Vocabulary Learning and Instruction
Volume 7, Issue 1, December 2018
http://vli-journal.org

Examining the Word Family through Word Lists

Dale Brown
Institute of Liberal Arts and Science, Kanazawa University, Kanazawa, Japan

doi: https://doi.org/10.7820/vli.v07.1.brown

Abstract

The choice of lexical unit has important consequences for L2 vocabulary research, testing and instruction. In recent years, the most widely used lexical unit has been the word family. This study examines the characteristics of word lists based on the word family and explores the levels of text coverage such lists may provide should the assumption that learners can deal with word families be incorrect. This is pursued through the detailed examination of a set of word-family-based word lists. The study finds that such word lists pose a number of challenges, including the number of word forms with multiple affixes, the number of word forms with more challenging affixes, and the number of word families in which the base word is not the most frequently occurring member. Moreover, the first thousand word families in particular are shown to be challenging. The study then demonstrates that if learners are unable to deal with the complexity of word families, even to a relatively small degree, word-family-based lists may provide far lower text coverage levels than may be assumed. It concludes that in work on second language vocabulary, careful consideration is needed of the appropriacy of the word family as the lexical unit and highlights the range of work based on the word family that may need reevaluating.

Keywords: lexical unit, word families, word lists, text coverage

1 Introduction

The choice of lexical unit has important consequences for work on second language vocabulary (Gardner, 2007), with implications for research, language testing, curriculum design, and teaching. In recent years, the word family has been the most widely used lexical unit. The word family, described by Bauer and Nation (1993), was intended as a flexible concept, with seven levels (see Table 1).

Despite the intention of flexibility, in practice Level 6 word families (i.e., including word forms featuring any of the Levels 1–6 affixes; referred to hereafter as WF6) have been used most often, and indeed, reference to "word families" in the literature largely refers to this specific level. This is primarily due to the free availability of word lists based on WF6 developed by Nation, but is also a result of the extensive work of Nation and colleagues featuring these word lists (e.g., Beglar, 2010; Laufer & Ravenhorst-Kalovski, 2010; Webb & Macalister, 2013), and the freely available Range (Nation & Heatley, 2002) software and its online equivalent at LexTutor (Cobb, no date) which make use of these lists.

Please cite this article as: Dale Brown (2018). Examining the Word Family through Word Lists. *Vocabulary Learning and Instruction, 7* (1), 51–65. doi: 10.7820/vli.v07.1.brown

Table 1. Summary of Bauer and Nation's (1993) Word Families Scheme

Level	Description	Number of affixes	Examples of affixes	Examples of forms
1	Each form is a different word.			
2	Inflectional suffixes.	8	*-ed*, *-s*	*heat-ed* *waste-s*
3	The most frequent and regular derivational affixes.	10	*-able*, *-er*, *non-* (each with restricted uses)	*heat-er*
4	Frequent, orthographically regular affixes.	11	*-al*, *-ful*, *in-* (each with restricted uses)	*waste-ful*
5	Regular but infrequent affixes.	50	*-age*, *-ally*, *ante-*	*wast-age*
6	Frequent but irregular affixes.	12	*-ee*, *-ic*, *pre-*	*pre-heat*
7	Classical roots and affixes.			

Alternatives to the word family include word types, lemmas, and flemmas. A word type is an individual word form. Thus, *act* and *acted* are different word types. A lemma is a base word and its inflections (i.e., paradigmatically related forms of the same word class). Thus, *act*$^{\text{verb}}$ and *acted*$^{\text{verb}}$ are part of a single lemma, while *act*$^{\text{noun}}$ belongs to a different lemma. A flemma (Pinchbeck, 2014, March) is a base word and inflected forms regardless of word class. Thus, *act*$^{\text{verb}}$, *act*$^{\text{noun}}$, and *acted*^{verb} are part of a single flemma, while *actor* belongs to a different flemma.

Nation (2006b; 2015) argues that different lexical units are appropriate for different purposes, but suggests word families are a good choice when considering receptive uses of language, except with learners who are very beginners. Nation cites two justifications: first, there is evidence that word families are psychologically real (Bertram, Laine & Virkalla, 2000; Nagy, Anderson, Schommer, Scott & Stallman, 1989), and second, once learners have some familiarity with a word family, they are able to deal with its various members with little difficulty when encountered in context.

There has, however, been some questioning of the word family. It has been pointed out that the research cited by Nation in justifying the word family was with L1 participants (McLean, 2017), while the compilers of two recently developed word lists (Brezina & Gablasova, 2015; Gardner & Davies, 2014) expressed concerns about the semantic distance that can exist between individual forms in a word family and about whether learners have the morphological skills necessary to deal with derivational word relationships, leading both to choose lemmas as their lexical unit.

More pointedly, several studies have shown that L2 learners do not necessarily find dealing with the types of word forms that word families contain a simple task. Mochizuki and Aizawa (2000) presented Japanese L2 learners with pseudowords featuring affixes in Levels 3–6 of Bauer and Nation's scheme and, in a multiple-choice format, asked learners to choose the meaning of the affix in the case of prefixes and the part of speech of the word in the case of suffixes. Knowledge of affixes correlated with vocabulary size, but remained partial even among learners with a vocabulary size estimated at over 5000 words. Furthermore, knowledge of particular affixes did not correspond with their position in Bauer and Nation's scheme (e.g., the best known prefix in the study was *re*-, yet this appears at Level 6 in the scheme). In a replication of Mochizuki and Aizawa's study with upper-intermediate Serbian learners, Danilović, Savić, and Dimitrijević (2013) likewise found partial knowledge of English affixes and an apparent order of acquisition at odds with Bauer and Nation's scheme. Ward and Chuenjundaeng (2009) had Thai learners of English give translations of base forms and of derived forms and found that in most cases where a base form was translated successfully, the derived form was not, and vice versa. They concluded that Thai learners do not in general make use of English word-building devices. Brown (2013) asked Japanese L2 learners to mark unknown words while reading and then investigated the characteristics of the marked words. This revealed that, in high-frequency word families, inflectional and derivational forms were relatively more likely to be marked than base words, suggesting such forms pose some additional difficulty. Reynolds (2015) conducted a study of incidental vocabulary learning with advanced learners in Taiwan and found better acquisition for words that occurred in the text in a single invariant form as compared with words whose occurrences displayed inflectional or derivational variation. Finally, McLean (2017), in a study with Japanese learners at various proficiency levels, tested the receptive understanding of a number of highly frequent words and multiple members of their word families. There was good comprehension of inflectional forms, but limited comprehension of derivational forms even among learners of advanced proficiency. There is, then, evidence, in the case of both learners with L1s that make use of derivation and those that do not, that learners up to advanced levels of proficiency find the degree of knowledge of derivations and word-building processes necessary to deal with the word family a considerable challenge.

The adoption of the word family as the lexical unit is not, however, without attractions: it enables higher levels of text coverage to be achieved with a smaller word list. Text coverage matters because it affects comprehension (Hu & Nation, 2000; Schmitt, Jiang & Grabe, 2011). One aim in developing word lists is therefore a desire to discover how higher coverage levels can be achieved most efficiently, that is, to identify what learners should learn in order to reach greater coverage levels as quickly and easily as possible. With an expanded concept of word, such as WF6, greater coverage can be achieved, and so word-family-based lists can make the vocabulary learning challenge appear relatively achievable and manageable.

The crucial issue is, however, whether learners can indeed deal with the various members of a word family when encountering them in context. If they cannot, the coverage levels that it is claimed can be reached are illusory. What is more, the assumption need only be somewhat incorrect for problems to arise. That is, learners who can deal with much of the challenge posed by word families, but not all, may nonetheless face problems. If, for example, learners can deal with 90% of what lies behind a word-family-based word list (i.e., learners have 90% of the knowledge needed to cope with word families), the true coverage level provided by that list for a given text may not be, say, 98%, but rather 88% (98% × 90%). Such a drop in coverage may seem insignificant, but small differences in coverage can substantially affect comprehension (Hu & Nation, 2000; Schmitt, Jiang & Grabe, 2011). This is easier to recognize by switching the perspective and considering that with 98% coverage 2% of words are unknown, while with 88% coverage 12% are unknown, six times more.

What is absent from the debate about the word family thus far is a thorough interrogation of word-family-based word lists that would plainly reveal their characteristics. The affixes permitted are described by Bauer and Nation (1993), and actual lists themselves can be freely downloaded and examined (see, e.g., http://www.victoria.ac.nz/lals/about/staff/paul-nation). Yet, it is not easy to gain an understanding of the nature of the lists: they are simply too large, containing thousands of forms, for any casual perusal to be informative. It has been pointed out, for example, that the base word of a family can be less frequent than other members of the family (Coniam, 1999). However, it is not clear if this is an isolated instance or a more generalized problem. A more systematic examination of word families would provide teachers and researchers with a better sense of the challenge that learners face in dealing with word families and of the consequences should the assumption that learners can deal with this challenge be mistaken.

Thus, this study does not replicate those cited above and explore how learners deal with word families. Instead, the aim is to conduct a thorough examination of a set of word-family-based word lists in order to provide a detailed characterization of such lists. Specifically, two questions are asked:

1. To what extent do WF6-based word lists have characteristics that may be challenging for learners?
2. If the assumption that learners can deal with WF6 is incorrect, what levels of text coverage might such lists provide?

The first question involves investigating the size of the word families in the lists and their complexity; the second question means looking at the frequency of individual forms within the word families that include affixes at different levels in the word families scheme.

2 Method

This study analyzes the higher frequency portion of Nation's (2006a) British National Corpus-based word lists. These lists consist of fourteen bands, each of 1000 word families. This study examines the first five bands, so as to concentrate on the bands that contain the vast majority of words in any text and which are the focus of vocabulary learning for the majority of L2 learners (Webb & Sasao, 2013).

Nation's BNC-based lists contain WF6 word families as established by Bauer and Nation. However, the lists also include some forms featuring affixes outside of the scheme (e.g., several forms including the affix *dis*- as in *dislike*; see the *Other affixes observed* column in the additional material online), along with irregular verb and noun forms (e.g., *became* within the BECOME family), abbreviated forms of base words (e.g., *ad* and *advert* within the ADVERTISE family), alternative spellings of base words (e.g., *center* within the CENTRE family), and compound forms (e.g., *backbone* within the BACK family).

The BNC-based lists were selected for study since they are based on a single source which can be easily accessed. This means that the frequency of the forms in the lists can be checked in the very source of the lists itself. The lists were downloaded from Nation’s website (http://www.victoria.ac.nz/lals/about/staff/paul-nation), and all searches of the BNC were conducted via the BYU-BNC (Davies, 2004) website (http://corpus.byu.edu/bnc). For the majority of the analyses, all 1000 word families in each of the five bands were examined. However, for the corpus-based investigation of the lists, a systematic random sample of 100 word families was taken from each band.

Frequency information was collected for each of the 2396 word forms in the five samples of 100 word families. In each case, the search was for the word form itself (i.e., no part of speech was specified).

In order to address research question 1, on the characteristics of WF6-based word lists, the analysis looks at:

- the number of word forms in the families across the five bands
- the number of word forms in each band that contain different numbers of affixes (i.e., the number of forms including a single affix, the number of forms including two affixes, and so on)
- the number of word forms in each band that include affixes at the various levels of Bauer and Nation’s scheme (i.e., the number of forms including Level 2 affixes only, the number of forms including affixes through to Level 3, and so on)
- the number of word families for which the base word is not the most frequent member of the family.

The analysis then addresses research question 2, on text coverage levels. This was a two-step process:

1. Based on the frequency in the BNC of the individual forms that comprise word families, a calculation was made of the mean proportion of a word family’s total occurrences that is provided by forms at different levels of the scheme (i.e., the proportion of a word family’s total occurrences accounted for by forms including Level 2 affixes only, by forms through to Level 3, and so on).
2. The above proportions were then used to make estimates of the varying degrees of text coverage that may be provided if learners are able to deal with different levels of the scheme (i.e., the text coverage that may be provided if learners can deal with Level 2 affixes only, with Level 3 affixes also, and so on).

It should be noted that despite the fact that, as reported earlier, the validity of the levels in the word families scheme has been questioned, in the absence of any comprehensive data on affix difficulty, the scheme’s levels were made use of in this analysis.

3 Results

3.1 Size and Complexity of the Lists

Table 2 gives the mean number of word forms that are included in each band. The number of word forms per family differs significantly across the

Table 2. Descriptive Statistics for the Size of Word Families

	Mean number of word forms	SD	Minimum	Maximum
1K	6.35	4.343	1	35
2K	5.59	3.536	1	22
3K	4.52	2.761	1	18
4K	4.29	2.700	1	19
5K	3.99	2.600	1	23

bands ($H(4) = 277.48$, $p < 0.001$, $\eta 2 = 0.05$), with the 1K families containing the most members and the number decreasing across subsequent bands. It can also be seen that some word families are very large, the largest, ORGANIZE, having 35 word forms, while some consist of just a single member (e.g., ABOUT in the 1K band, ABOVE in the 2K band, ABROAD in the 3K band, ALIKE in the 4K band, and ABOARD in the 5K band). The 1K band, despite having the highest mean number of members, contains 104 families consisting of a single member, approximately twice as many as the other bands (2K = 49; 3K = 50; 4K = 52; 5K = 51). This is due to the large number of function words in the 1K band. Accordingly, the multi-member word families in the 1K band contain an average of almost seven members per family, while those in the 5K band have just over four members.

Table 3 shows the number of word forms in each band containing different numbers of affixes. That is, there are some forms that feature one affix (e.g., *educat-ion*, in EDUCATE, a 1K family), some with two affixes (e.g., *educat-ion-al*), some with three affixes (e.g., *educat-ion-al-ist*), and some with four affixes (e.g., *educat-ion-al-ist-s*). A Pearson's chi-square shows a significant difference between the bands in this regard (χ^2 (16) = 602.09, $p < 0.001$, Cramer's $V = 0.09$, a small to medium effect), and examining the standardized residuals (which reveal the difference between the observed value in each cell of the table and the value that can be predicted on the basis of the overall figures) allows the location of these differences to be pinpointed. This first reveals that the considerable number of forms with zero affixes (irregular verb and noun forms, abbreviated forms, alternative spellings, and compound forms, as mentioned above) do not occur evenly across the bands. There are significantly more in the 1K band, primarily due to the presence of irregular verb and noun forms. Second, while in all five bands the majority of forms contain a single affix, there are differences across the bands. Specifically, there are significantly fewer forms with a single affix in the 1K band and significantly more such forms in bands 3K–5K. Corresponding with this, there are significantly more forms with two and three affixes in the 1K and 2K bands and significantly fewer such forms in bands 3K–5K. Finally, it may be pointed out that the number of forms with two or more affixes (5307) is not trivial, accounting for 26.9% of the forms overall, and that there are 1877 forms containing two or more derivational affixes.

Table 4 presents the number of forms within each band across different levels of the Bauer and Nation scheme. For example, the 1K band has 3040 word forms that feature only Level 2 affixes (e.g., *admitt-ed* and *clean-ing*), a further 1097 forms with Level 3 affixes only (e.g., *un-clean*) or Levels 2 and 3 affixes

Table 3. Number of Forms with Various Numbers of Affixes

	Forms with zero affixes (excluding base words)	Forms with one affix	Forms with two affixes	Forms with three affixes	Forms with four affixes
1K	433 (8.1)	3223 (60.3)	1495 (28.0)	188 (3.5)	9 (0.2)
2K	122 (2.7)	3068 (66.8)	1257 (27.4)	135 (2.9)	11 (0.2)
3K	99 (2.8)	2613 (74.3)	751 (21.4)	53 (1.5)	1 (0.0)
4K	72 (2.2)	2451 (74.5)	717 (21.9)	47 (1.4)	1 (0.0)
5K	65 (2.2)	2284 (76.3)	603 (20.2)	39 (1.3)	0 (0.0)
1–5K	791 (4.0)	13639 (69.1)	4823 (24.4)	462 (2.3)	22 (0.1)

Note: Brackets show the proportion of forms in each band (excluding base words) with each number of affixes.

Table 4. Number of Forms at Different Levels of Bauer and Nation's Scheme

	Forms with Level 2 affixes only	Additional forms through to Level 3 affixes	Additional forms through to Level 4 affixes	Additional forms through to Level 5 affixes	Additional forms through to Level 6 affixes[a]
1K	3040 (56.8)	1097 (20.5)	373 (7.0)	218 (4.1)	187 (3.5)
2K	2921 (63.6)	808 (17.6)	388 (8.4)	163 (3.5)	191 (4.2)
3K	2431 (69.1)	552 (15.7)	236 (6.7)	95 (2.7)	104 (3.0)
4K	2315 (70.4)	444 (13.5)	217 (6.6)	112 (3.4)	128 (3.9)
5K	2104 (70.3)	396 (13.2)	226 (7.6)	95 (3.2)	105 (3.5)
1–5K	12811 (64.9)	3297 (16.7)	1440 (7.3)	683 (3.5)	715 (3.6)

Note: Brackets show the proportion of forms at each band (excluding base words).
[a]Forms in the lists featuring affixes outside Bauer and Nation's scheme (see *Other affixes observed* column in the additional material online) are included here with Level 6 affixes.

(e.g., *admitt-ed-ly*), and so on. It should be noted that this analysis is based purely on the level of the affixes in the scheme, not on the number of affixes. Thus, for example, among the 3040 forms in the 1K band with Level 2 affixes only, there are 72 forms that feature two Level 2 affixes, such as *find-ing-s* and *low-er-ed*.

A Pearson's chi-square on the Table 4 figures finds a significant difference across the bands in the number of forms with different levels of affixes: χ^2 (16) = 210.08, $p < 0.001$, Cramer's $V = 0.05$, a small effect. The standardized residuals reveal that, in general, a similar proportion of the forms in each band are accounted for by the affixes through to Level 4, Level 5, and Level 6. Where the bands differ is in the proportion of forms with Level 2 affixes only and the proportion with affixes through to Level 3. The 1K band contains significantly fewer forms with Level 2 affixes only in comparison with the overall trend across the five bands, while the 3K–5K bands contain significantly more. As for forms with affixes through to Level 3, the 1K band contains significantly more such forms and the 4K and 5K bands significantly fewer.

A final aspect of the complexity of the word families concerns the number of families for which the base word is the most frequent member of the family and the number for which it is not. For example, in the BNC the word form *boy* is the most frequently occurring of the four forms in the BOY word family. There are

Table 5. Estimates of the Number of Families for Which the Base Word is the Most Frequent Member

	Base word is the most frequent member	Base word is not the most frequent member
1K	820	180
2K	740	260
3K	810	190
4K	760	240
5K	750	250
1–5K	3880	1120

Note: Estimates are based on samples of 100 word families from each band.

20 807 occurrences of the four forms: 12 714 (61.1%) for *boy*, 7790 (37.4%) for *boys*, 159 (0.8%) for *boyish*, and 144 (0.7%) for *boyhood*. In the ACTIVE word family, in contrast, the most frequently occurring of its nine forms is *activities*, with 11 476 (34.3%) out of 33 469 total occurrences for the family, while the form *active* itself has 7219 (21.6%) occurrences. As Table 5 shows, on the basis of samples of 100 word families from each band, for the majority of word families the base word is the most frequently occurring member. Nonetheless, in over 20% of the families, and fairly consistently across all five bands, another form occurs more frequently.

3.2 Coverage Levels

On the coverage levels that may be provided by the lists, Table 6 gives the mean proportion of the total occurrences in the BNC of all the forms in a word family that are accounted for by occurrences of the base word alone. For example, as explained above, in the BNC the base words of the BOY and ACTIVE word families account for 61.1% and 21.6%, respectively, of the total occurrences of all the forms in their families. As the table shows, on average the base words account for around three fifths of the occurrences of each family, with similar figures across all five bands. The standard deviations and minimum and maximum figures also show, however, that there is a great deal of variation among the families. There are families for which the base word accounts for 100% of occurrences, this obviously being the case for families consisting of a single member, and families for which the base word accounts for a very small proportion of the occurrences, the lowest being 0.8% for the GOVERN family.

Table 7 shows the proportion of occurrences accounted for by forms up to and including each level of Bauer and Nation's scheme. Note that other forms, that is, irregular verb and noun forms, abbreviated forms, and so on, have been included prior to any of the levels of affixation, though it is unclear if this is appropriate or not. As can be seen, the addition of the Level 2 forms results in a large increase in the proportion of occurrences accounted for, with smaller increases with the addition of each subsequent level.

Table 8 illustrates how the above figures can be used to revise coverage estimates. For example, if the word lists provide 95% coverage of a given text, but

Table 6. Estimates of the Proportion of the Occurrences of Families Accounted for by the Base Word only

	Mean	SD	Minimum	Maximum
1K	67.0	27.52	0.8	100
2K	56.7	26.22	0.9	100
3K	63.0	27.48	1.1	100
4K	62.9	28.53	2.1	100
5K	59.1	27.83	2.2	100
1–5K	62.0	27.46	0.8	100

Note: Estimates are based on samples of 100 word families from each band.

Table 7. Estimates of the Mean Proportions of the Occurrences of Families Accounted for by Forms at Different Word Family Levels

	Base word only	Plus other forms[a]	Plus Level 2 forms	Plus Level 3 forms	Plus Level 4 forms	Plus Level 5 forms	Plus Level 6 forms
1K	67.0	71.2	87.9	91.7	95.6	97.2	100.0
2K	56.7	57.5	82.5	89.4	92.7	95.1	100.0
3K	63.0	64.0	87.8	93.1	95.6	96.5	100.0
4K	62.9	63.1	86.6	91.0	93.1	94.7	100.0
5K	59.1	60.5	88.2	90.4	94.3	96.7	100.0
1–5K	62.0	63.2	86.6	91.1	94.3	96.0	100.0

Note: Estimates are based on samples of 100 word families from each band.
[a] Irregular verb and noun forms, abbreviated forms, alternative spellings, and compound forms.

Table 8. Estimates of How Assumed Coverage Levels are Affected by Different Levels of Affix Knowledge

Assumed coverage	If base words only are known	Plus other forms	Plus Level 2 forms	Plus Level 3 forms	Plus Level 4 forms	Plus Level 5 forms	Plus Level 6 forms
95	58.9	60.1	82.3	86.6	89.5	91.2	95.0
98	60.8	62.0	84.9	89.3	92.4	94.1	98.0

learners are only able to deal with Level 2 of the scheme, the actual coverage is estimated to be 82.3%. This is calculated by multiplying the assumed coverage figure (i.e., 95%) by the mean proportion of the occurrences accounted for by including up to Level 2 forms from Table 7 (i.e., 86.6).

As can be seen, if learners are unable to deal with the challenges of WF6, there is a considerable impact on the actual levels of coverage that may pertain. Even if learners are able to cope with affixes at Level 5, but not at Level 6, the impact is substantial: a text assumed to have 95% coverage (one in twenty words unknown) may in fact have only 91% coverage (almost one in ten words unknown), and a text assumed to have 98% coverage (one in fifty words unknown) may have only 94% coverage (three in fifty words unknown).

4 Discussion

The first question this study sought to answer was: To what extent do WF6-based word lists have characteristics that may be challenging for learners? The results above provide a number of insights into these characteristics.

It was found that the 5000 word families in the BNC-based lists contain almost 25 000 word forms, with the 1K band alone containing over 6000 forms despite 10% of its families being single-member families. Among these word forms, there are a number with zero affixes, consisting of irregular verb and noun forms, abbreviated forms, alternative spellings, and compound forms, which it is presumed are included on the assumption that they cause no problems for learners. There appears to be no research on whether this is actually the case or not. However, it must be said that while these forms may appear transparent, learners may have a different perspective.

Looking at the number of affixes in each word form, it was found that the majority contain just a single affix. Nevertheless, over 5000 forms, 26.9% of the total, contain multiple affixes, and there are almost 2000 forms with multiple derivational affixes. The extent to which learners are able to deal with such forms is an unexplored question. However, it seems likely that the more affixes a form contains the greater the difficulty for learners, since each additional affix makes the link between the base form and the derived form more obscure, both in terms of orthography/phonology and semantics. Some evidence of this was found by S. McLean in his 2017 study. He reports (personal communication, April 17th, 2018) that around a third of learners who demonstrated knowledge of the forms *use*, *reuse*, and *usable* could not do likewise for *reusable*. Also notable in this study is that the 1K and 2K bands contain significantly more forms with two and three affixes as compared with the 3K–5K bands.

Next, the results presented the number of forms accounted for by different levels of Bauer and Nation's word families scheme. This first revealed that among the forms that feature only Level 2 affixes, which are inflectional affixes, there are some containing two affixes (e.g., *find-ing-s* and *low-er-ed*). Thus, forms with inflectional (Level 2) affixes are not necessarily inflections (i.e., while *findings* and *lowered* include inflectional affixes, *findings* is not an inflection of *find* and *lowered* is not an inflection of *low*). This demonstrates that word families are a formal categorization, not a functional or grammatical categorization. Level 2 word families are, therefore, not lemmas nor flemmas, but a somewhat different unit.

Overall, around two-thirds of the word forms feature Level 2 affixes only. Nevertheless, there were significant differences across the bands, with only somewhat over half of the 1K forms featuring Level 2 affixes only, while around 70% do so in the 3K–5K bands.

There are therefore three findings that suggest the higher bands, and particularly the 1K band, may be especially challenging for learners. First, there is a significant difference across the bands in the number of forms. There are simply more forms in the 1K band. Second, the 1K band contains significantly fewer forms with a single affix (e.g., *accept-able*) and significantly more with two and three affixes (e.g., *accept-ab-ly* and *un-accept-ab-ly*). It is likely that such complex forms are more challenging for learners. Third, the 1K band contains significantly fewer forms with Level 2 affixes only (e.g., *accept-ed*) and significantly more with affixes through to Level 3 (e.g., *un-accept-able*; *un-* and *-able* being Level 3 affixes). The actual difficulty of the affixes in Bauer and Nation's scheme has been questioned (Danilović, Savić, & Dimitrijević, 2013; Mochizuki & Aizawa, 2000), but it is the case that the 1K band has proportionally fewer forms featuring affixes that the scheme itself regards as easier. The 1K band, then, which should presumably be the starting point for learners and might be expected to contain the easiest words, in fact may pose the greatest challenge for learners. This may not be a fault with the lists as such, but more a reflection of the nature of language. It does, however, prompt questions about the appropriateness of having fixed bands of 1000 word families (see Brown, 2017; Kremmel, 2016).

Finally in this section, it was seen that in around one-fifth of the word families, the base word is not the most frequent member of the family. Indeed, in around 10% of the families, a form other than the base form is more than twice as frequent as the base form itself. Notwithstanding the fact that in classroom-based learning the base word may be first encountered despite being of lower frequency, this means that for some word families a derived form is likely to be acquired first. Bauer and Nation (1993) suggest this is unproblematic: "once the base word or *even a derived word* [emphasis added] is known, the recognition of other members of the family requires little or no extra effort" (p. 253). Taking an earlier example, this suggests that if a learner knows *activities* (a derived form), little effort is required to understand *active* (the base form) or *actively* (another derived form).

It should be recalled, however, that Ward and Chuenjundaeng (2009) found that Thai learners who could give a translation of a derived form often could not translate the base form. Moreover, it is unclear how the process of dealing with other derived forms is envisaged. Is the assumption that learners are able to understand *actively* simply on the basis of their knowledge of *activities*? Are learners assumed to reverse-derive so-to-speak *active* from *activities* and then derive *actively* from *active*? Whatever the process, it seems likely that such cases do place an additional burden on learners. In an L1 study, Carlisle and Fleming (2003) discovered that children find it more difficult to recognize a base word within a derived form when the affix is unfamiliar (e.g., the unfamiliarity of the affix *-let* caused difficulty in recognizing the base word *tree* in the form *treelet*). By analogy, it may be that L2 learners find it difficult to recognize an affix within a derived form when the base word in that form is unfamiliar (i.e., a learner may struggle to recognize the affixes *-ity* and *-es* within *activities* if unfamiliar with the base word *active*). Learners may, then, not see a derived form as containing affixes at all. Thus, the 22% of word families in which the base word is not the most frequent form, and therefore less likely to be acquired first, may present learners with additional difficulties.

The second research question asked in this study was: If the assumption that learners can deal with WF6 is incorrect, what levels of text coverage might such lists provide? This was pursued by looking at the proportion of the occurrences of a word family that is accounted for by word forms at different levels of the word families scheme. This analysis was based on BNC data and thus is dependent on the

BNC being broadly representative of texts that learners may encounter. This is also of course true of the word lists themselves since they too were based on the BNC.

It was found that on average the base word alone accounts for 62% of the occurrences of all the members in a word family, Level 2 word forms account for around 23% of occurrences, with subsequent levels accounting for 2–5% of occurrences. These figures allow calculations to be made of the coverage levels that may actually ensue if the assumption that learners can easily deal with WF6 is mistaken. For example, if the word lists provide 95% coverage of a text, but learners are unable to cope with WF6, the actual coverage level may be substantially lower, from 59% if the base words alone are known, to 82% if other forms and Level 2 affixes can be dealt with, to 91% if Level 5 affixes are manageable. The point here is not that a narrower definition of the lexical unit would necessarily lead to lower coverage levels being achieved by a word list of a given size (see the coverage levels achieved by Brezina and Gablasova's (2015) lemma-based New General Service List). Rather the point is that if we assume that WF6 are suitable for learners, but it turns out that they are not, we are overestimating coverage to a degree that means learners are likely to experience far more difficulty than we imagine.

These findings have important implications for any work based on the word family. This includes research on vocabulary coverage and comprehension, vocabulary size requirements for a variety of tasks, incidental learning from reading, the balance between explicit and implicit vocabulary learning, estimating learners' vocabulary size, and setting vocabulary learning goals. For instance, a number of studies have explored coverage levels provided by word-family-based word lists and have attempted to estimate the vocabulary size needed to comprehend different types of text (e.g., Adolphs & Schmitt, 2003; Hsu, 2011; Nation, 2006b; Webb & Macalister, 2013). If learners cannot deal with word families, however, the coverage provided by a vocabulary of a given size may be substantially overestimated and the true vocabulary size needed may be far greater. Another example, as Kremmel (2016) and McLean (2017) have discussed, is the considerable number of vocabulary tests that use the word family as the lexical unit. Many tests assume that if a learner displays knowledge of one member of a word family (often the base form), they also have knowledge of all the other members. If this assumption is mistaken, such tests are considerably overstating the extent of learners' vocabulary knowledge. In short, findings and recommendations based on the word family require reevaluation with respect to whether the assumption that the word family is the most appropriate lexical unit is warranted.

5 Conclusion

This study has considered the choice of the word family as the lexical unit by investigating the characteristics of a set of word-family-based word lists. The lists predominantly contain relatively simple forms featuring a single affix from Level 2 of the Bauer and Nation scheme, but also include a large number of more challenging forms. There are many forms that feature multiple affixes or affixes from the higher levels of the scheme, and in many word families, the base word is not the most frequently occurring member of the family. It was also shown that if the assumption that learners can deal with the complexity of word families is incorrect, even to a relatively small degree, the apparent coverage levels the lists provide could be quite mistaken.

By laying out the characteristics of word-family-based word lists, this study provides teachers and researchers with a clearer view of the challenge such lists pose and the consequences if learners are unable to meet this challenge. The study's findings raise questions about research and recommendations based on the use of the word family which do not demonstrate that the learners concerned are able to deal with the scale of the challenge posed. This study also provides, in Table 7, a means of estimating the actual coverage levels that may be reached by the word lists for learners that are able to deal with different levels of the Bauer and Nation scheme. Finally, this study suggests that the complexity of the word family means that in many circumstances, alternative units, such as the lemma or flemma, may be more appropriate, even for learners of rather high proficiency (McLean, 2017).

In addition, the study highlights several matters about which information is lacking and which are crucial in attempts to determine the most appropriate lexical unit for learners. These include: (1) whether learners are able to deal with the various types of 'other forms' the lists contain (irregular verb and noun forms, abbreviated forms, alternative spellings, and compound forms); (2) whether learners are able to cope with an unfamiliar derived form when they are only familiar with a different derived form from that family rather than the base word; and (3) whether forms that contain more than one affix pose a particular challenge for learners.

Research on the above matters should provide direction and a more solid basis for the development of word lists. In addition, this study has revealed one other feature which future developers of word lists may consider: the fact that the 1K band appeared more challenging than the other bands in that it contains the most word forms, more forms with multiple affixes, and more forms with what the scheme considers to be more challenging affixes. It may be beneficial if future lists can find some means of making the highest frequency bands, the likely starting point for learners, less rather than more challenging than later frequency bands.

Acknowledgments

The author would like to thank Geoffrey G. Pinchbeck, Brandon Kramer, and Stuart McLean for their comments on an earlier draft of this article.

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics, 24*(4), 425–438. doi: 10.1093/applin/24.4.425

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253–279. doi: 10.1093/ijl/6.4.253

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101–118. doi: 10.1177/0265532209340194

Bertram, R., Laine, M., & Virkkala, M. M. (2000). The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology, 41*(4), 287–296. doi: 10.1111/1467-9450.00201

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics, 36*(1), 1–22. doi: 10.1093/applin/amt018

Brown, D. (2013). Types of words identified as unknown by L2 learners when reading. *System, 41*(4), 1043–1055. doi: 10.1016/j.system.2013.10.013

Brown, D. (2017). Coverage-based frequency bands: A proposal. *Vocabulary Learning and Instruction, 6*(2), 52–60. doi: 10.7820/vli.v06.2.Brown

Carlisle, J. F., & Fleming, J. (2003). Lexical processing of morphologically complex words in the elementary years. *Scientific Studies of Reading, 7*(3), 239–253. doi: 10.1207/s1532799xssr0703_3

Cobb, T. (no date). Web Vocabprofile, an adaptation of Heatley and Nation's (1994) *Range*. Retrieved from http://www.lextutor.ca/vp/

Coniam, D. (1999). An investigation into the use of word frequency lists in computing vocabulary profiles. *Hong Kong Journal of Applied Linguistics, 4*(1), 103–124.

Danilović, J., Savić, J. D., & Dimitrijević, M. (2013). Affix acquisition order in Serbian EFL learners. *Romanian Journal of English Studies, 10*(1), 77–88. doi: 10.2478/rjes-2013-0006

Davies, M. (2004). *BYU-BNC: The British National Corpus*. Retrieved from http://corpus.byu.edu/bnc

Gardner, D. (2007). Validating the construct of *word* in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28*(2), 241–265. doi: 10.1093/applin/amm010

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327. doi: 10.1093/applin/amt015

Hsu, W. (2011). The vocabulary thresholds of business textbooks and business research articles for EFL learners. *English for Specific Purposes, 30*(4), 247–257. doi: 10.1016/j.esp.2011.04.005

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly, 50*(4), 976–987. doi: 10.1002/tesq.329

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30.

McLean, S. (2017). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*. Advance online publication. doi: 10.1093/applin/amw050

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System, 28*(2), 291–304. doi: 10.1016/S0346-251X(00)00013-0

Nagy, W., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24*(3), 262–282.

Nation, I. S. P. (2006a). *BNC-based word lists.* Wellington: Victoria University of Wellington. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P. (2006b). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review/La revue canadienne des langues vivantes, 63*(1), 59–81. doi: 10.3138/cmlr.63.1.59

Nation, I. S. P., & Heatley, A. (2002). *Range: A program for the analysis of vocabulary in texts.* Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, P. (2015). Which words do you need? In J. Taylor (Ed.), *The Oxford Handbook of the Word* (pp. 568–581). Oxford: Oxford University Press.

Pinchbeck, G. (2014, March). *Lexical frequency profiling of a large sample of Canadian high school diploma exam expository writing: L1 and L2 academic English.* Paper presented at the American Association of Applied Linguistics, Portland, OR.

Reynolds, B. L. (2015). The effects of word form variation and frequency on second language incidental vocabulary acquisition through reading. *Applied Linguistics Review, 6*(4), 467–497. doi: 10.1515/applirev-2015-0021

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26–43. doi: 10.1111/j.1540-4781.2011.01146.x

Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System, 37*(3), 461–469. doi: 10.1016/j.system.2009.01.004

Webb, S., & Macalister, J. (2013). Is text written for children useful for L2 extensive reading? *TESOL Quarterly, 47*(2), 300–322. doi: 10.1002/tesq.70

Webb, S. A., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal, 44*(3), 263–277. doi: 10.1177/0033688213500582