Applied Linguistics 2020: 41/6: 971–998
doi:10.1093/applin/amz051 Advance Access published on 21 November 2019

How Much Knowledge of Derived Words Is Needed for Reading?

[1,*]BATIA LAUFER and [2]TOM COBB

[1]Department of English Language and Literature, University of Haifa, Haifa 3498838 and [2]Department of Didactique des langues, Faculty of Education, University of Quebec at Montreal, Montréal, QC H2L 2C4

*E-mail: batialau@research.haifa.ac.il

The study explores the usefulness of the word family as the unit of counting in studies of lexical coverage and comprehension. It determines the proportion of texts covered by the various members of a word family, that is, basewords, inflected words, and derived words, and analyzes the contribution of the affixed words to lexical thresholds. This exploration was performed by a text analysis computer program called Morpholex that analyzes the entire lexis of an entered text, pulling out all words bearing prefixes and suffixes and counting the unaffixed words as basewords. We analyzed a variety of texts, academic and narrative, authentic and simplified, and calculated the number and percentage of basewords and affixes in each text. We also located the most frequent affixes in our text corpus and demonstrated which affixes and how many contributed to 95 per cent and 98 per cent text coverages. Our results show that reaching the lexical thresholds for reading does not require the knowledge of most of the derived words in a word family since a small number of frequent affixes will provide the necessary coverage together with the basewords and inflections.

INTRODUCTION

Corpus-based L2 vocabulary research has been highly influential in informing language pedagogy about possible vocabulary learning targets and in providing data for the development of vocabulary tests, levels tests, and size tests. This research also has important implications for reading in a second language. Analyses of corpora show that a large percentage of the lexis of a language is accounted for or 'covered' by a relatively small number of frequent words. For example, the first 1,000 most frequent English word families cover 77.96 per cent of all the words in the British National Corpus (BNC; Oxford University Computing Services 2005), the second 1,000 cover 8.10 per cent, the third 4.36 per cent, and the fourth 1.77 per cent (Nation 2013). Since high-frequency vocabulary is more useful than low-frequency vocabulary for both comprehension and production of a new language, these are the words that should be learnt first. Over the years, this insight has given rise to several applications that have benefited not only researchers but also teachers, material writers, and test makers. These are various word frequency lists, text profiling, including profiling of learner texts, and vocabulary testing. Most, though not all lists, profiles, and tests have used word families as the unit of counting. 'Word family' refers to a baseword, its inflections, and its derived words with their respective affixes. An example of a word family is *avoid* (baseword), *avoids, avoided, avoiding* (inflections), *avoidance, avoidable, unavoidable* (derived words), and *avoidances* (inflection of a derived word).

The most influential word list was the 2,000 headword *General Service List of English Words* (West 1953), especially in the development of schemes for graded readers. A recent influential frequency list is the *BNC/COCA word family list* which includes 25,000-family lists based on frequency and range data. The Corpus of Contemporary American English (COCA) corpus is by Davies (2008) and available at english-corpora.org/coca; the BNC/COCA lists are available at Paul Nation's website at https://www.victoria.ac.nz/lals/about/staff/paul-nation.

The criteria used to make word families in the BNC/COCA lists were based on Bauer and Nation's (1993) hierarchy of affixes that consists of one level of inflectional and five levels of derivational affixes that are classified mainly on the basis of frequency, regularity of spelling, and productivity in forming new words. Examples of the derivational affixes at the most basic level are *~able, ~ly, and un~*; examples toward the more challenging and less frequent end of the scheme include *~esque*, *~ward*, and *hyper~*. There are roughly 100 affixes in the hierarchy. The full list can be seen in Online Appendix 1 (Supplementary material).

Frequency lists that use the word family unit have many applications. The lists are widely used for computerized lexical profiling of texts (available at https://lextutor.ca or from Laurence Anthony, https://www.laurenceanthony.net/). By matching a text with the lists, these programs show the proportions of word families at different frequency levels. An example of such an application is Nation's (2006) seminal study on the amount of vocabulary necessary for reading. He found that the most frequent 8,000–9,000 word families provided 98 per cent coverage of written discourse while the most frequent 6,000–7,000 word families covered 98 per cent coverage of spoken discourse. A timeline of the roughly 50 published studies employing the word family as their principal unit of word counting can be found in Nurmukhamedov and Webb (2019), beginning in the 1950s and accelerating with the development of computational tools. Of particular interest are the recent coverage studies that compare the vocabulary size of learners, or the percentage of words known by them in a written or spoken text, to the level of comprehension of the text (Stæhr 2008; Laufer and Ravenhorst-Kalovski 2010; Schmitt *et al.* 2011; van Zeeland and Schmitt 2013).

The lists that use word families as the counting unit have also been used in the random selection of test items at various levels of frequency for various vocabulary tests. Examples of some well-known tests are the Vocabulary Levels Test (Nation 1983; Schmitt *et al.* 2001), the Computer Adaptive Test of Size and Strength (CATSS; Laufer *et al.* 2004; Laufer and Goldstein 2004; Aviad-Levitzky *et al.* 2019), Vocabulary Size Test (VST; Nation and Beglar 2007), and the Updated Vocabulary Levels Test (Webb *et al.* 2017). As most size tests are tests of receptive knowledge, the underlying assumption behind a correct test answer on these receptive measures is that if the item (usually a baseword) is understood, many or all of the item's other family members can be understood as well (Bauer and Nation 1993).

Recently, however, questions have been raised regarding the validity of the word family as a counting unit (Gardner 2007; Kremmel 2016; McLean 2017). Arguments against measuring vocabulary size using word families have been leveled on the grounds that such tests overestimate learners' true vocabulary size. According to this view, receptive knowledge of basewords cannot be assumed to extend to knowledge of their derived forms.

Similarly, it has been argued that scores on tests of vocabulary size that use the family unit are misleading with regard to predicting learners' ability to read authentic materials. Research shows that learners need to know, receptively, 5,000 word families in order to read authentic texts with support, for example, with a dictionary, and 8,000 word families to become independent readers. (Hu and Nation 2000; Laufer and Ravenhorst-Kalovski 2010). However, if learners' performance on family-based vocabulary size tests cannot be extrapolated to their knowledge of derived words, and derived words contribute a substantial amount of text coverage, then measuring vocabulary knowledge in terms of word families is indeed questionable. To illustrate, we might ask: What does it mean if performance on a family-based vocabulary size test indicates that a learner's knowledge is at the 5,000 level? Can we assume that the learner understands 95 per cent of the words she meets in her reading and is able to comprehend texts at the 'with support' level outlined above? For those who question the family assumption, it is safer to say that the learner knows 5,000 lemmas and that 95 per cent criterion is not fully met. The lemma, it is argued by critics of the word family above, is a better unit for assessing vocabulary knowledge because it includes only a baseword and its inflections (e.g. *avoid*, *avoids*, *avoided*, *avoiding*), and it can safely be assumed that instructed learners would know these. Words derived from *avoid* (*avoidable, avoidance*), however, are different lemmas, with their own frequencies, and should not be assumed to be known just because *avoid* is known.

The objection to the family-based approach to assessing vocabulary size and predicting reading comprehension rests on two implicit assumptions. The first is that learners do not possess, or cannot use, the morphological knowledge that is necessary to infer the meaning of a derived word even if they know the meaning of its baseword. The second assumption is that derived words are so frequent in the texts read by learners that lack of their understanding will substantially reduce the lexical coverage of the text and hamper text comprehension. For example, if 20 per cent of the words in a given text were derived forms, and learners could not be assumed to know or infer any of these, then knowledge of all the basewords and inflected forms in the text would provide only 80 per cent coverage, and this would be insufficient for comprehension.

The empirical evidence for these two assumptions is quite limited. We will not dwell on the studies that have investigated how well learners understood derived words if basewords were familiar as the focus of our article is the second assumption, namely, that affixed words are highly frequent in the texts that learners read, and, therefore, are an obstacle to comprehension if unknown.

To our knowledge, there is one study that investigated the amount of coverage that basewords and affixed words provide. Brown (2018) examined the first five BNC lists of 5,000 word families and calculated the varying degrees of coverage that may be provided within the 100 million word corpus from which they were drawn, in a series of conditions: first, if learners are familiar with basewords only, then with inflections, then with derived words at each affix level of the Bauer and Nation affix hierarchy. Brown demonstrates that the full range of morphemes both inflectional and derivational are widely represented in the corpus, and from this he infers that if learners are unable to deal with the derived words even to a relatively small degree, word family-based lists may provide far lower text coverage than has been supposed. For example, an assumed text coverage of 95 per cent by the first 5,000 word families may actually amount to 82 per cent in the BNC—if learners have difficulties with all the derivational affixes and can rely only on basewords and their inflections when they read. The coverage increases to 89.5 per cent if derivational affixes at Levels 3 and 4 of the Bauer and Nation scheme are known to the learners. It reaches 95 per cent with the knowledge of affixes at Levels 5 and 6.

If the BNC reflects the kind of reading students do, then the above findings should be similar in the specific texts they might choose or be asked to read. However, we may find that particular texts differ from a large corpus comprising many texts in terms of the contribution of affixed words. Our research addresses this issue. We examine the distribution of affixed words in a variety of texts and their effect on the lexical coverage required for reading at 95 per cent and 98 per cent levels. The study also examines text coverages of affixes at the various Bauer and Nation levels. In addition, the frequency of individual affixes irrespective of level is investigated to determine whether specific affixes contribute differently to text coverage and whether their contribution is indeed in clusters of affix levels, as suggested by Bauer and Nation (1993). The overall goal is to determine the proportion of texts covered by the various members of a word family, that is, basewords, inflected words, and derived words, and to analyze the contribution of the affixed words to lexical thresholds.

In our definition, a word family consists of a baseword, its inflections, and all the derived words (with their inflections) formed from four types of derivational affixes. In this, we follow Bauer and Nation's (1993) classification, which has seven levels of which we have used the first six: Level 1—basewords, Level 2—inflectional affixes, and four further levels (3–6) of derivational affixes based on frequency, regularity, and productivity. Level 7 is not applicable to our argument because it is mainly affixations of Greek or Latin roots which are not basewords that many learners would know.

We ask the following research questions:

1. How many basewords and affixed word tokens (inflected and derived) are there in a variety of authentic texts that reflect reading that learners of English might undertake at various stages of their development?

2. How are the affixed words distributed in these texts in terms of Bauer and Nation affix levels?

3 What are the most frequent derivational affixes in the texts (across the various levels)?

4. What knowledge of derived words is necessary to reach 95 per cent and 98 per cent of text coverage?

METHODS

Morpholex: text analysis features and tool development

This exploration was performed by a text analysis computer program called Morpholex (available for use at https://www.lextutor.ca/cgi-bin/morpho/lex/). The program analyzes the entire lexis of an entered text, first eliminating proper nouns, then pulling out all words bearing Bauer and Nation prefixes and suffixes, and counting the unaffixed words as basewords. In addition to the unaffixed content words resulting from this process, basewords include figures, function words, bound morphemes with affixes, and irregular inflected forms (*got, begun, bent*). The latter two were treated as basewords as they cannot be understood by decomposition into a base and affix(es).

Clearly, not all instances of ~*ly* or ~*age* etc. are suffixes (e.g. *fly, wage*), so to avoid classing such words as derivations, the program strips the prefix or suffix (or both) from each input word and checks that what remains is a real word. To do this, it checks the presence of the remainder of the item in a test lexicon of 92,620 items (available at https://lextutor.ca/cgi-bin/morpho/bnc_coca_25/all_25k.txt). This test lexicon has been augmented with additional items to accommodate standard orthographic changes involved in affixation. For instance, all words ending in ~*e* were also supplemented by a version of the word without this ending (*close* and *clos*) so that when ~*ing* is removed from *closing*, the program will find the stem *clos* and accept the affixation. Words ending in ~*y* (like *beauty*) were given a second version ending in *i* so that *beautiful* would be accepted. About a dozen such wholesale modifications of the test lexicon were made, to accommodate consonant doubling, vowel deletion, spelling variation (~*s* and ~*es*), and others on a principled basis.

Some items in Bauer and Nation's affix lists themselves were also modified to avoid over- or under-classifying words whether as inflections or derivations. For example, ~*er* is an inflection when added to one-syllable adjectives (*bigger, smarter*) or a derivational affix when added to verbs (*driver, thinker*). This affix occurs as ∼*or* in longer words (*invigilator)*. Two modifications to the framework were made to handle this suffix: First, ∼*or* (omitted in the original framework) was added as a derivation. Furthermore, the ∼*er* suffix was divided into inflectional and derivational categories. A short stop list of the inflectional ∼*er* words was drawn up, which the program consults when it finds an ∼*er* word attached to a real stem, triaging the item into either Level 2 (inflections) or Level 3 (the first layer of derivations). The ∼*ible* variant of suffix ∼*able* was not present in the original scheme and so was added.

Morpholex was extensively trained and modified over several hundred runs with a variety of text types. Human checks were performed to make sure that each stem+affix identified is a true affixed word. For example, the remainder checks methodology described above allows *figure* to count as an affixation inasmuch as *fig* is a real word, but *fig* and *figure* have no relationship. As such cases were found, false basewords were added to a pair of stop lists which the program uses to check every word of an entered text. These stop lists currently comprise 152 blocks for prefixes (e.g. *til* is a prefix block, so that *until* will not be counted as an affixed word) and 363 for suffixes (*stud* is a suffix block, so that *study* will not be counted). A further human check was performed against the inflation of false basewords by having Morpholex include in its output a listing of all the words it had counted as basewords, so that any errors could be corrected or incorporated into the test lexicon or stop lists. Improvements are ongoing. The error rate in August 2019 is less than 1 per cent of classified words.

Examples of Morpholex input and output

Figure 1 shows a text that was input into Morpholex as an example. The text is a 201-word text from Wikipedia providing the legal definition of 'mail fraud'. Legal texts are notoriously dense and Latinate (the Greco-Latin side of English provides the majority of derivations). It has been chosen as a test case because of its lexical and morphological diversity. In terms of its lexical coverage, as measured by one of the lexical frequency profile instruments mentioned above, 95 per cent coverage corresponds to knowing 7,000 word families (5,000 is more typical) and 98 per cent to 10,000 word families (8,000 is more typical), so this is a lexically 'difficult' text. But here, we are interested in morphological coverage—the number of affixes a reader must know in addition to basewords to decode the text. Figure 2 is a screenshot showing the affix profile and coverage for the text. Figure 3 shows the baseword check and the basewords with their frequencies in the text.

In Figure 2, we see the number of words in the text that are inflected or derived forms. As the first line of the output shows, 46 of the 201 words are affixed, amounting to over 20 per cent. It then gives a profile of the proportion of affixed words from Bauer and Nation's Levels 2 through 7 and the coverage these provide in the overall text. (Level 7 is included only to assure that the total is 100 per cent but its output is not included in the analysis.) As shown in

Whoever, having devised or intending to devise any scheme or artifice to defraud, or for obtaining money or property by means of false or fraudulent pretenses, representations, or promises, or to sell, dispose of, loan, exchange, alter, give away, distribute, supply, or furnish or procure for unlawful use any counterfeit or spurious coin, obligation, security, or other article, or anything represented to be or intimated or held out to be such counterfeit or spurious article, for the purpose of executing such scheme or artifice or attempting so to do, places in any post office or authorized depository for mail matter, any matter or thing whatever to be sent or delivered by the Postal Service, or deposits or causes to be deposited any matter or thing whatever to be sent or delivered by any private or commercial interstate carrier, or takes or receives therefrom, any such matter or thing, or knowingly causes to be delivered by mail or such carrier according to the direction thereon, or at the place at which it is directed to be delivered by the person to whom it is addressed, any such matter or thing, shall be fined under this title or imprisoned not more than 20 years, or both.

From https://en.wikipedia.org/wiki/Mail_and_wire_fraud

Figure 1: Input text—mail fraud definition
Source: Wikipedia contributors.

the upper portion of Figure 2, basewords in this text comprise 77.1 per cent of tokens; basewords and inflected words comprise 91.5 per cent; and so on. The Bauer and Nation levels at which coverage reaches (or exceeds) 95 per cent and 98 per cent are shown in bold, which for this text are Levels 4 and 6, respectively. The right side of the profile gives the specific derivational affixes by frequency from any level, in addition to basewords and inflections, that were needed to reach (or exceed) these coverage criteria (e.g. three affixes in six instances were required to make 95 per cent, including two each of the suffixes ∼*er* and ∼ion and the prefix *re*∼, and so on). Then in the columns below the profile, all the affixations identified by the program are listed by level, with color-coding on the affixed part. Visual output reveals any problem in need of repair, for instance, here that *depository* has been included twice, once for suffix ∼y and once for ∼ory, since stripping these suffixes has still left behind a real word (*deposit* + *-ory* and *depositor* + *-y*). It also reveals any artifacts of processing the Bauer and Nation levels separately, for example, here that *unlawful* has been counted twice because *un*∼ is Level 3 and ∼*ful* is Level 4. So with the present methodology, the derivations count may be slightly higher than it should be.

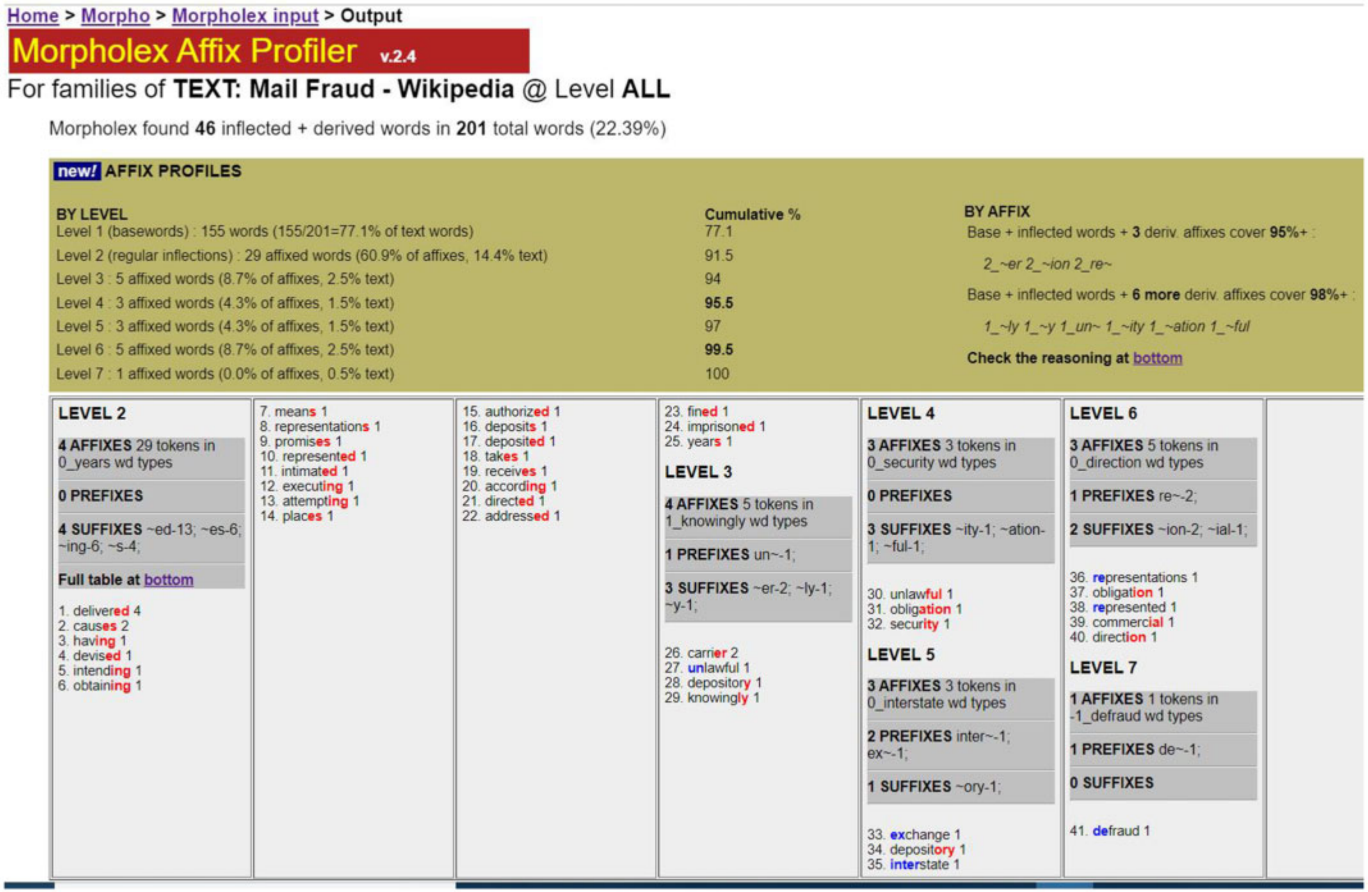

Figure 2: Morpholex analysis of mail fraud text

BASEWORD CHECK:

These words have been classified as basewords

or to any be by such the matter for thing of scheme artifice counterfeit spurious article mail

whatever sent at it is whoever devise money property false fraudulent pretenses sell dispose

loan alter give away distribute supply furnish procure use coin other anything held out purpose

so do in service office private therefrom thereon place which person whom shall under this title

not more than both

Figure 3: Program's baseword check for mail fraud text

Any further problems with particular words are revealed by the baseword check, shown in Figure 3, which lists all the words the program has regarded as nonaffixations. Like the column output, this check is stored over multiple runs and periodically fed back into the program to 'teach' it what we understand by affixation. This particular baseword check raises an important methodological point. The words *devise, dispose, distribute, spurious, property*, and *pretenses* do indeed include affixes (*de∼, dis∼, pro∼, pre∼*, and *∼ous*) but we are not considering these words as affixations in the present context. We are counting only cases where a baseword could be potentially known to a learner yet its derivations not known, or not recognized (the case that is stated or implied in the critique of the family concept). Clearly the hypothetical basewords that remain with the removal of these affixes (*∼vise, ∼pose, ∼tribute, spur∼, ∼pert, ∼tense*) are either nonwords or bound morphemes in current English (though they may be words in Latin or Greek) or are different words unrelated to the derivation in question. A further issue is ambiguous cases, such as the prefix *re∼* in *represent.* Though knowing *represent* as a variant of *present* seems improbable, our analysis nonetheless counts such cases as affixations—with a view to avoiding potential underestimation of derived words.

To answer our research questions, we entered the following types and lengths of preselected texts into Morpholex and recorded the results: five applied linguistics articles of typical length [1], five quality news articles (in the league of the *New York Times*, *Washington Post*, and *The Guardian*), five classic unsimplified novels for native speakers, and six simplified novels for language learners (one for each of six proficiency levels). (Information on the texts can be found in the Supplementary materials). Each text was analyzed by the program with the output for each including the numbers and percentages of basewords, inflected words, and derived words at each of the six Bauer and Nation levels, and of individual affixes and the coverage that each added to its text.

RESULTS

Table 1 shows the results of these text analyses in terms of the components described in the mail fraud example. Column 1 gives the text types and titles; Column 2 gives the number of words (tokens) in each text; Column 3 gives the percentage of basewords, Columns 4–8 give the cumulative percentages at each Bauer and Nation affix level, with the 95 per cent criterion of text coverage indicated in light gray shading and 98 per cent criterion in dark gray; Column 9 (per cent Derivs) gives the total coverage percentage of derived words for the particular text. Below each text type follows the mean coverage for that type at each level, with standard deviations, and below the final text type the total mini-corpus size (243,731 word tokens) and cumulative coverages.

With the information presented in Table 1, we are in a position to answer our first two research questions, concerning the type of texts learners of English might be required or inclined to read at various stages of their development.

1. How many basewords and affixed word tokens (inflected and derived) are there in a variety of texts?

The averages and standard deviations at the bottom of Table 1 show a reasonably regular pattern in the proportion of basewords and inflections despite the variety of text types. Basewords comprise 75–87 per cent of the total words in most texts (Mean = 81.65%, SD = 5.29%), and basewords plus inflections are over 94 per cent (Mean = 94.17%) even more consistently (SD = 2.34%). The remaining roughly 6 per cent of words are the derived words (Bauer and Nation Levels 3–6), although the percentage varies by text type. The averages for each type are shown individually for each text type in Table 1, in a steady decline from 7.78 per cent and 7.88 per cent for academic articles and quality news (SD = 1.40 and 1.53, respectively) to 5.04 per cent for classic novels (SD = 0.65), and 3.17 per cent for graded novels (SD = 0.98).

2. How are the affixed words distributed in a text in terms of B&N affix level?

In the bottom rows of Table 1, we see that Morpholex has been able to categorize 99.77 per cent of the total word tokens in this sample (the remainder being URLs, email addresses, and the like). Subtracting the mean percentage of basewords (81.65 per cent) from this figure reveals that the proportion of all affixed words taken together amounts to 18.12 per cent. Further subtractions show that this breaks down into 12.52 per cent for inflections and 5.60 per cent for derivations. The mean percentages for derivation coverages by Bauer and Nation level decline evenly, with Level 3 at (96.44 − 94.17=) 2.27 per cent, Level 4 at (97.73 − 96.44=) 1.29 per cent, Level 5 at (98.87 − 97.73=)

Table 1: Affix proportions for a range of texts, with 95 per cent and 98 per cent coverages highlighted

Text type/Title	Number of words	B + N levels (cumulative per cent coverage by level) 1 (base)	2 (inflect)	3	4	5	6	Per cent Derivs
Academic								
Laufer and Ravenhorst-Kalovski (2010)	6,855	75.3	92.5	94.0	95.1	98.1	99.9	7.5
Nation (2006)	6,896	76.6	94.3	96.5	97.3	99.2	99.9	5.7
Schmitt *et al.* (2011)	9,108	74.3	91.2	94.3	95.6	98.5	99.7	8.8
Pawley and Syder (1983)	12,694	77.7	90.7	93.8	95.7	97.0	99.6	9.3
Horst *et al.* (2005)	8,661	72.8	92.4	95.0	96.7	98.2	99.9	7.6
Total	**44,214**							
Mean		**75.34**	**92.22**	94.72	96.08	98.2	**99.8**	**7.78**
SD		**1.92**	**1.40**	1.09	0.90	0.80	**0.14**	**1.40**
Newspapers (April 2019)								
The Globe and Mail (Wente)	748	79.5	91.3	93.7	96.8	97.7	99.8	8.7
NYT (Brookes)	824	80.6	94.7	96.1	97.8	98.6	99.7	5.3
The Washington Post (Slater)	1,364	75.4	92.3	94	96.2	97.8	99.7	7.7
The Guardian (Gambino)	828	77.4	91.4	94.3	96.6	97.9	99.8	8.6
National Post (Murphy)	999	77.8	90.9	93.5	96.7	98.4	98.9	9.1
Total	**4,763**							
Mean		**78.14**	**92.12**	94.32	96.84	98.10	99.60	**7.88**
SD		**2.00**	**1.53**	1.04	0.59	0.37	0.40	**1.53**
Classic novels (max. 25k words)								
The Call of the Wild (1903)	21,883	82.8	95.8	98.4	99	99.5	99.8	4.2

Text type/Title	Number of words	B + N levels (cumulative per cent coverage by level)						Per cent
		1 (base)	2 (inflect)	3	4	5	6	Derivs
Lady Chatterley (1928)	25,000	85.3	94.2	97	98.4	99.1	99.7	5.8
Hard Times Bk. I (1905)	25,000	85.1	95.4	97.3	98.1	98.9	99.7	4.6
The Great Gatsby (1925)	25,000	83.2	94.9	97.3	98.2	99.5	100	5.1
The Turn of the Screw (1898)	25,000	86.6	94.5	97.1	98.0	99.4	99.9	5.5
Total	121,883							
Mean		**84.60**	**94.96**	97.42	98.34	99.28	99.82	**5.04**
SD		**1.58**	**0.65**	0.56	0.40	0.27	0.13	**0.65**
Graded readers								
Witches of Pendle (Bookworms 1)	4,913	89.2	97.7	99.3	99.5	99.7	99.8	2.3
Speckled Band (Bookworms 2)	5,307	89.5	97.9	99.1	99.4	99.9	100	2.1
Love Story (Bookworms 3)	7,160	88.1	97.1	98.7	99.5	99.9	100	2.9
Lord Jim (Bookworms 4)	17,900	84.9	95.5	98.6	99.1	99.6	99.7	4.5
The Bride Price (Bookworms 5)	17,504	87.1	97	99	99.5	99.8	100	3
Cold Comfort Farm (Bookworms 6)	25,000	85.5	95.8	98.3	99.1	99.5	99.7	4.2
Total	72,871							
Mean		**87.38**	**96.83**	**98.83**	**99.35**	**99.73**	**99.87**	**3.17**
SD		**1.90**	**0.98**	**0.37**	**0.20**	**0.16**	**0.15**	**0.98**
Cumulative total	243,731							
Overall cumulative means		81.65	94.17	96.44	97.73	98.87	99.77	5.60
Overall SD		5.29	2.34	2.09	1.43	0.84	0.23	

Note: Boldface is used to to highlight trends.

1.14 per cent, and Level 6 at (99.77 − 98.87=) 0.90 per cent. To be noted is that there are differences for text types, as can be seen in the leftward drift of the gray coverage threshold highlighting from top to bottom in Table 1.

3. What are the most frequent derivational affixes in authentic texts (across levels)?

Table 1 shows the levels of the Bauer and Nation affix hierarchy that are needed to reach the 95 per cent and 98 per cent criteria of text coverage (e.g. for the first academic paper, Level 4 is needed to reach 95 per cent and Level 5 for 98 per cent). However, our data analysis also indicates that just a few affix types from these levels are highly repeated. Table 2 shows the specific affixes (across levels) with their frequencies in each text that are needed to reach each coverage criterion. The first four columns in Table 2 include figures already shown in Table 1; Columns 5 and 6 further specify the number of types, tokens, and the specific derivations needed to reach or exceed 95 per cent; Columns 7 and 8 provide the additional (+) types and derivations needed to reach or exceed 98 per cent. For example, for the first academic paper, Table 2 shows that basewords and inflections provide 92.5 per cent coverage, with just three further derivations needed to reach 95 per cent and another seven for 98 per cent.

To find out which affixes were most frequent, we assembled all the affixes from Table 2 that had been needed to reach or exceed both the 95 per cent and 98 per cent criteria and calculated the number of tokens and types. The result was 8,168 tokens in 36 types, but only a small handful was recurrent. This is shown in Tables 3 and 4. Table 3 gives the affix frequencies for the entire corpus; Table 4 gives the affix frequencies by text type. (The percentages in the tables are based on the total number of affixes, not the total number of words in the texts, i.e. they are not coverages.)

Table 3 shows that across the corpus just three suffixes account for more than 50 per cent of the total derivational affixations (*∼ly*, *∼ion*, and *∼er*); just 17 account for over 95 per cent. The distribution varies to some extent with text type as shown in Table 4.

While academic texts and newspapers include derived words with 25 and 26 different derivational affixes (types), respectively, the novels comprise only 13 different derivational affixes and graded readers only seven. But even in the derivation-heavy texts, just a few affixes form slightly more than half of the derived words. In academic texts these are *∼ly, ∼ion, ∼age, and ∼al* (55 per cent of all affixed words, Table 4) and in newspapers they are *∼ly, ∼ion, ∼al*, and *∼ment* (55.58 per cent).

In terms of distribution, many affixes are well represented across three or more text types (*∼ly, ∼ion, ∼er, re∼, ∼un, ∼able*) while others are quite restricted (*∼age* appears only in academic text) and in different ranks or proportions (*∼ful* and ∼ness are plentiful in stories of either type, less so in academic or press).

Table 2: Derivational affixes needed to reach 95 per cent and 96 per cent text coverage

Text type/Title	Words	B + N (Cumulative per cent)		Affixes needed to reach ...			
		1	2	95 per cent		98 per cent	
				Types	Tokens	Types	Tokens
Academic							
Laufer and Ravenhorst (2010)	6,855	75.3	92.5	3	*121_~age 47_~ion 45_~ly*	+7	*37_re~ 34_~al 29_~ic 21_~ent 16_~er 15_~y 15_~ship*
Nation (2006)	6,896	76.6	94.3	1	*78_~age*	+8	*64_~ly 29_~ion 26_un~ 13_~ation 12_~able 12_pro~ 11_~ity 10_~er*
Schmitt *et al.* (2011)	9,108	74.3	91.2	3	*165_~age 122_~ly 61_~ion*	+11	*32_~ity 30_~able 29_~ship 28_non~ 27_~al 24_~er 22_~y 22_un~ 19_in~ 18_~ment 18_~ic*
Pawley and Syder (1983)	12,694	77.7	90.7	4	*183_~ly 181_~ion 115_~al 70_~er*	+10	*47_~ity 47_~ic 44_~ive 40_~y 40_un~ 40_~ation 31_~able 28_re~ 28_~ally 26_~ar*
Horst *et al.* (2005)	8,661	72.8	92.4	4	*97_~ly 54_~ion 48_~er 47_~al*	+11	*31_~able 31_~ition 28_~ity 25_~ance 22_~ful 22_~ic 21_re~ 20_~ive 19_in~ 18_~ary 18_~ent*
Total	**44,214**						
Mean		**75.32**	**92.22**	**3**		**+9.4**	
SD		**1.92**	**1.4**	**1.22**		**1.80**	

Text type/Title	Words	B + N (Cumulative per cent) 1	2	Affixes needed to reach ... 95 per cent Types	Tokens	98 per cent Types	Tokens
Newspapers (April 2019)							
The Globe and Mail (Wente)	748	79.5	91.3	4	*11_~al 7_~ly 7_~ion 5_~ment*	+7	*4_~ity 4_~ive 3_~y 3_~ful 3_~ial 2_~ness 2_~er 2_~ship 1_~or*
NYT (Brookes)	824	80.6	94.7	1	*9_~ly*	+4	*7_~ity 5_~al 3_re~ 3_~ion*
The Washington Post (Slater)	1,364	75.4	92.3	4	*12_~ion 12_~ial 8_~ity 7_~ly*	+8	*6_~er 6_~al 6_ex~ 5_~ment 4_un~ 4_~ist 4_~ent 4_inter~*
The Guardian (Gambino)	828	77.4	91.4	3	*13_~ment 10_~ly 9_~ion*	+5	*6_~ent 5_~able 4_~al 4_~ify 3_un~*
National Post (Murphy)	999	77.8	90.9	4	*17_~ly 11_~al 10_~ment 7_~et*	+8	*5_~ally 4_un~ 4_~ation 4_~ion 3_~ous 3_~ure 2_~ness 2_~ity*
Total	4,763						
Mean		**78.14**	**92.12**	**3.20**		**+6.4**	
SD		**2.00**	**1.53**	**1.30**		**1.8**	
Classic novels (max. 25k words)							
The Call of the Wild (1903)	21,883	82.8	95.8	0		+5	*276_~ly 81_~er 67_un~ 60_~y 48_~ness*
Lady Chatterley (1928)	25,000	85.3	94.2	1	*360_~ly*	+10	

Text type/Title	Words	B + N (Cumulative per cent)		Affixes needed to reach . . .			
		1	2	95 per cent		98 per cent	
				Types	Tokens	Types	Tokens
							92_~ness 78_~ion 75_~y 74_~al 66_~er 54_in~ 45_~ity 44_~less 44_un~ 38_~ful
Hard Times Bk. I (1905)	25,000	85.1	95.4	0		+9	*242_~ly 82_~ion 71_~y 69_re~ 59_~al 53_~er 49_un~ 33_~ation 32_~ness*
The Great Gatsby (1925)	25,000	83.2	94.9	1	*361_~ly*	+7	*86_~ion 73_~er 68_~y 54_un~ 52_re~ 46_~ness 44_~al*
The Turn of the Screw (1898)	25,000	86.6	94.5	1	*339_ly*	+9	*132_~ion 80_~ness 59_re~ 44_~y 41_~ity 41_~ation 37_~er 36_~ful 32_un~*
Total	**121,883**						
Mean		**84.60**	**94.96**	**0.60**		**+8**	
SD		**1.58**	**0.65**	**0.55**		**2.00**	
Graded readers							
Witches of Pendle (Bookworms 1)	4,913	89.2	97.7	0		+1	*48_~ly*
The Speckled Band (Bookworms 2)	5,307	89.5	97.9	0		+0	
Love Story (Bookworms 3)	7,160	88.1	97.1	0		+1	*68_~ly*

Text type/Title	Words	B + N (Cumulative per cent)		Affixes needed to reach ...			
		1	2	95 per cent		98 per cent	
				Types	Tokens	Types	Tokens
Lord Jim (Bookworms 4)	17,900	84.9	95.5	0		+4	*296_~ly 59_~y 48_~er 42_~ness*
The Bride Price (Bookworms 5)	17,504	87.1	97	0		+2	*162_~ly 62_~er*
Cold Comfort Farm (Bookworms 6)	25,000	85.5	95.8	0		+5	*408_~ly 63_~y 54_~ful 42_un~ 41_~ion*
Total	**72,871**						
Mean		**87.38**	**96.83**	**0.00**		**+2.2**	
SD		**1.90**	**0.98**	**0.00**		**1.9**	
Cumulative total	243,731						
Overall cumulative means		81.65	94.16	1.62		+6.5	
Overall SD		5.30	2.35	1.69		3.11	

Note: Boldface is used to to highlight trends.

Table 3: Derivational affixes by frequency (entire corpus)

	Affix	Frequency	Per cent	Cumulative per cent
1	~ly	3121	38.21	38.21
2	~ion	826	10.11	48.32
3	~er	596	7.30	55.62
4	~y	520	6.37	61.99
5	~al	437	5.35	67.34
6	un~	387	4.74	72.07
7	~age	364	4.46	76.53
8	~ness	344	4.21	80.74
9	re~	269	3.29	84.04
10	~ity	225	2.75	86.79
11	~ful	153	1.87	88.66
12	~ation	131	1.60	90.27
13	~ic	116	1.42	91.69
14	~able	109	1.33	93.02
15	in~	92	1.13	94.15
16	~ive	68	0.83	94.98
17	~ment	51	0.62	95.60
18	~ent	49	0.60	96.20
19	~ship	46	0.56	96.77
20	~less	44	0.54	97.31
21	~ally	33	0.40	97.71
22	~ition	31	0.38	98.09
23	non~	28	0.34	98.43
24	~ar	26	0.32	98.75
25	~ance	25	0.31	99.06
26	~ary	18	0.22	99.28
27	~ial	15	0.18	99.46
28	pro~	12	0.15	99.61
29	~et	7	0.09	99.69
30	ex~	6	0.07	99.77
31	~ify	4	0.05	99.82
32	~ist	4	0.05	99.87
33	inter~	4	0.05	99.91
34	~ous	3	0.04	99.95
35	~ure	3	0.04	99.99
36	~or	1	0.01	100
	Total	8168		

Table 4: Proportions of derivational affixes by text type

	Academic (44,214 words)				News (4,763 words)				Novels (121,883 words)				Graded (72,871 words)			
#	Affix	#	Per cent	C per cent	Affix	#	Per cent	C per cent	Affix	#	Per cent	C per cent	Affix	#	Per cent	C per cent
1	~ly	511	19.12	19.12	~ly	50	17.61	17.61	~ly	1578	41.34	41.30	~ly	982	70.50	70.50
2	~ion	372	13.92	33.04	~al	37	13.03	30.64	~ion	378	9.90	51.20	~y	122	8.76	79.26
3	~age	364	13.62	46.65	~ion	35	12.32	42.96	~y	318	8.33	59.53	~er	110	7.90	87.15
4	~al	223	8.34	55.00	~ment	33	11.62	54.58	~er	310	8.12	67.66	~ful	54	3.88	91.03
5	~er	168	6.29	61.28	~ity	21	7.39	61.98	~ness	298	7.81	75.46	~ness	42	3.02	94.05
6	~ity	118	4.41	65.70	~ial	15	5.28	67.26	un~	246	6.44	81.91	un~	42	3.02	97.06
7	~ic	116	4.34	70.04	un~	11	3.87	71.13	re~	180	4.72	86.62	~ion	41	2.94	100
8	~able	104	3.89	73.93	~ent	10	3.52	74.65	~al	177	4.64	91.26				
9	un~	88	3.29	77.22	~er	8	2.82	77.47	~ity	86	2.25	93.51				
10	re~	86	3.22	80.44	~et	7	2.46	79.93	~ation	74	1.94	95.45				
11	~y	77	2.88	83.32	ex~	6	2.11	82.05	~ful	74	1.94	97.39				
12	~ive	64	2.39	85.71	~able	5	1.76	83.81	in~	54	1.41	98.81				
13	~ation	53	1.98	87.69	~ally	5	1.76	85.57	~less	44	1.15	100				
14	~ship	44	1.65	89.34	~ation	4	1.41	86.98								
15	~ent	39	1.46	90.80	~ify	4	1.41	88.38								
16	in~	38	1.42	92.22	~ist	4	1.41	89.79								
17	~ition	31	1.16	93.38	~ive	4	1.41	91.20								
18	~ally	28	1.05	94.43	~ness	4	1.41	92.61								
19	non~	28	1.05	95.48	inter~	4	1.41	94.02								
20	~ar	26	0.97	96.45	~ful	3	1.06	95.07								
21	~ance	25	0.94	97.38	~ous	3	1.06	96.13								

#	Academic (44,214 words)				News (4,763 words)				Novels (121,883 words)				Graded (72,871 words)			
	Affix	#	Per cent	C per cent	Affix	#	Per cent	C per cent	Affix	#	Per cent	C per cent	Affix	#	Per cent	C per cent
22	~ful	22	0.82	98.21	~ure	3	1.06	97.19								
23	~ary	18	0.67	98.88	~y	3	1.06	98.24								
24	~ment	18	0.67	99.55	re~	3	1.06	99.30								
25	pro~	12	0.45	100	~ship	2	0.70	100								
26					~or	1	0.35	100								
	Totals	2673	100			284	100			3817	100			1393	100	

Note: To allow side-by-side comparison, percentages are rounded to save space and the following abbreviations are used: #, number; C percent, cumulative percent.

How does our affix frequency compare with Bauer and Nation affix levels? Table 5 lists the top 10 affixes in the entire corpus, that is, affixes that constitute more than 2 per cent of all affix tokens, along with their classification in the Bauer and Nation hierarchy. Four of the 10 affixes do not fit the hierarchy in the sense that they are frequent in our analysis but were assigned a level in the Bauer and Nation scheme that is associated with infrequency and/or potential difficulty. One of these is ~*age*, which was very frequent in these particular academic texts, but not in the other sub-corpora; the others are ~*ion*, ~*al, and re*~. These incongruities are not surprising given that in addition to frequency, the Bauer and Nation framework was intended to reflect learnability factors that were not taken into account here (such as transparency of meaning and generativity). There is possibly a case for revising the framework to move re~ to Level 3, as being frequent, transparent in meaning, and imposing little change on its baseword—indeed Brown (2018) has already done this in his analysis.

4. What morphological knowledge is necessary to reach 95 per cent and 98 per cent of text coverage?

In terms of the number of affixes needed to reach criterion coverages, Table 2 shows that 95 per cent for some types of texts can be reached with no derivational knowledge at all, just basewords and inflections. This is the case for graded stories and even some classic novels written for native speakers. But even in the case of both academic and quality press articles, Table 2 shows that only an average of three different affixes (SD = 1.39 and 1.30, respectively, across texts) must be known to meet 95 per cent coverage. For novels, Table 2 shows that, on average, 94.96 per cent of the lexis consists of just basewords and inflected words. Typically, just one affix, ~*ly*, raises coverage to over 95

Table 5: Most frequent derivational affixes by B&N level

	Affix	Frequency	Per cent of tokens	Cumulative per cent	B&N level
1	~ly	3121	38.21	38.21	3
2	~ion	826	10.11	48.32	6
3	~er	596	7.30	55.62	3
4	~y	520	6.37	61.99	3
5	~al	437	5.35	67.34	5
6	re~	269	3.29	70.63	6
7	un~	451	4.75	75.38	3
8	~age	364	4.46	79.64	5
9	~ness	344	4.21	84.05	3
10	~ity	225	2.75	86.55	4

per cent. To reach 98 per cent coverage requires slightly more knowledge, though the average number of additional affix types here is still an average of nine and eight for academic articles and novels, respectively, decreasing to six for press and two for graded stories (figures are rounded).

Table 6 shows what affixes are needed to reach 95 per cent and 98 per cent coverages not in individual texts but in each text type. It includes information from Tables 2 and 4. Table 2 shows the number of affixes needed to reach 95 per cent and 98 per cent coverage levels in each text type (Columns 5 and 7, respectively), and Table 4 presents the affix frequencies for each text type. For example, in academic texts, three affixes are needed to raise the 92.22 per cent (basewords + inflected words) to 95 per cent and additional 9.4 affixes to raise 95–98 per cent. Based on this information, Table 6 lists the three most frequent affixes needed for 95 per cent coverage and then the additional nine affixes for 98 per cent coverage. The same procedure is followed for all four text types.

Table 6 shows that a relatively small amount of morphology knowledge is required to reach the comprehension criteria, and that this varies to some extent by text type. While *∼ly* is powerful across the text types and *∼ly* and *∼ion* across the first three, *∼age* appears to be important in these particular academic texts (*coverage, usage, percentage*) while *∼al* is common in news (*governmental, societal, national, cultural*). Adjective-making *∼y* is common in stories of both types (*rainy, scary, brainy*), but rare in news.

DISCUSSION

Our study addressed two related issues. First, we wanted to find out the quantity and the distribution of affixed words in a variety of texts that represent the kinds of reading learners of English can be expected to do. To this end, we analyzed a variety of texts, academic and narrative, authentic and simplified,

Table 6: Derivational affixes needed to reach the 95 per cent and 98 per cent coverage levels

Text type	Basewords (per cent)	Basewords + inflections (per cent)	95 per cent coverage	98 per cent coverage
Academic	75.32	92.22	∼ly ∼ion ∼age	∼al ∼er ∼ity ∼ic ∼able un∼ re∼∼y ∼ive
News	78.14	92.12	∼ly ∼al ∼ion	∼ment ∼ity ∼ial ∼un∼ ∼ent ∼er
Novels	84.60	94.96	∼ly	∼ion ∼y ∼er ∼ness un∼ re∼ ∼al ∼ity
Graded readers	87.38	96.83	–	∼ly ∼y

and calculated the number and percentage of basewords, and affixes, in tokens and types in each text (RQ 1). We also showed how the affixes were distributed in each text according to the affix hierarchy suggested by Bauer and Nation (RQ 2).

Second, like Bauer and Nation (1993), we felt that not all word family members were equally important for text comprehension since learners would not encounter some of them as frequently as the others. For example, if the baseword *desire* appears as *desire* 30,839 times in the 560 million word COCA and as *desirable* 5,922 times, as *undesirable* 1,734 times, as *desirability* 1,039 times, and as *undesirability* 28 times, we cannot argue that learners would benefit from knowing all the family members in the same way. Bauer and Nation's hierarchy of affixes is based on the affix coverage they found in the Lancaster–Oslo–Bergen corpus. We wanted to see whether the contribution of affixes to the coverage of specific texts of the type learners might be expected to read was similar. Therefore, we located the most frequent affixes in our text corpus and checked whether the frequency rank order we found corresponded to Bauer and Nation (RQ 3). On the basis of the token counts of basewords, inflections, and individual affixes, we demonstrated which affixes and how many contributed to the 95 per cent and 98 per cent text coverages (RQ 4).

The answer to RQ 1 shows that the average percentage of derived words in texts is 7.78 per cent in academic texts, 7.88 per cent in newspaper articles, 5.04 per cent in authentic novels, and 3.17 per cent in graded readers. This is in striking contrast to the proportion of the number of derived words in large corpora or in frequency lists derived from large corpora. When we entered BNC/COCA family lists into Morpholex, we found that the derived words constituted 30 per cent of each of the first three 1,000 lists, though declining slightly thereafter (27.6 per cent of the fourth 1,000, 25 per cent of the fifth). Lextutor's Familizer/Lemmatizer further shows that the first 1,000 word families includes 6,853 individual word forms, or is equivalent to 4,737 lemmas at a ratio of just under one to five (e.g. one family *avoid* is equivalent to four lemmas *avoid, avoidance, avoidable*, and *unavoidable*, plus inflections, in terms of the earlier example). These figures suggest that knowing a word family is equivalent to knowing nearly five times as many lemmas (baseword and their inflections) if we want to make valid predictions about the number of words required for reading. These figures may also explain why some test-makers insist on using lemmas as the unit of measure. Their argument is that learners' knowledge of a baseword may not extend to knowledge of all its constituent lemmas. However, as our analysis shows, not all lemmas are represented equally in all texts. Therefore such complete knowledge is not normally necessary for reading texts.

Similarly, our results also show that the distribution of derived words in a particular text is different from their distribution in the assembled texts of a large corpus like the BNC. Brown (2018) found that basewords and inflections of the 5,000 most frequent words constituted 86.6 per cent of the BNC corpus.

The other 13.3 per cent were derived words. However, our analysis yielded an average of 5.60 per cent derived words, which is less than half of Brown's corpus figure. The reason for this discrepancy may be explained in two ways. First, it may be explained by Brown's analysis of high-frequency words only, the most frequent 5,000 families. The number of derived words of less frequent basewords is much smaller than the number for high-frequency basewords. As shown above, the proportion of derived words declines as one progresses up the frequency lists. The first 1,000 most frequent word families have 2,564 derived words, the second have 2,437, while the seventh have 1,038, and the tenth only 715 (these figures were obtained by running BNC 1000-family lists through Morpholex). Since texts typically include words beyond the most frequent 5,000, the overall percentage of derived words in them is smaller than is implied by Brown's analysis.

The second reason for the discrepancy could be that the vast collection of texts and text types that make up a corpus are virtually guaranteed to include most or all of the possible affixations of many different words, but this is not to say that all of these affixations will be present in a particular text or even text type. We have seen (in Tables 2 and 4) that different text types appear to have their own 'morphological footprints' or profiles with typical over-, under-, and nonuse of particular affixes (though this would have to be confirmed with a much larger corpus of texts and broader range of text types). We have also shown that the number appears to be quite small and pedagogically manageable. There is no reason that learners should be expected to work with every possible affixation in early or intermediate stages of learning, nor that tests and other measures should be devised that assume or imply such an objective. Indeed the value of Bauer and Nation (1993) and Sasao and Webb (2017) is suggesting which affixes would be useful to know at different stages of learning, and our work addresses the same question, but from a text analysis perspective.

The finding that the frequency of derived words in passages is low represents an important contribution to our understanding of how learners experience reading in their second language. Even more important in terms of its pedagogical significance is the finding that the number of different affixes that make up the derived words in texts is small. Put differently, reaching the lexical thresholds for reading does not require the knowledge of most of the derived words in a word family since a small number of frequent affixes will provide the necessary coverage together with the basewords and inflections. Thus, it is possible to reach 95 per cent of text coverage with three or four derivational affixes in academic and newspaper texts, one affix (*∼ly*) in novels, and none at all in graded readers. Of the five most frequent affixes in our corpus, three correspond to Level 3 in the Bauer and Nation framework (the most basic level of derivations), one to Level 5 and one to Level 6.

How can our rather optimistic findings be reconciled with studies that express concerns about the negative effect of learners' insufficient knowledge of derived words? To our knowledge, no study of learners' derivational knowledge concludes that they do not know *any* of the derived words. Rather, the studies claim that learners do not know *many* of the derived words that were tested in the research or do not know them with total confidence. It is plausible, therefore, that a smaller percentage of the derived words in texts than was previously assumed along with learners' partial knowledge of word families may suffice for comprehension. Let us examine this proposition in light of two types of evidence: classification of affix difficulty that is based on learners' knowledge (Sasao and Webb 2017) and the connection between text coverage, learners' vocabulary size, and reading comprehension scores (Laufer and Ravenhorst-Kalovski 2010).

Sasao and Webb (2017) selected 118 derivative affixes based on frequency data from the BNC and tested L2 learners on three aspects of affix knowledge: recognition of written affix forms, knowledge of affix meanings, and knowledge of the syntactic properties of affixes. On the basis of the results, the researchers classified the affixes into three levels of difficulty: beginner, intermediate, and advanced. Thus the affixes in the 'beginner' group were familiar to most learners among the participants. In our study, text analyses showed that across our corpus most derived words were constructed with 10 affixes *∼ly, ∼ion, ∼er, ∼y, ∼al, re∼, un∼, ∼age, ∼ness*, and *∼ity* (Table 5). Six of these were classified as 'beginner' level by Sasao and Webb, one as 'intermediate' and three as 'advanced'. This means that most of the frequent affixes in our corpus can be assumed to be familiar to most learners. And breaking down our corpus by text type it is indeed the case that graded readers contain only beginner affixes by this classification (*∼ly, ∼ness*), novels mainly these plus an intermediate (*∼ful*), with only news and academic texts containing advanced affixes (*∼ship, ∼age*).

A further support for this familiarity is that family-based vocabulary tests do indeed predict comprehension. The study by Laufer and Ravenhorst-Kalovski (2010) combines data on learners' vocabulary size in word families, lexical coverage of texts they were tested on, and reading comprehension scores. Participants took the Vocabulary Levels Test (Schmitt *et al.* 2001) and the VST (Nation and Beglar 2007), both of which are family-based tests. They were also tested on several expository texts where 5,000 word families plus proper nouns provided coverage of 95 per cent, and 8,000 plus proper nouns provided 98 per cent. Participants who, according to vocabulary tests, knew 8,000 word families did so well on the reading comprehension test that they were exempted from further English for Academic Purposes courses; that is, they were regarded as independent readers. Participants who knew 5,000 word families did well enough to be placed in an advanced reading course, after which they were expected to have become independent readers. Thus, the research shows that the family-based vocabulary size tests were accurate in diagnosing the vocabulary knowledge required to reach the lexical coverage of the texts they were tested on. One of the texts used in the study can be found in an appendix to the specific paper. Our analysis of this literary text using Morpholex showed that the percentage of the derived word tokens was 5.06 per cent (almost identical to the 5.16 per cent identified for novels in our study); we can safely assume that the other texts read in this study were similar—that is, 'normal' in terms of their lexical characteristics. The family-based vocabulary test scores in that study did not show how much morphological knowledge learners had; indeed, it is reasonable to think that they had less than full knowledge of many derived forms. But it is abundantly clear that they had enough morphological knowledge for successful comprehension of the texts.

Earlier we pointed out that the concern about vocabulary tests based on word family knowledge is that they may overestimate the lexical knowledge that learners can apply to reading. Based on the evidence from the two studies above and our text analysis by Morpholex, we contend that this concern is exaggerated and further that there is little reason to reconsider the large amount of useful and influential research that is based on the word family as the unit of counting. Even if a correct answer on a baseword item does not imply knowledge of the entire word family, knowledge of the most frequent affixes that appear in our corpus probably suffices for comprehending most texts.

CONCLUDING REMARKS

In the title of our article, we ask how much knowledge of derived words is needed for reading. The obvious answer is 'as much as is required to understand specific texts'. Our analysis of texts showed that what is required is not knowledge of the entire word families of basewords, but the inflections of basewords and a limited number of the most common affixes that participate in the construction of some derived words. Moreover, most derived words appear in quality news and academic articles, not in novels and not in graded readers. If learners who read newspapers and academic texts have not yet fully mastered the morphological patterns of English, we believe that in view of the limited number of essential affixes, they can rely on partial morphological knowledge for comprehension. We also believe, therefore, that vocabulary tests and predictions of coverage and comprehension based on them that are based on word families are not misleading as has previously been suggested.

Though we have analyzed a corpus of almost a quarter-million words representing four text genres, further studies should be conducted to challenge or supplement our findings by analyzing additional texts and additional genres, for example, scientific texts and conversations. Studies could also check empirically whether learners are in fact more familiar with the affixes we found to be most frequent in texts. On a more practical note, teacher–researchers can use Morpholex to analyze specific written and spoken texts they are teaching in various courses and incorporate the most essential affixes into prereading activities and course syllabi. For example, a teacher leading a class through Bookworm's *Cold Comfort Farm* could ask questions that focus attention on learners' understanding of *∼ly, ∼y, ∼ful, un∼*, and *∼ion* the first few times these were met, since some of these had not appeared in earlier works in the series but appear extensively here (~*ly* 408 times, ~*y* 63 times, ~*ful* 54 times, *un*~ 42 times, and ~*ion* 41 times, as shown in Table 2).

To our knowledge, no study has analyzed the contribution of affixes to real texts as we have tried to do and no tool similar to Morpholex has been previously developed. We hope that our work will inspire useful research on lemmas, word families, and the role of derived forms in comprehension.

NOTE

1 For academic texts we chose articles in *Applied Linguistics* since, based on the second author's experience with Morpholex text analysis, Applied Linguistics papers almost invariably have the highest proportion of derived words.

SUPPLEMENTARY DATA

Supplementary material is available at *Applied Linguistics* online.

Conflict of interest statement. None declared.

REFERENCES

Aviad-Levitzky, T., B. Laufer, and Z. Goldstein. 2019. 'The new Computer Adaptive Test of Size and Strength (CATSS): Development and validation,' *Language Assessment Quarterly* 16: 345–68.

Bauer, L. and P. Nation. 1993. 'Word families,' *International Journal of Lexicography* 6: 253–79.

Brown, D. 2018. 'Examining the word family through word lists,' *Vocabulary Learning and Instruction* 7: 51–65.

Davies, M. 2008. [computer program]. *Corpus of Contemporary American English*. Consulted 13 May 2019. Available at https://www.english-corpora.org/coca/.

Gardner, D. 2007. 'Validating the construct of 'word' in applied corpus-based vocabulary research—A critical survey,' *Applied Linguistics* 28: 241–65.

Horst, M., T. Cobb and I. Nicolae. 2005. 'Expanding academic vocabulary with an interactive on-line database,' *Language Learning & Technology* 9: 90–110.

Hu, M. and **I. S. P. Nation**. 2000. 'Unknown vocabulary density and reading comprehension,' *Reading in a Foreign Language* 13: 403–30.

Kremmel, B. 2016. 'Word families and frequency bands in vocabulary tests: Challenging conventions,' *TESOL Quarterly* 50: 976–87.

Laufer, B. and Z. Goldstein. 2004. 'Testing vocabulary knowledge: Size, strength, and computer adaptiveness,' *Language Learning* 54: 399–436.

Laufer, B., C. Elder, K. Hill, and P. Congdon. 2004. 'Size and strength: Do we need both to measure vocabulary knowledge?' *Language Testing* 21: 202–26.

Laufer, B. and G. Ravenhorst-Kalovski. 2010. 'Lexical threshold revisited: Lexical text coverage, learner's vocabulary size and reading comprehension,' *Reading in a Foreign Language* 22: 15–30.

McLean, S. 2017. 'Evidence for the adoption of the flemma as an appropriate word counting unit,' *Applied Linguistics 39: 823–45*.

Nation, I. S. P. 1983. 'Testing and teaching vocabulary,' Guidelines 5: 12–25.

Nation, I. S. P. 2006. 'How large a vocabulary is needed for reading and listening?' The Canadian Modern Language Review 6: 59–82.

Nation, I. S. P. 2013. Learning Vocabulary in Another Language. 2nd edn. Cambridge University Press.

Nation, I. S. P. and D. Beglar. 2007. 'A vocabulary size test,' *The Language Teacher* 31: 9–13.

Nurmukhamedov, U. and S. Webb. 2019. 'Lexical coverage and profiling,' *Language Teaching* 52: 188–200.

Oxford University Computing Services 1995. The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. Available at: http://www.natcorp.ox.ac.uk/.

Pawley, A. and F.H. Syder. 1983. 'Two puzzles for linguistic theory: Nativelike selection and nativelike fluency,' in J.C. Richards and R.W. Schmidt (eds): *Language and Communication* Longman, pp. 191–226.

Sasao, Y. and S. Webb. 2017. 'The Word Part Levels Test,' *Language Teaching Research* 21: 12–30.

Schmitt, N., X. Jiang, and W. Grabe. 2011. 'The percentage of words known in a text and reading comprehension,' *Modern Language Journal* 95: 26–43.

Schmitt, N., D. Schmitt, and C. Clapham. 2001. 'Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test,' *Language Testing* 18: 55–88.

Stæhr, L. 2008. 'Vocabulary size and the skills of listening, reading and writing,' *Language Learning Journal* 36: 139–52.

Webb, S., Y. Sasao, and O. Ballance. 2017. 'The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT,' *ITL - International Journal of Applied Linguistics* 168: 33–70.

West, M. 1953. *A General Service List of English Words*. Longman, Green and Co.

Wikipedia contributors. 2019. Mail and wire fraud. Wikipedia, The Free Encyclopedia. Available at https://en.wikipedia.org/w/index.php?title=Mail_and_wire_fraud&oldid=916890963. Accessed 26 August 2019.

van Zeeland, H. and N. Schmitt. 2013. 'Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension?' *Applied Linguistics* 34: 457–79.