word frequency lists as well as provide statistical information on the distribution of different parts of speech across individual frequency levels for use by the designers of EFL vocabulary size tests. Concerning the parts of speech distribution, we aim to answer the following questions:

1. What is the overall distribution of different parts of speech in the first 14 levels of the BNC-COCA word frequency lists?
2. Are there differences in the distribution of different parts of speech between individual levels?

3. Methodology

3.1. The BNC-COCA Word Lists

The first fourteen frequency-based lists from the 28 lists of the BNC-COCA word lists (Nation, 2017) were chosen for analysis. The limitation to the first fourteen levels was motivated by the common practice of testing the first fourteen levels in well-known English vocabulary size tests that use the BNC-COCA list (e.g. Nation & Beglar's VST (2007); Laufer & Goldstein's CATTS (2004)).

The fourteen lists are sequenced by frequency such that the first list (1K) contains words with the highest frequency with each subsequent list containing words of lower frequency than the previous list. The lists include only the base words (e.g. *happy*), excluding other family members (e.g. *happier, happiest, unhappy, happiness, unhappiness, happily*) although they were constructed on the basis of word families. The present study was also limited to basewords as it is usually the basewords that are tested in vocabulary size tests rather than one of the family members. Each frequency level contains around 1,000 words. The internal organisation of individual lists is alphabetical. The lists were downloaded from the Lextutor website (https://www.lextutor.ca/list_learn/bnc_coca/).

3.2. Coding

Each word in the list was coded with a single part of speech category. The following categories were used: noun, verb, adjective, and function word (function adverb, determiner, preposition, conjunction, modal, interjection, abbreviation). A distinction is made here between *content adverbs* and *function adverbs* corresponding to Quirk et al.'s (1985) *open class adverbs* vs *closed class adverbs* (p.73). *Content adverbs* are regularly derived from adjectives through the addition of the -ly suffix (e.g. *peacefully, reluctantly, regularly*). They were not included as a separate category in the present study because they do not appear in the base word lists. They are family members of the corresponding adjective base word and only rarely are base words themselves (e.g. *especially, probably*). *Function adverbs* are not based on adjectives, and like other function words have less semantic content (e.g. *so, too, enough, already, indeed*, etc.). These are counted under the function word category.

Parts of speech coding was done using an online parts of speech tagger (https://linguakit.com/en/part-of-speech tagging). Each one-thousand-word sublist was analysed separately. In Linguakit site, a given list is copied and pasted in the space provided and the analyser returns an automatically tagged version. It turned out, however, that the tagging was not always accurate. Linguakit is designed to work with texts, and without textual clues about a word's part of speech the assignment of parts of speech labels is carried out only on the basis of morphological clues, which can be misleading or ambiguous. For this reason, the lists were hand-checked and miscodings were corrected. In case of doubt about a word's part of speech, three sources were consulted: Adam Kilgariff's BNC word lists (https://www.kilgarriff.co.uk/BNClists/lemma.al), the Longman Dictionary of Contemporary English Online and Collins Online Dictionary. The two dictionaries were preferred for their emphasis on frequency. When a word has multiple parts of speech (e.g. *dream* as a verb and as a noun), the more frequent part of speech in Kilgariff's lists or the first cited part of speech, assumingly the most frequent, in the aforementioned dictionaries has been adopted as the word's part of speech (noun in the case of *dream*). Assignment of parts of speech on the basis of frequency can be justified on the grounds that learners are more likely to know the more frequent part of speech for a target word.

4. Results

4.1. Overall Distribution of PoS in the BNC-COCA lists

The overall results of the coding across the fourteen frequency levels are presented in Table 3 below. Of the 14,007 words examined, 13,978 words were unambiguously assigned into a parts of speech category. Only 29 words (e.g. *fÃœhrer* (10K), *unced* (12K), *eau* (13K), *auf* (14K)) could not be coded as they were not included in the dictionaries consulted, and most of these were in the lower frequency levels (i.e. 10K-14K). As a whole, there was a greater number of nouns than words of any other part of speech. In fact, more than half of the words in the list (58%) were nouns. Function words consisted of a tiny proportion of the words with merely 3%. There were significantly more verbs (21%) than adjectives (18%) in the entire list (X^2 (1, 5522)= 33.17, p<.0001).

Table 3.

Overall distribution of PoS categories in the BNC-COCA lists.

Part of Speech	N	%
Noun	8,030	58%
Verb	2,982	21%
Adjective	2,543	18%
Other	426	3%
Total	**13,978**	**100%**

4.2. Within-Level Distribution

The results of the coding for individual frequency levels are given in Table 4 below. These suggest that overall tendencies observed above also hold in individual frequency levels, with the exception of the 1K level. This level has an unusually high percentage of function words (19%). This is even higher than adjectives (16%), although a chi-square test of independence did not reveal this difference to be significant (X^2 (1, 352) = 1.63, p=.20). On the other hand, the number of verbs was significantly higher than function words (X^2 (1, 422) = 5.01, p=.02). Unlike other levels, the percentage of nouns in the 1K level was less than half the words in this level (42%) due to the high percentage of function words. When computed without function words, the percentage of nouns rises to 51%, as in the other levels.

Table 4.

Parts of Speech distribution of words in individual frequency levels (1K-14K) of the BNC-COCA lists.

Frequency Level	Noun	Verb	Adjective	Function Word	Total
1K	420 **42 %**	234 **23 %**	164 **16 %**	188 **19 %**	1,006 **100 %**
2K	571 **57 %**	271 **27 %**	138 **14 %**	20 **2 %**	1,000 **100 %**
3K	523 **52 %**	288 **29 %**	174 **17 %**	15 **2 %**	1,000 **100 %**

4K	568 **57 %**	266 **27 %**	149 **15 %**	17 **2 %**	1,000 **100 %**
5K	596 **60 %**	237 **24 %**	158 **16 %**	10 **1 %**	1,001 **100 %**
6K	543 **54 %**	285 **29 %**	161 **16 %**	11 **1 %**	1,000 **100 %**
7K	547 **55 %**	230 **23 %**	208 **21 %**	17 **2 %**	1,000 **100 %**
8K	587 **59 %**	218 **22 %**	179 **18 %**	16 **2 %**	1,000 **100 %**
9K	561 **56 %**	221 **22 %**	199 **20 %**	19 **2 %**	1,000 **100 %**
10K	594 **60 %**	178 **18 %**	215 **22 %**	11 **1 %**	998* **100 %**
11K	599 **60 %**	180 **18 %**	198 **20 %**	23 **2 %**	1,000 **100 %**
12K	613 **61 %**	163 **16 %**	193 **19 %**	26 **3 %**	995* **99 %**
13K	662 **66 %**	107 **11 %**	199 **20 %**	23 **2 %**	991* **99 %**
14K	646 **65 %**	104 **10 %**	208 **21 %**	29 **3 %**	987* **99 %**

*Some words were unclassifiable.

In the remaining frequency levels, nouns had the highest percentage (between 52% and 66%) and function words the lowest (1%-3%). The number of verbs was higher than adjectives in all levels up to the 10K level. From 10K level on, the opposite was the case. These differences between verbs and adjectives were statistically significant in the first six levels (1K-6K) and in the last two levels (13K-14K). In between these two extremes (7K-12K levels), there was no statistically significant difference between the number of verbs and adjectives (cf. Table 5 for the results of the chi-square tests of independence).

Table 5.

Statistical differences between the number of verbs and adjectives in 1K-14K levels.

Level	**Degrees of Freedom**	**X^2**	**p**
1K	1, 398	12.31	p<.0005*
2K	1, 409	43.24	p<.0001*
3K	1, 462	28.13	p<.0001*
4K	1, 415	32.98	p<.0001*
5K	1, 395	15.80	p<.0001*
6K	1, 446	34.47	p<.0001*

7K	1, 438	1.10	p=.29
8K	1, 397	3.83	p=.05
9K	1, 420	1.15	p=.28
10K	1, 393	3.48	p=.06
11K	1, 378	0.86	p=.35
12K	1, 356	2.53	p=.11
13K	1, 306	27.66	p<.0001*
14K	1, 312	34.67	p<.0001*

4.3. Cross-Level Distribution

Figure 1 displays the distribution of each part of speech category across the fourteen frequency levels. Nouns seem to gradually increase in number from 42% in the 1K level to 65% in the 14K level. The differences in the number of nouns across the fourteen frequency levels were statistically significant (1K-14K: (X^2 (13, 8030) = 77.31, p<.0001). Although there was a sharp increase from the 1K to the 2K level (from 42% to 57%), this does not seem to be responsible for the significant differences. When the 1K level was removed from the analysis, the difference between the remaining thirteen levels was still significant (2K-14K: (X^2 (12, 7610) = 32.37, p=.36).

The number of verbs significantly decreases with decreasing frequency (X^2 (13, 2982) = 217.40, p<.0001), whereas adjectives significantly increase in number as the frequency level decreases (X^2 (13, 2543) = 43.12, p<.0001). This interaction between verbs and adjectives is seen as the crossing of the lines for verbs and adjectives in the 10K level in Figure 1.

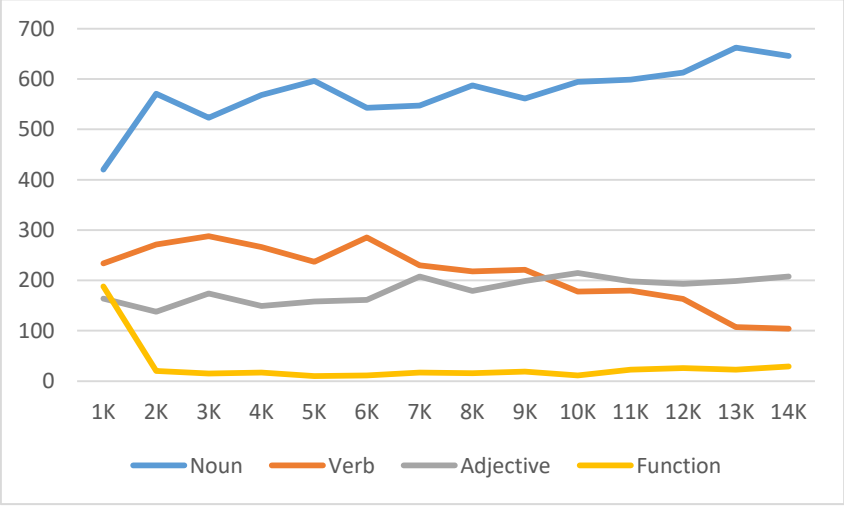

Figure 1. Parts of speech distribution across frequency levels

5. Discussion

The present study investigated the distribution of the parts of speech categories in the first 14 one-thousand-word levels of the BNC-COCA word frequency list. Given the comprehensiveness of the list as well as the corpora on which it is based, the distribution found in the present study could be argued to be fairly representative of the English lexicon as a whole.

Not surprisingly, the present data revealed an uneven distribution of different parts of speech in the list. Nouns formed the greatest category amounting to more than half of the words in the list, and function words the smallest. There were more verbs than adjectives although the sizes of these categories were very close (21% vs 18%). The overall results from the present study are compared to the results obtained in previous studies in Table 6 below. For comparability, Brysbaert et al.'s data were recomputed after the category for *Names*, which was missing in the present data as well as in Hudson's (1994), was removed, and everything else was included under the category '*Other'*.

Table 6.

Comparison of results to previous studies.

Part of Speech	**Present Study**	**Brysbaert et al. (2012)**	**Hudson (1994)**
Noun	58%	55%	37%
Verb	21%	25%	18%
Adjective	18%	15%	7%
Other	3%	5%	38%

The results of the present study are remarkably similar to Brysbaert et al.'s (2012) figures. This is not surprising as both are list counts. In spite of the differences in corpus content and unit of counting, the general tendencies seem to be similar. Hudson 's text count (1994), on the other hand, has lower percentages for content word categories and a much higher percentage of function words (38%). This is also not surprising as there are a small number of function words in English, but they are used extremely frequently, hence their higher percentage in texts. This further illustrates that corpus counts of parts of speech are not reliable guides in target word selection for size tests.

The distribution of parts of speech in the various versions of the Vocabulary Levels Test is also rather similar to the present study, except that verbs are overrepresented in the Levels Tests, and function words are non-existent. This comparison is provided in Table 7 below.

Table 7.

Representativeness of the Levels Tests.

Part of Speech	**Levels Tests**	**Present Study**
Noun	50%-50%	58%
Verb	33%-30%	21%
Adjective	17%-20%	18%
Other	-	3%

Another finding of the present study is that the relative proportions of PoS categories in individual levels do not remain the same across different frequency levels. Verbs and adjectives are particularly problematic as their proportions are effectively reversed as the frequency level decreases. This recommends against using the same ratios across individual frequency levels in size tests. Table 8

proposes item numbers for major parts of speech categories in individual frequency levels in a vocabulary size test of 140 items in accordance with the variable ratios obtained in the present data.

Table 8.

Proposed item numbers for a 140-item size test.

Level	Noun	Verb	Adjective	Function
1K	4	3	1	2
2K	6	3	1	0
3K	5	3	2	0
4K	6	3	1	0
5K	6	3	1	0
6K	5	3	2	0
7K	6	2	2	0
8K	6	2	2	0
9K	6	2	2	0
10K	6	2	2	0
11K	6	2	2	0
12K	6	2	2	0
13K	7	1	2	0
14K	7	1	2	0
TOTAL	82	32	24	2

6. Conclusion

The present study made a count of the parts of speech categories in the first fourteen baseword lists of the BNC-COCA word frequency list. Additionally, the lists were augmented with parts of speech categories, and made available in a Word Excel worksheet as supplementary material to the present paper. Used together with the lists, the present count is expected to inform future vocabulary tests enabling more systematic and valid selection of the target vocabulary and more accurate estimates of vocabulary size.

References

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press.

Barclay, S., & Pellicer-Sánchez, A. (2021). Exploring the learning burden and decay of foreign language vocabulary knowledge: The effect of part of speech and word length. *ITL-International Journal of Applied Linguistics*, *172*(2), 259-289.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, 44(4), 991-997.

Cook, V. (Accessed February 23, 2018). List of English structure (function) words. Available at: http://www.viviancook.uk/Words/StructureWordsList.htm)

Coxhead, A., Nation, P., & Sim, D. (2014). Creating and trialling six versions of the Vocabulary Size Test. *The TESOLANZ Journal*, 22, 13–27.

Dang, T. N. Y., & Webb, S. (2016). Making an essential word list for beginners. In I. S. P. Nation, *Making and Using Word Lists for Language Learning and Testing* (pp. 153-167, 188-195). Amsterdam: John Benjamins.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning*, 43(4), 559-617.

Glanzer, M. (1962). Grammatical category: A rote learning and word association analysis. *Journal of Verbal learning and Verbal Behavior*, 1(1), 31-41.

Hashimoto, B. J. (2021). Is frequency enough?: The frequency model in vocabulary size testing. *Language Assessment Quarterly*, *18*(2), 171-187.

Hashimoto, B. J., & Egbert, J. (2019). More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning*, *69*(4), 839-872.

Higgins, J. & Higgins, M. (1994). Function words. Accessed on 14 October, 2019 at: http://www.marlodge.net/Function-words/

Horst, M., & Meara, P. (1999). Test of a model for predicting second language lexical growth through reading. *Canadian Modern Language Review*, 56(2), 308-328.

Hudson, R. (1994). About 37% of word-tokens are nouns. *Language*, 70(2), 331-339.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language learning*, *54*(3), 399-436.

Ludington, J. D. (2015). Effects of word class and training method on vocabulary learning in a second language. *Research in Language*, 13(4), 426-449.

McLean, S., & Kramer, B. (2015). The creation of a new Vocabulary Levels Test. *Shiken*, 19(2), 1–11.

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, 19(6), 741–760. doi: 10.1177/1362168814567889

Morgan, C. L., & Bonham, D. N. (1944). Difficulty of vocabulary learning as affected by parts of speech. *Journal of Educational Psychology*, 35(6), 369.

Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. Boston: Heinle & Heinle.

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

Nation, I.S.P. & Davies, M. (2012). The BNC/COCA word family lists. Available from: https://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Information-on-the-BNC_COCA-word-family-lists.pdf

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, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31(7), 9–13.

O'Shea, J. (Accessed on 15 October, 2019). Function word lists. Available at: https://semanticsimilarity.wordpress.com/function-word-lists/

Ozturk, M. (2019). Function words in word frequency lists. *Research Gate*. DOI:10.13140/RG.2.2.15097.62565

Öztürk, M. (2020). Are function words in English adequate? *Research Gate*, DOI: 10.13140/RG.2.2.15355.87840

Puimège, E., & Peters, E. (2019). Learning L2 vocabulary from audiovisual input: An exploratory study into incidental learning of single words and formulaic sequences. *The Language Learning Journal*, 47(4), 424-438.

Quirk R., Greenbaum. S., Leech G., and Svartvik J. 1985. *A Comprehensive Grammar of English Language*. London Pearson Longman.

Rodgers, T. S. (1967). Measuring vocabulary difficulty: An analysis of item variables in learning Russian-English and Japanese-English vocabulary pairs. *Technical Report No.124*, Standford University.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18(1), 55-88.

Thorndike, E.L. and Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University.

Tschichold, C. (2013). Using CASLR to estimate individual word difficulty. Paper presented at *WorldCALL Conference on Global Perspectives on Computer-Assisted Language Learning*, Glasgow. pp.345-347.

Van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, 41(3), 609-624.

Webb, S., Sasao, Y., & Balance, O. (2017). The updated vocabulary levels test: Developing and validating two new forms of the VLT. *ITL–International Journal of Applied Linguistics*, 168 (1), 33–69.

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

Appendix

Data: Parts of Speech in bnc-coca lists

Freely available from OSF at the following URL:

https://osf.io/zbnv8/?view_only=897cf72789444576a1b6ccd2a51d8af6