



What's in a Word Family? The Assumptions of Lexical Units

Phil Bennett 

University of Niigata Prefecture, Japan
philbennett59@gmail.com

Abstract

Lemmas, flemmas, and level 6 word families (WF6) are three commonly discussed lexical units. Because each makes differing assumptions about learner knowledge, the selection of one unit over another in research or pedagogy has a great impact on interpretations of the lexical challenge. It is therefore important to fully understand these assumptions so that practitioners can select the most suitable unit for a given purpose. This study introduces an enhanced version of Nation's BNC-COCA word lists that can be used to quantify several features of lexical units. The original WF6 lists were adapted by including flemma and lemma groupings, part-of-speech (POS) tags, morphological codings, frequency data, and an expanded list of proper nouns. Analyses of lexical unit composition reveal that, owing to their much greater inclusivity than flemmas or lemmas, WF6 units provide rapid corpus coverage over the 1-2k bands, and that irregular forms make up a considerable proportion of 1k tokens regardless of the unit chosen. Accuracy checks then suggest that POS-tagged lists offer an improvement over untagged lists due to the latter's overestimation of coverage and blocking of homographic concepts. Finally, lexical and morphological profiling shows that threshold coverage values are unlikely to be reached without knowledge of at

least mid-frequency headwords and tens of derivational affixes in most genres. The word list and supplementary materials are available at <https://osf.io/4mz6y/>.

Keywords: lexical units, word lists, morphology, frequency profiling

Background

Lexical units are the fundamental building block of vocabulary analysis. Put simply, these units are groups of related word forms that are considered equivalent in the mental lexicon, with demonstrated comprehension of one unit member being taken as evidence that other members are likely known as well. For example, a learner with knowledge of the form *train* (v) might also be assumed to know *trained* (v), *trainer* (n) or even *untrained* (adj) and *trainee* (n). Estimates of broader vocabulary knowledge can be obtained based on the number of known units in a range of frequency bands. Quantifying vocabulary knowledge in this manner enables straightforward interpretation of measurements and facilitates comparisons between findings. Lexical units are thus implicit features of word list development, lexical frequency profiling, testing, and pedagogy.

These units can be ordered by their inclusivity, as shown in Table 1. The least inclusive unit is the *type*, which treats different orthographic forms or POS as separate units; no assumptions are made for knowledge of morphology or grammatical function. The *lemma* assumes knowledge of inflectional affixes but distinguishes separate POS. The *flemma*, or *family-based lemma* (Nation, 2016; Pinchbeck, 2014), groups inflected forms from separate POS provided that their orthographic form matches the headword or one of its own inflections¹. Finally, *word families* are a group of lexical units that can include different POS as well as derivational affixes such as the prefix *un-* and the suffixes *-er* and *-ee*.

Bauer and Nation's (1993) word family schema has served as a template for research into lexical units. This study sets out a rationale for how affixes might be ranked based on their frequency, regularity, productivity, and predictability (see Table 2). Here, types appear at level 1, while lemmas are closest to level 2. Flemmas might be slotted in between levels 2 and 3 owing to their inclusion of other parts of speech. The word family from Table 1 comes from level 6, as it includes the suffix *-ee*. Note that irregular inflectional forms such as *took* and *taken* are not discussed in this schema but are commonly grouped with their headwords in word lists.

¹ Note that flemma-based word lists implicitly permit a small number of derived forms related to the headword. This is because certain affixes that typically act as verbal inflection markers (e.g., *-ing*, *-ed*, *-en*) can also appear on nouns or adjectives. For example, the flemma for the headword *read* would include *reading* as both a verbal inflection and a gerund. Likewise, the flemmas for *excite* and *break* would include the forms *excited* and *broken* as both verbal inflections and adjectives.

Table 1 *Lexical Units for Words Related to “Train”*

Type	Lemma		Flemma		Word Family (WF6)	
	Headword	Members	Headword	Members	Headword	Members
1. train (n)	1. train (n)		1. train (n)		1. train (n)	
2. trains (n)		trains (n)		trains (n)		trains (n)
3. train (v)	2. train (v)			train (v)		train (v)
4. trained (v)		trained (v)		trained (v)		trained (v)
5. training (v)		training (v)		training (v)		training (v)
6. trains (v)		trains (v)		trains (v)		trains (v)
7. trained (adj)	3. trained (adj)			trained (adj)		trained (adj)
8. training (n)	4. training (n)			training (n)		training (n)
9. untrained (adj)	5. untrained (adj)		2. untrained (adj)			untrained (adj)
10. trainee (n)	6. trainee (n)		3. trainee (n)			trainee (n)
11. trainees (n)		trainees (n)		trainees (n)		trainees (n)
12. trainer (n)	7. trainer (n)		4. trainer (n)			trainer (n)
13. trainers (n)		trainers (n)		trainers (n)		trainers (n)

Table 2 *Word Family Levels in Bauer and Nation (1993)*

Level	Description	Example affixes
1	Each form is a different word	/
2	Regularly inflected words are part of the same family	-s (n), -s (v), -ing, -ed, -er (adj), -est, possessive
3	Family includes the most frequent and regular derivational affixes	-able, -er, -less, -ly (adv), -ness, -th (ordinal numbers), un-
4	Family includes frequent, orthographically regular affixes	-al (adj), -ful, -ism, -ist, -ity, -ize, -ment, -ous
5	Family includes regular but infrequent affixes	-age, -dom, -en (adj), -en (v), -ese, -hood, -most
6	Family includes frequent but irregular affixes	-ee, -ic, -ify, -ive, -th (n), -y (n), pre-, re-
7	Classical roots and affixes	auto-, -arch, mono-, proto-

The degree to which meaning connections can be seen between related forms is the crux of the lexical unit issue. Undoubtedly, lexical, morphological, and encyclopaedic knowledge can all facilitate comprehension when an unfamiliar word is encountered.

Everyday discourse requires the application of such faculties whenever novel forms are encountered, as the following examples show:

- *Several years ago, shortly after Clyde Tombaugh died at the age of ninety, Pluto's planethood was thrown into question.*
- *...it's unclear whether conservatives, who generally believe that a Zuckerbergian Deep State is tasked with policing their speech on the platform... (COCA)*

The degree to which this ability is employed by language learners, however, may be more limited. Qualitative studies have revealed the challenges learners face in determining the correct meanings of forms related to known words. Ishii et al. (2021) investigated learner comprehension of 12 target-word pairs that could function as either nouns and verbs or nouns and adjectives. It was found that learners differed greatly in their ability to deduce meaning even when one member of a word pair was known. Semantic shifts across POS boundaries and degree of L1 correspondence were identified as factors that influenced comprehension. Similarly, William Grabe, while studying Portuguese in Brazil, experienced difficulty in pinpointing word meanings even when related forms were known, leading him to doubt the reliability of word families as a tool for pedagogy (Grabe & Stoller, 1997). Cross-linguistic differences notwithstanding, that a renowned linguist and leading authority on reading comprehension struggled to identify meanings even when word relationships were apparent suggests that the challenges other learners face should not be underestimated.

Although there seems to be general agreement that the choice of unit will depend on learner proficiency and the purpose of measurement (Dang et al., 2017; Nation, 2016), debates have arisen over the utility of different lexical units (e.g., Brown et al., 2022; Webb, 2021). Arguments have typically drawn on either learner knowledge or corpus coverage as evidence.

For many years, research has raised concerns that learners lack awareness of derivational affixes (Marcos Miguel, 2020; McLean, 2018; Mochizuki & Aizawa, 2000; Schmitt & Meara, 1997; Ward & Chuenjundaeng, 2009), and this appears to be the case whether or not words appear in context (Matthews et al., 2025; Milliner et al., 2024; Stoeckel et al., 2024). A smaller number of studies have examined learner comprehension of identical word forms in different parts of speech (the assumption that marks the difference between lemmas and flemmas) and found only a limited degree of success (Ishii et al., 2021; Stoeckel et al., 2020). Imperfect knowledge is thus very much the norm for findings in this branch of research, but perhaps it is also important to consider how such results should be interpreted. Laufer et al. (2021) found that CEFR-scale B1 level learners who recognized a given headword also knew one of its derived forms 60% of the time, while B2-level learners recognized a tested derivational form 84.3% of the time. These authors argue that since derivational forms typically comprise only around 3–8% of the tokens in running text, even learners with only 60% knowledge of derivational forms will suffer little loss in vocabulary coverage and are unlikely to experience any great deficit in comprehension.

The question of how impactful deficiencies in derivational knowledge are likely to be has also been investigated using corpora. Brown (2018) examined the first five

1,000-WF6 bands in Nation's (2006) British National Corpus word lists. Unlike Laufer et al. (2021), Brown's findings suggest a large drop-off in vocabulary coverage should learners lack comprehension of derivational affixes. Taking a different approach, Laufer and Cobb (2020) calculated the percentage of headwords, inflected forms, and derived forms in a range of texts and determined the number of affixes that must be known in order to reach threshold vocabulary coverage levels of 95% (for comprehension with some level of support) or 98% (for unsupported comprehension). Laufer and Cobb concluded that a limited number of derivational affixes are necessary to reach these coverage levels, ranging from zero affixes required for 95% coverage of graded readers to 12 affixes for 98% coverage of academic texts. A further approach was taken by Stoeckel and McLean (2022), who found that Laufer and Cobb's figures could only be reached if all headwords from all frequency levels were assumed to be known. They proposed an alternative method of assessing coverage that treats lexical and morphological knowledge as separate scalable constructs. This allows for various coverage estimates to be obtained for a text depending on learners' likely knowledge of headwords and affixes. For example, 95% coverage of a text might be achieved with knowledge of either 6,000 headwords and three derivational affixes or 3,000 headwords and four derivational affixes.

This paper aims to fulfil a dual-purpose. The first is to introduce an enhanced version of the BNC-COCA WF6 lists (Nation, 2017) with additional features to facilitate future studies into word properties, lexical and morphological coverage, and word lists. The second purpose is to use this list and a balanced corpus reflecting a range of genres to quantify the assumptions of lemma, flemma, and WF6 units.

As Dang and Webb (2016) point out, comparing word lists that are based on different lexical units is challenging. The more inclusive a unit, the more rapidly it will provide coverage of a text, while simultaneously making greater assumptions about word knowledge. One goal of this study is to provide a word list that can estimate coverage in lemmas, flemmas, or WF6. Since lemmas contain only forms from the same grammatical class, it is necessary to provide POS-tagging to distinguish words from different classes that share an orthographic form. As described below, POS-tags were added to word forms in WF6, flemma, and lemma units. This provided an opportunity to consider the advantages and disadvantages of POS-tagged and untagged word lists.

The treatment of proper nouns is another conundrum in lexical frequency profiling. As Nation (2016) notes, the prototypical proper noun is a personal name which individuates, rather than classifies. Other than yielding hints as to perhaps gender or background (i.e., *Margaret, Jacques*), knowledge of a personal name provides little information about its referent. Consequently, it may be assumed that proper nouns pose little difficulty for learners and the coverage they provide can be treated as additional to that of word bands ranked by frequency (Nation, 2006). Caveats are important here. Firstly, other entities may be considered as proper nouns. Nation (2016) lists locations, businesses, events, and artefacts as being less prototypical than personal names, and in these cases, knowledge of the form entails some degree of referential information about an underlying concept. Furthermore, even personal names may occasionally take on referential usage. This can occur when an individual is so prominent that further explanation beyond their name is not deemed necessary or when the

name itself becomes a byword for a particular attribute, as in the cases of Einstein with intelligence, Mozart with musical ability, and Houdini with acts of escape. Lastly, concerns have been raised over the lexical demands of proper nouns for language learners (Klassen, 2021) and their treatment in lexical coverage studies (Brown, 2010). Proper nouns may be particularly challenging during listening activities, where the lack of visual clues (i.e., capitalization) makes their recognition more problematic (Kobeleva, 2012). Furthermore, dealing with unlisted proper nouns is a common challenge of lexical frequency profiling (Nation, 2016). The original BNC-COCA list already contains 22,409 entries of both prototypical and non-prototypical proper nouns, including related derived forms in other POS. However, since they are the largest group of word types, with more than a million unique surnames in the US alone (Hanks, 2013), it was felt that there would be value in expanding the proper nouns list.

Finally, lexical units are typically defined in terms of their assumptions regarding knowledge of affixation or POS (Bauer & Nation, 1993; Nation, 2016). But as Brown (2018) notes, when one examines word lists, other assumptions become apparent. The most obvious of these are irregular forms, including standard but irregular inflections, abbreviations, and a wide range of contracted, casual, dialectal, or archaic variants. Building on Brown's work, this study will provide a detailed description of the composition of lexical units across frequency bands.

The specific research questions addressed in the study are as follows:

1. How do lemma, flemma, and WF6 units differ in size, content, and coverage?
2. How do the changes to the original lists (POS-tagging and expansion of proper nouns list) affect accuracy of measurement and coverage estimates?
3. How many and which derivational affixes are required to reach coverage thresholds in a range of genres?

Methods

Source Word Lists

The word list in this study is based on Nation's BNC-COCA WF6 list (v.2) (2017). The first ten frequency-ranked bands (10,000 WF6 units), together with the non-frequency-ranked bands containing proper nouns, compound words, acronyms, and marginal words were selected.

Reference Corpus

Frequency information for the word list is based on an approximately 111.5-million-token corpus, balanced between US and British English and across five genres: academic, fiction, magazine, news, and spoken (see Table 3). The corpus consists of data drawn from the following sources: Corpus of Contemporary American English (COCA) (49%), British National Corpus (BNC1994) (39%), British National Corpus (BNC2014) (10%), Michigan Corpus of Academic Spoken English (MICASE) (1%), and British Academic Spoken English Corpus (BASE) (1%). All data were sampled randomly from the source corpora and reflected the compositional balance of the original corpus where such information was available. Files were cleaned of corpus

annotations, markup, and extraneous symbols before being POS-tagged with the spaCy tagset (small model) using TagAnt (Anthony, 2024b). Token frequencies were obtained with AntWordProfiler (Anthony, 2024a). Full details of corpus development are available in the “Corpus Data Handling” supplementary materials.

Table 3 *Reference Corpus Token Counts by Section*

Sections by			Sub-section			Total
Variety	US	British				111,473,927
	55,789,432	55,684,495				
Genre	Academic	Fiction	Magazine	News	Spoken	111,473,927
	22,699,465	21,986,101	22,302,332	22,442,047	22,043,982	
Medium	Written	Spoken				111,473,927
	87,429,701	24,044,226				

Word List Expansion

Because the expanded word list is a complex dataset with a great many codings added, a separate supplementary file has been prepared to explain each group of codings in detail. What follows is a summary of the major expansions.

POS-Tags

POS-tags are necessary to create a lemma-based list. The spaCy tagset codings from TagAnt were manually added to list members in each lexical unit in all bands, including both frequency-ranked and non-frequency-ranked bands (see *Source Word Lists*).

Since POS-tagging can lead to inaccuracies, forms that appeared among the reference corpus off-list types in AntWordProfiler were checked to see if they had been mistagged. All types with a token frequency of at least 100 were checked and where necessary, placed into the correct lexical unit. Placement decisions were based on an inspection of concordance lines. Samples of up to 100 hits were checked in AntConc (Anthony, 2023), and forms were placed into whichever lexical unit was most frequent in the sample.

Proper Nouns

The list of proper nouns was expanded using data on commonly occurring family names from the US Census (Comenetz, 2016) and given names, locations, and organizations from the Never-Ending Language Learning Project (NELL) (Carlson et al., 2010). Plural forms were also included. These changes increased the list to 348,538 entries. As with other list members, POS-tags were added to each form, and inspections of concordance lines were used to decide on correct word placements (e.g., *frank* (adj) was placed into the *frankly* WF, while *Frank* (name) appears in the proper nouns list.

Lexical Units

Each WF6 unit in the list was sub-divided into its constituent flemmas and lemmas. Decisions were based on published sources (Bauer & Nation, 1993, 2020) and entries in the Oxford (via Google) and Merriam-Webster Dictionaries. The following forms were placed into the same lemma and flemma units as their regular headwords:

- Forms listed as contractions in dictionaries,
- Forms with irregular inflectional affixes (e.g., *beaten*),
- Forms with standard but irregular root elements (e.g., *ate*),
- Forms with standard spelling variants (e.g., *customise*, *customize*).

The following forms were placed into separate lemmas and flemmas within each WF:

- Abbreviations (e.g., *advert* and *ad* in the *advertise* WF),
- Compound words within non-compound WFs (e.g., *sometimes*, *somebody*, *someone*, *something*, and *somewhere* in the *some* WF),
- Non-standard irregular roots (forms marked as archaic, historical, informal, dialect, or specialized use in dictionaries) (e.g., *whaddya* and *whatcha* in the *what* WF, *twas* in the *be* WF).

Morphological Codings

Morphological codings were added for prefixes, suffixes, and inflections. First, the orthographic forms of words were separated into their morphological elements (see Table 4). Decisions on the presence of morphemes were based on published and online sources (Affixes.org; Bauer & Nation, 2020; Bauer et al., 2013; Etymonline.com).

Table 4 Orthographic Forms in the “Document” WF Divided into their Morphological Elements

Form	Prefix 1	Root	Suffix 1	Suffix 2	Inflection
document (n)		DOCU	MENT		
documents (n)		DOCU	MENT		S
documented (adj)		DOCU	MENT	ED	
documenting (n)		DOCU	MENT	ING	
document (v)		DOCU	MENT		
documented (v)		DOCU	MENT		ED
documenting (v)		DOCU	MENT		ING
documents (v)		DOCU	MENT		S
undocumented (adj)	UN	DOCU	MENT	ED	
documentation (n)		DOCU	MENT	ATION	

Note. Actual word list codings contain columns for four prefixes, four roots, six suffixes, and one inflection.

Each affix element was then assigned a code indicating (1) prefix, suffix, or inflection, (2) affix type, and (3) level in the Bauer and Nation (1993) affix schema (see Table 5). Cases where different affixes shared an orthographic form were distinguished with additions to the codings. For example, the form *-en* as a marker of adjectives (e.g., *wooden*) was given the code S_EN(ADJ_5), while the homographic verb-forming *-en* (e.g., *lengthen*) was coded S_EN(V_5).

As noted by Bauer and Nation (1993), the distinction between the inflectional and derivational functions of affixes such as *-ing* and *-ed* is not always easily drawn. In accord with other works on morphology (Bauer & Nation, 2020; Bauer et al., 2013), codings in this study treated these affixes on forms with verbal POS-tags as inflectional and those tagged as nouns or adjectives as derivational.

Codings were added for almost 300 affix types, including all those in the Bauer and Nation schema. The full list of affixes, together with their codings and descriptions, is available in the supplementary word list file.²

Table 5 Morphological Codings for Members of the “Document” WF

Form	Prefix 1	Root	Suffix 1	Suffix 2	Inflection
document (n)		DOCU	S_MENT(4)		
documents (n)		DOCU	S_MENT(4)		INF_NOUN_S(2)
documented (adj)		DOCU	S_MENT(4)	S_ED(2)	
documenting (n)		DOCU	S_MENT(4)	S_ING(2)	
document (v)		DOCU	S_MENT(4)		
documented (v)		DOCU	S_MENT(4)		INF_VERB_ED(2)
documenting (v)		DOCU	S_MENT(4)		INF_VERB_ING(2)
documents (v)		DOCU	S_MENT(4)		INF_VERB_S(2)
undocumented (adj)	P_UN(3)	DOCU	S_MENT(4)	S_ED(2)	
documentation (n)		DOCU	S_MENT(4)	S_ATION(4)	

Note. Morphological codings contain three parts: 1. P/S/INF = Prefix/Suffix/Inflection, 2. Affix type, 3. Number in parentheses = level in Bauer and Nation scheme.

Frequencies

The reference corpus was used to determine token frequencies for each list member and the lexical units to which it belongs in the whole corpus and each subcorpus

² Extra codings provided in the supplementary word list file indicate whether affixes are additional to the morphological structure of other lexical unit members (as with *un-*, *-ation*, *-ed*, *-ing*, and *-s* in Table 5) or whether they are etymologically present but common to all unit members (as with *-ment* in Table 5). This distinction may be useful for future studies comparing the frequency of affixes as they are conceived of in lexical units (i.e., the ‘additional’ counts) with learner exposure to affixes as they appear in word forms (i.e., both ‘additional’ and ‘etymologically present’ affixes). The RQs in the present study are all based on counts of affixes that are ‘additional’. For further details, see the supplementary materials.

(academic, fiction, magazine, news, spoken). Zipf values (van Heuven et al., 2014) were calculated to classify units into broad frequency ranges (e.g., high, medium, low).

List Organization

During list development, a number of organizational changes were made.

- The addition of POS-tags meant that some homographs could be moved to other lexical units or word bands. For example, in the original list, the forms *box* and *boxes* appear in the WF referring to a container. With POS-tags added, the verbal forms of these words could be added to the *boxing* WF. This applied to homographic word forms in both the frequency-ranked (e.g., *close*, *left*, *miss*) and non-frequency-ranked (e.g., *Sue*, *Brown*, *Baker*) lists. In all cases, inspections of concordance line samples were used to determine the best placement for a POS-tagged word form, as described in the section on POS-tags above.
- In some cases, no appropriate lexical unit existed for homographs to be moved to, so new word bands had to be created. These were as follows:
 - Reclassified words. For example, the adjective *alight* meaning ‘burning’ was moved from the verbal *alight* WF into this list as it lacked a suitable unit to join in the BNC-COCA 1-25k lists.
 - Genitives. The possessive marker *'s*, tagged S_POS, was moved to its own band.
 - Foreign words. Forms whose most frequent usage (based on a sample of concordance lines) was in non-English sentences were placed into a separate list.
- Other bands were created to accommodate words with special characteristics.
 - Offensive words. In the original lists, most offensive words appear in the marginal words list. Since these forms exhibit the properties of regular words (i.e., conceptual reference, inflection, derivation, compounding), it was decided to move them to their own band.
 - Numbers. Since some affixes can appear with digits (e.g., *1960s*, *7th*), a list of non-alphabetic numbers was added.
- Forms that contained misspellings were fixed. The Oxford and Merriam-Webster Dictionaries were consulted to confirm correct spellings.
- Apart from numbers and proper nouns, no new forms were added even if they appeared to be oversights. For example, the lists contain the forms *beautify*, *beautified*, and *beautifies*, but no entry for *beautifying*.

Irregularity Codings

Codings were added for irregular forms in the following cases:

- Contractions (e.g., *'m*, *'re*, and *'s* as contracted forms of the *be*-verb),
- Abbreviations (e.g., *doc* or *dr* as abbreviated forms of *doctor*),
- Compound words that are members of a non-compound WF (e.g., *anybody*, *anyone*, *anything*, *anyhow*, *anymore*, *anyway*, and *anywhere* as members of the *any* WF). Note that this does not include compound forms in WFs whose

- other members share the same compound elements, as is the case in the non-frequency ranked list of compound words such as *airbag* and *bathroom*,
- Forms whose root elements differ from those of the headword without being a contraction, abbreviation, or spelling variant (e.g., *spoke* in the *speak* WF),
 - Forms whose root elements match those of the headword but also contain affixes that are not mentioned in the Bauer and Nation schema (1993) and are not classical Greco-Latin affixes (e.g., *eaten* (-en), *children* (-ren)).

Results

RQ1: Content and Coverage of Lexical Units

Lexical unit sizes over the 1-10k WF6 bands are shown in Table 6. Type per unit counts reflect the inclusivity of lexical units. Since lemmas contain only inflectional affixes, there is little scope for them to change in size over the frequency range. The other units are less restricted, and the size of WF6 units, particularly in the highest frequency bands, is evident.

Table 6 Comparison of Lexical Unit Sizes in BNC-COCA 1-10k Lists

Word band	Types	WF6	Flemmas	Lemmas	Flemmas per WF6	Lemmas per WF6	Lemmas per flemma	Types per WF6	Types per flemma	Types per lemma
WF6 1k	11,451	1,000	3,428	5,961	3.43	5.96	1.74	11.45	3.34	1.92
WF6 2k	10,782	1,000	3,062	5,403	3.06	5.40	1.76	10.78	3.52	2.00
WF6 3k	9,392	1,000	2,854	4,643	2.85	4.64	1.63	9.39	3.29	2.02
WF6 4k	7,972	1,000	2,256	3,825	2.26	3.83	1.70	7.97	3.53	2.08
WF6 5k	6,993	1,000	1,960	3,307	1.96	3.31	1.69	6.99	3.57	2.11
WF6 6k	6,548	1,000	1,883	3,116	1.88	3.12	1.65	6.55	3.48	2.10
WF6 7k	5,675	999	1,754	2,731	1.76	2.73	1.56	5.68	3.24	2.08
WF6 8k	5,236	998	1,654	2,544	1.66	2.55	1.54	5.25	3.17	2.06
WF6 9k	4,734	1,000	1,568	2,343	1.57	2.34	1.49	4.73	3.02	2.02
WF6 10k	4,170	1,000	1,523	2,089	1.52	2.09	1.37	4.17	2.74	2.00

Note. Discrepancies in the 7k and 8k WF6 totals are due to three WFs being moved to the offensive words list (see supplementary materials for details).

Another way to examine the content of lexical units is to consider how morphological constituents change over the frequency range. Table 7 shows the percentage of tokens in four morphological categories over the 1-10k bands. In this analysis, forms containing both inflectional and derivational affixes were counted as derivational. Here, the unique characteristics of the 1k band can be seen, with comparatively low proportions of inflectional and derivational forms but very high rates of unaffixed regular and irregular forms. Thereafter, irregular forms comprise only a small proportion of tokens, while forms containing inflectional affixes comprise around 20-25% of tokens. Derivational affixes appear to be most prominent in the post-1k to mid-frequency range, and after this, WF6 units become less morphologically complex.

Table 7 *Percentage of Tokens in 1–10k WF6 Bands from Four Morphological Categories*

WF6 band	Regular unaffixed forms	Irregular unaffixed forms	Inflectional forms	Derivational forms
1k	75.20%	11.89%	9.48%	3.43%
2k	54.31%	0.90%	26.47%	18.32%
3k	55.31%	0.89%	24.16%	19.64%
4k	57.06%	1.39%	25.75%	15.80%
5k	60.95%	1.63%	24.84%	12.58%
6k	58.25%	0.97%	25.44%	15.34%
7k	63.55%	0.90%	23.18%	12.37%
8k	66.36%	1.19%	21.20%	11.25%
9k	63.23%	0.53%	24.54%	11.70%
10k	67.31%	1.03%	21.49%	10.17%
1–10k	71.64%	9.86%	12.44%	6.07%

Tables 8 and 9 show the same analyses for flemma and lemma units. These units make fewer assumptions regarding knowledge of derivational affixes (none in the case of lemmas), which is reflected in the increased proportions of regular unaffixed forms. Again, irregular forms are largely concentrated in the 1k bands.

Table 8 *Percentage of Tokens in 1–10k Flemma Bands from Four Morphological Categories*

Flemma band	Regular unaffixed forms	Irregular unaffixed forms	Inflectional forms	Derivational forms
1k	78.36%	10.43%	10.40%	0.80%
2k	72.83%	0.45%	24.52%	2.20%
3k	72.75%	0.43%	24.15%	2.67%
4k	75.04%	0.17%	22.60%	2.19%
5k	75.98%	0.35%	21.53%	2.15%
6k	77.78%	0.22%	20.15%	1.85%
7k	77.32%	0.11%	20.64%	1.93%
8k	77.57%	0.28%	20.43%	1.72%
9k	79.79%	0.38%	18.26%	1.57%
10k	78.61%	0.24%	18.74%	2.41%
1–10k	77.60%	8.51%	12.81%	1.07%

Table 9 *Percentage of Tokens in 1-10k Lemma Bands from Four Morphological Categories*

Lemma band	Regular unaffixed forms	Irregular unaffixed forms	Inflectional forms
1k	82.65%	7.64%	9.70%
2k	74.70%	0.53%	24.77%
3k	76.14%	0.55%	23.31%
4k	75.27%	0.35%	24.38%
5k	76.20%	0.19%	23.60%
6k	76.45%	0.28%	23.26%
7k	79.21%	0.23%	20.56%
8k	77.75%	0.16%	22.09%
9k	76.02%	0.35%	23.63%
10k	78.79%	0.29%	20.92%
1-10k	81.21%	6.13%	12.66%

A closer examination of irregular forms reinforces the view that irregularity is a significant component of mastering the WF6 1k band (see Table 10). In particular, contractions and forms containing irregular affixes or roots comprise a large share of all 1k tokens. Such forms are typically addressed in the early stages of language instruction, but the coverage figures here make clear how detrimental to comprehension gaps in knowledge are likely to be.

Table 10 *Irregular Forms as a Percentage of All Tokens in WF6 Units in BNC-COCA 1-10k Lists*

Word band	Contractions	Abbreviations	Compound words in WF	Irregular affixes or roots	Total (% of all tokens)
WF6 1k	2.39%	0.25%	0.69%	13.27%	16.59%
WF6 2k	0.00%	0.77%	0.01%	0.46%	1.24%
WF6 3k	0.00%	0.26%	0.15%	0.66%	1.08%
WF6 4k	0.00%	0.94%	0.02%	0.89%	1.85%
WF6 5k	0.35%	1.06%	0.00%	0.95%	2.36%
WF6 6k	0.00%	0.86%	0.10%	0.41%	1.36%
WF6 7k	0.00%	0.58%	0.00%	0.60%	1.18%
WF6 8k	0.00%	0.22%	0.65%	0.85%	1.72%
WF6 9k	0.00%	0.05%	0.00%	0.72%	0.77%
WF6 10k	0.00%	0.59%	0.00%	0.73%	1.32%
WF6 1-10k	1.95%	0.32%	0.57%	10.91%	13.75%

Note. Some irregular forms (e.g., the abbreviation *advert*s or the adjective *spoken*) contain inflectional or derivational affixes. In Table 7, these were counted as inflectional or derivational forms. Thus, the percentages of irregular unaffixed forms in Table 7 do not match the row totals in this table.

In order to compare WF6, flemmas, and lemmas in terms of text coverage, it was necessary to re-rank all units. This was to ensure a fair comparison; the flemma and lemma units could only be ranked using the frequency data from this study, but the original WF6 bands had been ranked using Nation's own criteria and corpus frequencies. Furthermore, some of the non-frequency-ranked bands (compound, offensive, and reclassified words) contain forms that refer to concepts, and it has been argued that knowledge of proper nouns cannot always be assumed. Therefore, it was decided to re-rank lexical units from these bands as well as the frequency-ranked bands for this analysis. All lexical units (WF6, flemma, and lemma) belonging to the frequency-ranked, proper noun, compound word, offensive word, and reclassified word lists were re-ranked so that their coverage could be compared using frequencies from the same corpus. These rankings were then divided into 20 new 1,000-unit bands for each lexical unit.

Figure 1 shows the coverage comparison for the re-ranked WF6, flemma, and lemma units across the whole corpus, and Table 11 presents the numbers of lexical units required to reach coverage threshold values of 95% and 98% in the whole corpus and each subcorpus. As a rule of thumb, the number of flemma bands required for 95% coverage is about 1.5 times that of WF6 bands, while for lemmas, it is around double the number of WF6 bands.

A notable feature of Figure 1 is how rapidly the point is reached at which flemma and lemma units begin to narrow the gap to WF6. Over the first two bands, WF6

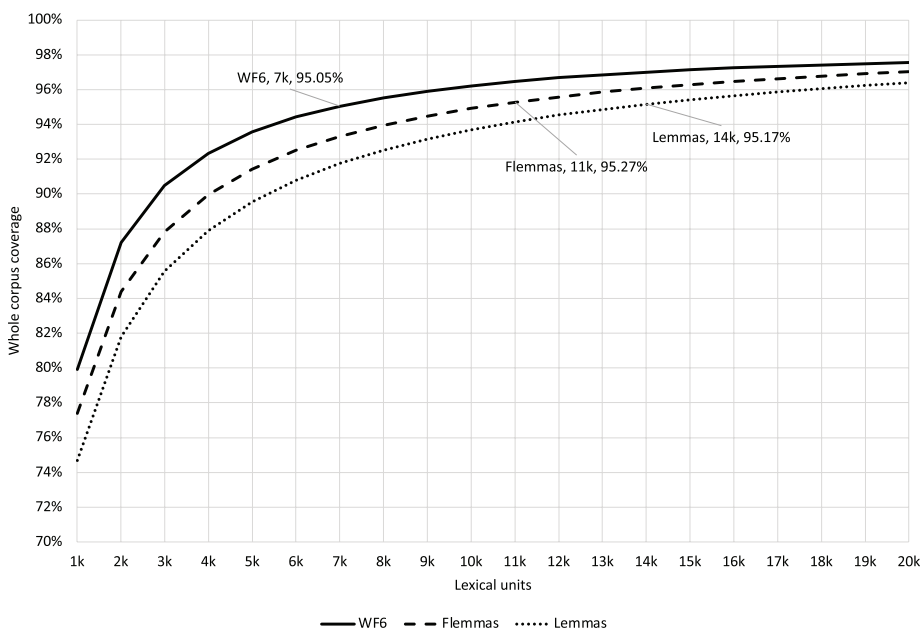


Figure 1 *Cumulative Coverage for Lexical Units.*

Note. Coverage from marginal words, acronyms, foreign words, genitives, and numbers has been added to the coverage values for each lexical unit.

coverage exceeds that of the other units, but from 3k on, flemma bands contain more tokens than WF6 bands, and lemma band coverage also outpaces that of flemmas. Clearly, this is due to the relative sizes of the units; as Table 6 shows, in the 1-2k range, WF6 units are typically over three times the size of flemmas and over five times larger than lemmas. This inclusivity exhausts the number of high frequency types more rapidly in WF6, whereas in the smaller units, high- and medium-frequency words are spread over a greater number of bands.

Table 11 *Lexical Unit Bands Required to Reach Threshold Coverage Values*

Genre	WF6		Flemma		Lemma	
	95%	98%	95%	98%	95%	98%
Whole	7k	/	11k	/	14k	/
Academic	10k	/	15k	/	20k	/
Fiction	7k	19k	10k	/	14k	/
Magazine	9k	/	13k	/	17k	/
News	9k	/	13k	/	16k	/
Spoken	3k	8k	4k	11k	5k	14k

Note. / = Coverage threshold cannot be reached with 20k lexical units.

RQ2: Accuracy and Coverage of POS-Tagged and Untagged Word Lists

Lemma analysis requires the addition of POS-tags to a word list. This raises the question of accuracy since tagging is dependent on the reliability of software algorithms. POS-tagging accuracy was investigated using three approaches.

First, 50 100-token samples of text were randomly selected from the corpus and their POS-tags coded for accuracy. These samples were balanced to reflect the composition of the whole corpus in terms of genre, language variety, and medium. The results are given in Table 12. In the initial (strict) analysis, there was noticeable variation in results across corpus genres and between written and spoken data. Accuracies were also in most cases lower than the reported figure of 97% (<https://spacy.io/usage/facts-figures>). Data inspections revealed that much of this inaccuracy was due proper noun tags being assigned to capitalized non-proper nouns. Therefore, it was decided to also apply a more lenient standard in which regular nouns that had been tagged as proper nouns would be coded as correct. In this second analysis, tagging accuracy became more evenly balanced across genres, varieties, and media.

Accuracy data were then grouped into individual POS categories. The accuracy for verbs was noticeably lower than that of other content word classes (see Table 13). Accordingly, the codings for verbs were then broken down further into separate POS-tags (see Table 14). Here, it was evident that the main causes of inaccuracy in the tagging of verbs were the present (VBG) and past (VBN) participle forms, which were sometimes applied to gerunds or participle adjectives.

Table 12 *Accuracy of POS-Tagging by Corpus Section*

Genre	Tokens	Correct (strict)	Accuracy (strict)	Correct (lenient)	Accuracy (lenient)
ACAD	1,038	1,002	96.53%	1,009	97.21%
FIC	1,006	985	97.91%	992	98.61%
MAG	1,008	964	95.63%	988	98.02%
NEWS	960	898	93.54%	941	98.02%
SPOK	988	959	97.06%	967	97.87%
Whole	5,000	4,808	96.16%	4,897	97.94%
Variety	Tokens	Correct (strict)	Accuracy (strict)	Correct (lenient)	Accuracy (lenient)
BR	2,589	2,488	96.10%	2,536	97.95%
US	2,411	2,320	96.23%	2,361	97.93%
Medium	Tokens	Correct (strict)	Accuracy (strict)	Correct (lenient)	Accuracy (lenient)
WRITTEN	3,829	3,672	95.90%	3,752	97.99%
SPOKEN	1,171	1,136	97.01%	1,145	97.78%

Table 13 *Tagging Accuracy by Part of Speech*

Broad POS	Accuracy
Conjunction	98.80%
Number	100.00%
Determiner	99.82%
Existential there	100.00%
Foreign word	100.00%
Preposition	99.23%
Adjective	97.71%
Modal	100.00%
Noun	97.66%
Proper noun	70.15% / 96.72%
Possessive	89.74%
Pronoun	100.00%
Adverb	98.17%
Particle	100.00%
Infinitive	97.30%
Interjection	96.88%
Verb	94.26%

Note. The two values for proper nouns are under strict and lenient conditions.

Table 14 *Tagging Accuracy for Verb Forms*

Verb POS-tag	Accuracy
VB	96.30%
VBD	99.54%
VBG	66.33%
VCN	95.65%
VBP	96.72%
VBZ	98.52%

The above analysis followed a standard approach for measuring tagging accuracy (Salton, 1989), but it may give an unduly positive impression. This is because, in any contextualized sample, a significant proportion of tokens will be function words that have only one possible POS-tag. In a study aimed at lexical issues, there is value in focussing on the tagging accuracy of content words. Therefore, in the second analysis of POS-tagging accuracy, 10% of all flemmas containing both nouns and verbs were sampled from the word list. For each flemma, up to 100 concordance lines in each POS were coded for accuracy.

The values in Table 15 indicate that while high frequency forms are tagged with accuracy rates close to or above reported values, those at lower frequencies, particularly verbs, are more likely to be mistagged. This was again due to gerunds and participle adjectives being mistaken for verbal forms.

Table 15 *Accuracy of POS-Tagging for Nouns and Verbs Grouped by Zipf Value*

Nouns (Zipf value)	Frequency in sample	Coded correctly	Accuracy
High/very high frequency	10,200	10,034	98.37%
Medium frequency	13,400	12,691	94.71%
Low/very low frequency	4,291	3,921	91.38%
All nouns	27,891	26,646	95.54%
Verbs (Zipf value)	Frequency in sample	Coded correctly	Accuracy
High/very high frequency	5,800	5,587	96.33%
Medium frequency	13,200	12,252	92.82%
Low/very low frequency	6,499	5,484	84.38%
All verbs	25,499	23,323	91.47%

The final analysis of POS-tag accuracy examined the possibility that tagging might lead to forms incorrectly being counted as off-list items and therefore underestimating coverage. As mentioned in the methodology, off-list types with a frequency of at least 100 tokens had already been checked for their correct placement. The orthographic forms of all remaining off-list types were compared to forms in the original untagged version of the BNC-COCA lists. The bands to which these forms potentially belonged could then be identified. The results appear in Table 16, where it can be seen that approximately 91% of off-list tokens would definitely not belong on the word list as they were either potential members of frequency bands beyond 10k or not present on the lists at all. The remaining 8.66% of tokens shared an orthographic form with a member of the lists used in this study. If all of these tokens were found to belong to the list, they would increase corpus coverage by 0.15%. It should be noted, however, that this is an unlikely scenario; many of these tokens are homographs of list members. For example, in the spoken section of the corpus, there are many instances where speakers pause mid-sentence. This leads to data such as the following:

- ...*the learning curve is st way too steep...* (BNC2014 Spoken)

The form 'st' appears in the 1k band as the abbreviated form of *street*. Thus, placing all forms such as these into the list would be incorrect.

Table 16 *Potential Placement of Off-List Forms*

Off-list word potential placement	% Off-list tokens	Coverage increase
Current 1-10k bands	8.66%	0.15%
Frequency bands beyond 10k	33.90%	/
Not present on BNC-COCA lists	57.44%	/

Accuracy in untagged word lists was then investigated. In an untagged word list, each orthographic form only appears in one place. Homographic forms therefore represent a source of error since some proportion of the tokens for each homograph refer to a different concept than that of the unit it has been placed into. With POS-tags added, many homographic forms that could only be placed into a single unit in the untagged lists could now be moved to a different lexical unit. Quantifying this movement of tokens allows for an estimate of accuracy in untagged lists to be obtained (see Table 17).

In total, 2,488,711 tokens, 2.23% of the entire corpus, had to be moved to other word families as a result of POS-tags being added. In other words, the accuracy of the untagged lists could be taken as approximately 97.77%, a value that falls between the strict and lenient measures of accuracy across the whole corpus for the POS-tagged lists given in Table 12.

As Table 17 shows, in the untagged lists, almost 69% of the moved tokens shifted from one frequency-ranked band into a lower frequency band (e.g., moved from 1k to 2k or lower-frequency bands, including off-list words) or into one of the non-frequency-ranked bands (e.g., proper nouns or compound words). Just 6.09% of tokens

were moved into higher frequency bands. This suggests that overall, untagged word lists inflate coverage estimates by including in higher-frequency bands many tokens that actually refer to lower-frequency concepts.

Table 17 *Movement of Homographs Due to Addition of POS-Tags*

Movement	% Moved tokens	% of corpus
1-10k to lower frequency band or off-list	23.94%	0.53%
1-10k to non-frequency band	44.89%	1.00%
1-10k to higher frequency band	6.09%	0.14%
Moved within frequency band	16.05%	0.36%
Non-frequency band to non-frequency ranked band	9.02%	0.20%
Total	100.00%	2.23%

The effects of adding POS-tags can be seen in Figure 2, which plots coverage values for the whole corpus by the POS-tagged lists from this study and the untagged original lists, both with and without coverage from the non-frequency-ranked bands. The addition of POS-tags reduces coverage by approximately 1% across the frequency bands. As Table 17 shows, much of this can be explained by forms being moved to lower-frequency or non-frequency-ranked bands. However, owing to the expansion of the proper nouns list, when non-frequency-ranked band coverage is added, the POS-tagged list coverage exceeds that of the original lists by approximately 0.5% across the frequency bands.

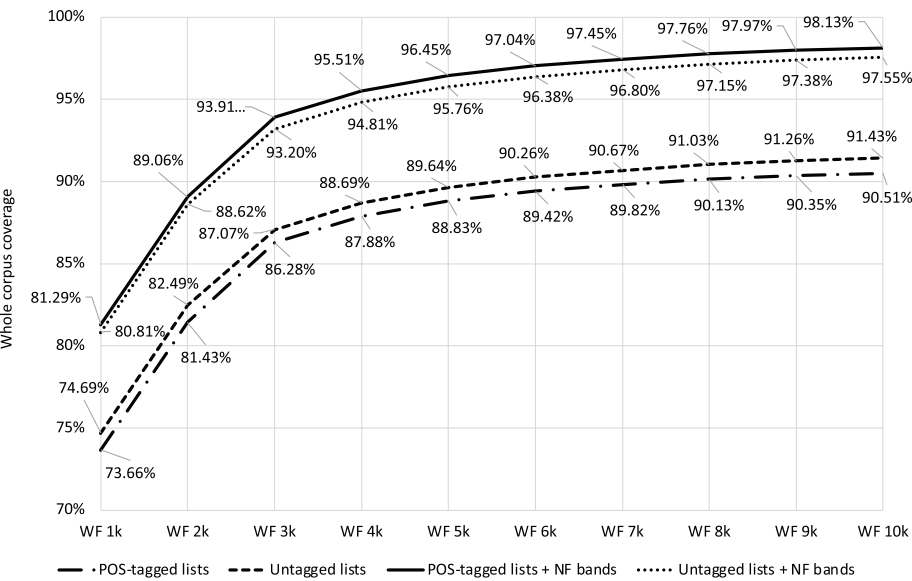


Figure 2 *Cumulative Coverage from POS-Tagged and Untagged Lists With and Without Non-Frequency Ranked (NF) Bands.*

RQ3: The Contribution of Derivational Affixes to Corpus and Subcorpus Coverage

Stoeckel and McLean’s (2022) approach was used to measure the lexical and morphological components of coverage. This approach measures coverage along two scales: frequency bands (rows) and morphological knowledge (columns) (see Table 18). Frequency bands follow the ordering of the word lists (i.e., 1k precedes 2k), and affixes are ordered with inflections preceding derivations and then by their token frequency in the corpus. In the table, all inflections have been collapsed into a single category.

The table shows a portion of the results for the Fiction subcorpus. Values are cumulative both vertically and horizontally; the coverage figure at a particular point assumes comprehension of all unaffixed forms in that row and those above as well as all morphological categories in that column and those to the left. Coverage values of 95% or higher are indicated in bold. It is possible to reach 95% coverage with knowledge of only a single derivational affix (adverb-forming *-ly*) provided that a minimum of 8,000 unaffixed or inflected forms are known. Alternatively, knowledge of two derivational affixes (*-ly* and gerund or participle adjective-forming *-ing*) means that only 7,000 unaffixed or inflected forms are required to reach 95%. Similarly, the coverage threshold can be reached with 5,000 or 6,000 unaffixed/inflected forms and seven or three derivational affixes respectively.

Table 18 *Lexical and Morphological Coverage of the Fiction Subcorpus*

Word Band	Regular unaffixed	Irregular unaffixed	Inflections	S_LY(3)	S_ING(2)	S_ED(2)	S_ER(3)	P_UN(3)	S_Y(3)	S_ACTION(4)	S_AL(4)
WF6 1k	63.46%	74.80%	83.59%	84.07%	84.37%	84.50%	84.66%	84.73%	84.75%	84.80%	84.84%
WF6 2k	66.51%	77.91%	88.35%	88.97%	89.36%	89.56%	89.78%	89.88%	89.96%	90.03%	90.10%
WF6 3k	67.98%	79.39%	90.49%	91.18%	91.60%	91.82%	92.07%	92.18%	92.28%	92.36%	92.44%
WF6 4k	68.87%	80.30%	91.96%	92.69%	93.14%	93.39%	93.66%	93.78%	93.90%	93.98%	94.07%
WF6 5k	69.50%	80.95%	92.95%	93.70%	94.16%	94.42%	94.70%	94.83%	94.96%	95.04%	95.14%
WF6 6k	69.86%	81.32%	93.55%	94.31%	94.79%	95.06%	95.35%	95.48%	95.61%	95.70%	95.80%
WF6 7k	70.12%	81.58%	93.93%	94.71%	95.20%	95.48%	95.76%	95.90%	96.03%	96.12%	96.22%
WF6 8k	70.32%	81.78%	94.23%	95.01%	95.51%	95.79%	96.08%	96.23%	96.36%	96.45%	96.55%
WF6 9k	70.48%	81.94%	94.47%	95.26%	95.76%	96.05%	96.34%	96.49%	96.62%	96.71%	96.81%
WF6 10k	70.61%	82.07%	94.66%	95.45%	95.95%	96.24%	96.53%	96.68%	96.81%	96.91%	97.00%

Note. Non-frequency-ranked band coverage has been added to all values.

Table 19 provides a summary of this analysis. Here, the rows indicate the number of unaffixed (both regular and irregular) and inflectional forms known, and the columns refer to the number of derivational affixes required to reach threshold coverage values of 95% or 98% in different subcorpora. Unsurprisingly, spoken discourse requires the fewest derivational affixes, while academic vocabulary is the most morphologically complex. While it is true that 95% coverage can be achieved with a small number of derivational affixes in most genres, this is only possible when several thousands of unaffixed words are known.

Clearly, across all genres, the step up from 95% to 98% coverage is a significant hurdle, requiring gains of at least 2,000—and typically more—extra unaffixed forms and tens of derivational affixes. Thus, fluent unaided comprehension of these genres requires vocabularies that stretch well into the mid-frequency range accompanied by knowledge of many tens of derivational affixes.

Table 19 *Derivational Affixes Required to Reach Coverage Thresholds in a Range of Genres for Given Levels of Unaffixed and Inflected Form Knowledge*

Unaffixed + inflected forms known	Derivational affixes required to reach coverage thresholds											
	Whole corpus		Academic subcorpus		Fiction subcorpus		Magazine subcorpus		News subcorpus		Spoken subcorpus	
	95%	98%	95%	98%	95%	98%	95%	98%	95%	98%	95%	98%
1k	/	/	/	/	/	/	/	/	/	/	/	/
2k	/	/	/	/	/	/	/	/	/	/	/	/
3k	/	/	/	/	/	/	/	/	/	/	1	/
4k	32	/	/	/	25	/	/	/	29	/	0	/
5k	13	/	50	/	7	/	31	/	13	/	0	23
6k	9	/	27	/	3	/	16	/	9	/	0	15
7k	7	/	20	/	2	/	12	/	8	/	0	12
8k	6	/	17	/	1	60	9	/	6	/	0	10
9k	6	/	15	/	1	33	8	/	6	61	0	9
10k	5	54	14	/	1	25	8	/	5	44	0	8

Note. Unaffixed includes both regular and irregular forms.

Note. / = Threshold coverage not possible.

Table 20 *Most Frequent Affixes Across Genres*

Rank	Academic		Fiction		Magazine		News		Spoken	
	Affix	Coverage	Affix	Coverage	Affix	Coverage	Affix	Coverage	Affix	Coverage
1	S_ATION(4)	1.10%	S_LY(3)	0.94%	S_LY(3)	0.97%	S_ER(3)	0.91%	S_LY(3)	0.68%
2	S_LY(3)	1.08%	S_ING(2)	0.52%	S_ING(2)	0.85%	S_ING(2)	0.75%	S_ING(2)	0.39%
3	S_ING(2)	1.00%	S_ED(2)	0.33%	S_ER(3)	0.70%	S_LY(3)	0.73%	S_ER(3)	0.28%
4	S_AL(4)	0.85%	S_ER(3)	0.29%	S_ATION(4)	0.49%	S_ATION(4)	0.43%	S_ED(2)	0.17%
5	S_ER(3)	0.65%	S_Y(3)	0.15%	S_AL(4)	0.46%	S_AL(4)	0.42%	S_ATION(4)	0.14%
6	S_ION(6)	0.64%	S_ATION(4)	0.15%	S_ED(2)	0.35%	S_MENT(4)	0.33%	S_AL(4)	0.13%
7	S_MENT(4)	0.56%	P_UN(3)	0.12%	S_MENT(4)	0.31%	S_ED(2)	0.32%	S_AN(5)	0.12%
8	S_ITY(4)	0.45%	S_AL(4)	0.12%	S_ION(6)	0.29%	S_ION(6)	0.27%	S_Y(3)	0.11%
9	S_ED(2)	0.35%	S_FUL(4)	0.11%	S_AN(5)	0.21%	S_AN(5)	0.24%	S_MENT(4)	0.10%
10	S_IC(6)	0.32%	S_ION(6)	0.11%	S_ITY(4)	0.20%	S_ITY(4)	0.18%	S_ION(6)	0.09%
Subcorpus coverage		6.99%		2.86%		4.82%		4.58%		2.20%
%derivational tokens		78.02%		74.06%		74.14%		74.49%		73.32%

While this may suggest that learner comprehension of unmodified discourse is a daunting challenge, an examination of the contributions of individual affixes towards coverage offers some hope. Table 20 lists the ten most frequent derivational affixes by token count in each genre. Their contributions to subcorpus coverage vary, but in each genre, the ten most frequent derivational affixes account for around 75% of all derivational tokens. There is also considerable overlap in the affixes themselves, with only 14 unique types represented, seven of which appear in all five genres.

Discussion

The first analysis highlighted the differences in size and composition of three lexical units. WF6 units are much more inclusive than flemmas or lemmas, particularly in the highest frequency bands. Irrespective of the lexical unit though, irregular tokens, especially standard forms with non-regular affixes or roots, are an important component of high-frequency vocabulary, comprising 8-12% of tokens at the 1k level. Derivational affixes, meanwhile, are a key element of mastering the high- to mid-frequency range in WF6 units. In the lower-frequency bands, the three lexical units become more similar, although WF6 is still around double the size of a lemma even in the 10k band.

Regarding coverage, a ratio of 1:1.5:2 seems to be a reasonable approximation of the number of bands required to achieve 95% coverage with WF6, flemmas, or lemmas. The results in Figure 1 also have important implications for the interpretation and application of vocabulary tests. For example, a test might indicate that learners know around 4,000 WF6 units. If those learners in fact lack the knowledge required to make connections between tested word forms and other WF6 members, then estimates of their ability to comprehend certain materials will be inaccurate. If flemmas or lemmas were a more appropriate unit for the learners' level of knowledge, then an estimated lexicon of 2,000-3,000 WF6 units is probably more reflective of their ability.

Adding POS-tags has a number of knock-on effects for word lists. Distinguishing word classes is essential for lemma-based measurement, but beyond this, POS-tagging also offers a finer-grained level of analysis that has implications regarding list organization and accuracy. In general terms, the accuracy of tagged and untagged lists is comparable. Across the whole corpus, POS-tagging was between 96.16% and 97.94% accurate, while the movement of list members prompted by the addition of POS-tags implied that untagged lists are roughly 97.77% accurate.

Caveats are important in both cases. The accuracy of POS-tagging was notably lower for proper nouns and participle forms, especially among lower frequency words. This suggests that manual data treatment such as concordance line checking is necessary to ensure that a POS-tagged list is as accurate as possible. Nevertheless, most POS-tagging errors are only at a lemma level. For untagged lists, meanwhile, inaccuracies due to homography are more consequential. These cases involve a shift between word families and therefore a change in concept. For example, the orthographic form *kind* appears in the untagged lists as a noun (i.e., *a kind of...*) because this form is more frequent than the adjective *kind* (as in *kindness*). Therefore, any instances of the form *kind* in a text that actually refer to kindness are being classified as a different concept. In other words, less frequent members of homographic pairs are blocked from appearing in the list. In an untagged list, concepts such as *bear* (animal), *minute* (tiny), *mean*

(average), *pop* (music), *row* (line), and *bow* (knot), are all blocked by more frequent homographs.

Finally, derivational forms were shown to be an important element of coverage. In most genres, less than 5% of tokens were derivational, but apart from spoken discourse, learners are unlikely to reach 95% coverage thresholds without knowledge of 4,000–5,000 unaffixed forms and 20–30 derivational affixes. More optimistically, the study found that approximately 75% of derivational tokens were accounted for by 10 affixes in each genre, which suggests that learners could benefit from focussed instruction in this area.

Conclusion

This study used an expanded version of Nation's (2017) BNC-COCA WF6 lists to quantify the assumptions of three lexical units: WF6, flemmas, and lemmas. The much greater inclusivity of WF6 units accounts for their rapid corpus coverage across the 1–2k frequency bands, but thereafter, coverage from flemma and lemma units outpaces that of WF6. In WF6, almost all high-frequency word types are covered by the first 2,000 units, while for the smaller units, they are spread over a wider range of bands. The three lexical units make similar assumptions regarding knowledge of irregular forms, which comprise 8–12% of 1k tokens in all cases. An analysis of how accurately untagged and POS-tagged lists place forms into lexical units suggested that in they are broadly comparable. POS-tagged word lists require manual identification of erroneous tags in order to ensure high levels of accuracy, but they are also able to perform tasks that are not possible with untagged lists. These include the ability to analyse texts using lemmas and the distinction of many, but not all, homographic forms. Finally, it was shown that while derivational affixes make up a relatively small proportion of tokens, coverage threshold values for written discourse are unlikely to be reached without knowledge of 20–30 affixes and of headwords to at least mid-frequency level.

One important limitation is that the flemma and lemma lists used in the study were limited to only those units present in the 1–10k and non-frequency-ranked (proper noun, compound word, marginal word, and acronym) original WF6 bands. There are undoubtedly some flemmas and lemmas from the 11k WF6 band or beyond with frequencies greater than those whose presence on the list is due to a more frequent WF6 member. While it is not expected that this will have a dramatic impact on results, coverage values for flemmas and lemmas are being underestimated here. Another issue is the representativeness of the corpus used in the study. Since the original word lists were based on data from the BNC and COCA, there is a close match with the corpus used in this study, for which 98% of the data were derived from the same corpora. However, studies have suggested that word frequency rankings from other genres may be more closely related to learners' word knowledge (Pinchbeck et al., 2022). Certainly, a useful future development would be lists of WF6, flemma, and lemma units, each constructed independently and based on a broad balanced corpus.

Despite these limitations, it is hoped that the word list developed for this study will be of use in future explorations of lexical units. The enhanced 1–10k BNC-COCA word lists are available at <https://osf.io/4mz6y/>. One future project that is already in development is a profiler that will enable analyses of texts for their lexical

and morphological content. There remains much to explore regarding how learners incorporate awareness of morphology into their developing lexical competence, and the codings provided in the list should allow for further analyses to be conducted.

References

- Anthony, L. (2023). *AntConc* (Version 4.2.4) [Computer Software]. Waseda University. <https://www.laurenceanthony.net/software/AntConc>
- Anthony, L. (2024a). *AntWordProfiler* (Version 2.2.1) [Computer Software]. Waseda University. <https://www.laurenceanthony.net/software/AntWordProfiler>
- Anthony, L. (2024b). *TagAnt* (Version 2.1.1) [Computer Software]. Waseda University. <https://www.laurenceanthony.net/software/TagAnt>
- Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford reference guide to English morphology*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198747062.001.0001>
- Bauer, L. & Nation, P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253–279. <https://doi.org/10.1093/ijl/6.4.253>
- Bauer, L. & Nation, I. S. P. (2020). *English morphology for the language teaching profession*. Routledge. <https://doi.org/10.4324/9780367855222>
- Brown, D. (2010). An improper assumption? The treatment of proper nouns in text coverage studies. *Reading in a Foreign Language*, 22(2), 355–361. <https://doi.org/10.64152/10125/66842>
- Brown, D. (2018). Examining the word family through word lists. *Vocabulary Learning and Instruction*, 7(1), 51–65. <https://doi.org/10.7820/vli.v07.1.brown>
- Brown, D., Stoeckel, T., McLean, S., & Stewart, J. (2022). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, 43(3), 596–602. <https://doi.org/10.1093/applin/amaa061>
- Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr., E. R., & Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, (pp. 1306–1313). <https://doi.org/10.1609/aaai.v24i1.7519>
- Comenetz, J. (2016). *Frequently occurring surnames in the 2010 census*. U.S. Census Bureau. <https://www2.census.gov/topics/genealogy/2010surnames/surnames.pdf>
- Dang, T. N. Y. & Webb, S. (2016). Evaluating lists of high-frequency words. *International Journal of Applied Linguistics*, 167(2), 132–158. <https://doi.org/10.1075/itl.167.2.02dan>
- Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The Academic Spoken Word List. *Language Learning*, 67(4), 959–997. <https://doi.org/10.1111/lang.12253>
- Grabe, W. & Stoller, F. L. (1997). Reading and vocabulary development in a second language: A case study. In J. Coady & T. Huckin (Eds.) *Second language vocabulary acquisition: A rationale for pedagogy*. Cambridge University Press.
- Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. MIT Press.
- Ishii, T., Bennett, P., & Stoeckel, T. (2021). Challenges in the assumptions of using a flemma-based word-counting unit. *Vocabulary Learning and Instruction*, 10(1), 1–15. <https://doi.org/10.7820/vli.v10.1.Ishii>

- Klassen, K. (2021). Proper name theory and implications for second language reading. *Language Teaching*, 55(2), 149–155. <https://doi.org/10.1017/S026144482100015X>
- Laufer, B. & Cobb, T. (2020). How much knowledge of derived words is needed for reading? *Applied Linguistics*, 41(6), 971–998. <https://doi.org/10.1093/applin/amz051>
- Laufer, B., Webb, S., Kim, S. K., & Yohanan, B. (2021). How well do learners know derived words in a second language? The effect of proficiency, word frequency and type of affix. *International Journal of Applied Linguistics*, 172(2), 229–258. <https://doi.org/10.1075/itl.20020.lau>
- Marcos Miguel, N. (2020). Analyzing morphology-related strategies in Spanish L2 lexical inferencing: How do suffixes matter? *International Review of Applied Linguistics in Language Teaching*, 58(3), 351–377. <https://doi.org/10.1515/iral-2016-0091>
- Matthews, J., Milliner, B., & McLean, S. (2025). Can learners understand words with derivational affixes and does presence of context make a difference? *RELC Journal*, 56(2), 273–286. <https://doi.org/10.1177/00336882231222034>
- McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, 39(6), 823–845. <https://doi.org/10.1093/applin/amw050>
- Milliner, B., Lange, K., Matthews, J., & Umeki, R. (2024). Examining EFL learners' comprehension of derivational forms: The role of overlap with base word knowledge, word frequency, and contextual support. *Language Teaching Research*. Advance online publication. <https://doi.org/10.1177/13621688231225704>
- Mochizuki, M. & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System*, 28(2), 291–304. [https://doi.org/10.1016/S0346-251X\(00\)00013-0](https://doi.org/10.1016/S0346-251X(00)00013-0)
- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1), 59–81. <https://doi.org/10.3138/cmlr.63.1.59>
- Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company. <https://doi.org/10.1075/z.208>
- Nation, I. S. P. (2017). *The BNC/COCA Level 6 word family lists* (Version 2.0.0) [Data file].
- Pinchbeck, G. G. (2014). *Lexical frequency profiling of a large sample of Canadian high school diploma exam expository writing: L1 and L2 academic English*. Roundtable presentation at American Association of Applied Linguistics. Portland, OR, USA.
- Pinchbeck, G. G., Brown, D., McLean, S., & Kramer, B. (2022). Validating word lists that represent learner knowledge in EFL contexts: The impact of the definition of word and the choice of source corpora. *System*, 106, Article 102771. <https://doi.org/10.1016/j.system.2022.102771>
- Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Addison-Wesley.
- Schmitt, N. & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition* 19(1), 17–36. <https://doi.org/10.1017/S0272263197001022>

- Stoeckel, T., Ishii, T., & Bennett, P. (2020). Is the lemma more appropriate than the flemma as a word-counting unit? *Applied Linguistics*, 41(4), 601–606. <https://doi.org/10.1093/applin/amy059>
- Stoeckel, T., & McLean, S. (2022). The case for combining lexical and morphological text profiling: A response to Cobb. *Reading in a Foreign Language* 34(1), 172–183. <http://hdl.handle.net/10125/67418>
- Stoeckel, T., McLean, S., Kim, Y. A., Shinhara, Y. & Spencer Patterson, S. (2024). The role of context in the word family debate. *System*, 126, Article 103453. <https://doi.org/10.1016/j.system.2024.103453>
- van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67(6), 1176–1190. <https://doi.org/10.1080/17470218.2013.850521>
- Ward, J. & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System*, 37(3), 461–469. <https://doi.org/10.1016/j.system.2009.01.004>
- Webb, S. (2021). Word families and lemmas, not a real dilemma: Investigating lexical units. *Studies in Second Language Acquisition*, 43(5), 973–984. <https://doi.org/10.1017/S0272263121000760>