「論文」

Measuring Similarities Within Word Families: A Word-embedding Approach Using word2vec

Satoru UCHIDA and Mitsuhiro MORITA

Abstract

The word family is a useful concept to determine the lexical aspects of English learners and has been widely used in vocabulary studies. However, it has been criticized, especially because elementary foreign language learners do not have a full command of its derivational operations. In addition, it remains unclear as to which member in a word family is challenging for the learners. This study examines the similarities between each member of the word family by using word2vec, a widely used natural language processing application. Based on the similarity scores between the word forms generated by the application using 7,540 pairs of words created from the CEFR-J wordlist and BNC/COCA family lists, this study argues that teachers and learners must especially focus on word families with low similarity scores. Furthermore, these results are useful for determining the difficulty level of affixes and discovering specific word forms that require special treatment in the classroom.

1. Introduction

A word family is a group of words with a word base or stem. For example, “kindly,” “kindness,” and “unkind” are grouped into one word family with “kind” as the common base. The concept of word family is important in English vocabulary studies, especially with regard to vocabulary assessment and text coverage.

In the context of English as a foreign language (EFL) classrooms, it should be considered that the members in a word family do not always exhibit the same difficulty level. For instance, it would be easy for several learners to infer the meaning of “flatten” from the meaning of “flat” with the knowledge of “en” as a verb suffix, but linking the meaning of “flat” and “flatly” would be more challenging, especially for a

novice learner who has just learned such phrases as “a flat stone.” This implies that teachers and learners must be aware that each member in a word family has an individuality. In fact, some studies have revealed the issues with using word families as the measuring unit for assessing the learners’ vocabulary size (Gardner, 2007; Kremmel, 2016). However, no attempts have been made to identify which words in each word family actually cause problems for learners.

The present study aims to evaluate the similarities among each member in a word family using word2vec (Mikolov et al., 2013), a powerful and influential application in natural language processing (NLP) that enables the assignment of numbers (vectors) to words; these can then be observed as a representation of word meanings. Subsequently, we can calculate how closely each word is related using cosine similarity scores. The current study hypothesizes that this score can be used to measure the relatedness and learnability of words within a word family, which eventually reveals peculiar members that need special attention in English education.

2. Literature Review

2.1 Word family and vocabulary learning

Word families have been used as counting units in vocabulary research, especially vocabulary knowledge assessment and text coverage studies. Popular vocabulary assessment measures have adapted word family counts, such as the Vocabulary Levels Test (Nation, 1983) and the Vocabulary Size Test (Nation & Beglar, 2007), among others. Nation’s (2006) influential text coverage study indicated that for 98% coverage, the most frequent 8,000 to 9,000 word families were necessary for written discourse, and the most frequent 6,000 to 7,000 word families were essential for spoken discourse. Nurmukhamedov and Webb (2019) reported that many text coverage studies adapted word families as counting units owing to the development of corpus analysis tools, which enabled researchers to create corpus-based word lists, such as the BNC/COCA (British National Corpus/Corpus of Contemporary American English) word family lists, and computerized text analysis tools based on word lists, such as Range. These studies all use a word family count with the expectation that “once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort” (Bauer & Nation, 1993: p. 253).

However, literature has also challenged the idea of the word family as a counting unit (Gardner, 2007; Kremmel, 2016; McLean, 2018; Reynolds, 2013; Stoeckel et al., 2021). Some empirical studies have indicated that learners of EFL experienced difficulty in learning derived words and provided evidence to support the challenge. Research with Japanese learners of English has shown that they had insufficient derivational knowledge (Schmitt & Meara,1997; Mochizuki & Aizawa, 2000; McLean, 2018). Among those studies, McLean (2018) examined 279 university-level Japanese learners of English regarding their knowledge of inflections and derivations. Participants were asked to write a Japanese translation for a target item. The accuracy rate for inflections was 98% when the participants knew the bases, whereas it was 54% for the derivations. Moreover, studies with Thai and Austrian learners of English revealed that knowing the base words did not guarantee that the learners would be aware of their derivations (Ward & Chunenjudaeng, 2009; Kremmel & Schmitt, 2016). Based on these studies, Brown et al. (2020) claimed that the lemma (the baseword and inflected forms of a word of a particular part of speech) or flemma (the base form and inflected forms of a word, regardless of part of speech) is the more appropriate counting unit for second language (L2) English learners than word family due to their limited derivational knowledge. Although Laufer (2021) argued strongly against the claim that word family is not suitable for the counting unit, it is proposed that the Nuclear Family List developed by Cobb and Laufer (2021) consisting of frequently used word family members, should be used for novice and intermediate learners to expose them to useful derived words. Thus, it is clear that some effort is required for L2 learners, at least in their early stages, to learn the members of a word family.

The current issue examines how English instruction can help learners effectively expand the members of word families. One way to accomplish this is to increase the learners' exposure to derived words in teaching and learning materials. However, it has been noted that this method provides limited input for the learners. Laufer and Cobb (2020) examined the frequency of prefixes and suffixes in graded readers as well as a limited number of academic articles, news articles, and novels; their results indicated that only a few prefixes and suffixes were required to read the texts. The graded readers examined in the study required "-ly" and "-y" suffixes to understand 98% of the text. Morita et al. (2019) investigated prefixes and suffixes in junior high school English textbooks in Japan to reveal that only limited types and tokens of prefixes and suffixes

were used. A similar result was found in high school English textbooks (Morita et al., 2021). While first language (L1) studies demonstrated that elementary school students encounter far more words with prefixes and suffixes (Anglin et al., 1993; Nagy & Anderson, 1984), graded readers and textbooks may not be sufficient for learners to grasp derivational words.

Another way to foster the learners' mastery of derivational words is to provide explicit instruction. While individual studies for explicit morphological instruction showed mixed results (e.g., Sritulanon, 2013 for not effective; Lin, 2019 for effective; Ross & Berwick, 1991 for effective in a limited domain), recent meta-analysis studies have found that explicit instruction regarding derivational affixes benefits both L1 and second language L2 English learners (Goodwin, 2016; Goodwin & Ahn, 2010; Kirby & Bowers, 2017). However, it is unclear which members in a word family are suitable for learning and teaching derivational forms. Therefore, the current study aims to effectively bridge this gap.

2.2 The present study

Extant literature has revealed that the exposure to affixes through learning materials is limited, although certain prefixes and suffixes occur more frequently. While it remains ambiguous as to which words in each word family are difficult or easy for learners, clearly some affixes are easier to master than others. One of the underlying factors is the combination of the base and affix. Specifically, "player" may be simple to learn, but the meaning of "sitter" is not directly drawn from "sit." This fact suggests that the difficulty level of a derivation should be judged word by word. Therefore, the present study attempts to prove the usefulness of word2vec, a widely used natural language processing (NLP) application, to reveal the derivations that need special treatment in teaching and learning English. It is anticipated that word forms that display unique behavior have lower similarity scores between the base form. For example, the usage of "lastly" differs from its base form "last" in that the former is used as a list marker while the latter can be used as either a verb or an adjective. It is expected that we may be able to identify word forms that require special treatment in education by observing the similarity scores.

3. Methodology

3.1 word2vec

According to the distributional hypothesis proposed by Harris (1954), words that denote similar meanings occur in similar contexts (see Sahlgren, 2008 for a detailed discussion). As a recent NLP technique, word embedding relies on this theory to map the word meanings to a set of numbers (vectors) using contextual information. A simplified model is described as follows to explain this process using the collocational information of the sample words.

Six nouns are selected here, one of which is masked for the purpose of demonstration ("apple," "car," "cat," "dog," "XXX" (masked), and "pencil"). Table 1 displays the frequencies of the verbs and adjectives ("buy," "drive," "eat," "fresh," "peel," "sharpen," and "stray") with the target nouns taken from the COCA using the following expressions: "[*verb*] [a] [*noun*]" and "[a] [*adjective*] [*noun*]," which allow us to include the inflected forms, such as "eats" and "ate." If we need to determine "XXX" in this table, one possible approach involves comparing the frequencies of the collocations to discover a word with a tendency similar to "XXX," presuming that such a word has something in common with the masked word. This simple method of using contextual (collocational) word information is based on the distributional hypothesis.

Table 1. The Sample Words' Verbal and Adjectival Collocations

	buy	drive	eat	fresh	peel	sharpen	stray
apple	68	0	140	112	19	0	0
car	1,118	1,428	1	5	0	0	1
cat	16	1	15	3	0	0	297
dog	93	1	32	9	0	0	533
XXX	9	2	28	245	41	0	1
pencil	3	1	3	4	0	17	1

However, it is difficult to identify the most similar word at a glance from the given table of raw frequencies. To mathematically calculate the distances between each word, the cosine similarity score is beneficial; it can be calculated using the following

formula where x_i and y_i denote the frequencies of each collocation of the target words:

$$\cos(x, y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

This score ranges from -1 to 1 and takes a value of 1 if all the numbers are identical. Table 2 displays the matrix of the cosine similarity scores between the sample nouns.

Table 2. Cosine Similarity Scores Between the Sample Nouns

	apple	car	cat	dog	XXX	pencil
apple	1.000	0.220	0.061	0.113	**0.679**	0.309
car		1.000	0.036	0.108	0.031	0.147
cat			1.000	0.993	0.021	0.075
dog				1.000	0.033	0.097
XXX					1.000	0.243
pencil						1.000

The highest score is 0.993 between “dog” and “cat,” given the simple fact that these two nouns are animals, and the others are not; naturally, the collocations of “dog” and “cat” are fairly similar. In the “XXX” column, the most similar word is “apple” (0.679). A natural guess is that “XXX” also denotes a type of food, or more specifically, a fruit. Actually, the answer is “orange.” Note that the meanings of “apple” and “orange” can be denoted as a set of numbers—or [68, 0, 140, 112, 19, 0, 0] and [9, 2, 28, 245, 41, 0, 1], respectively—which enables a calculation of the distance (similarity score) between them.

The word2vec application is based on this framework with mathematical sophistication. It was developed by Mikolov et al. (2013) and has been widely used in the NLP field, but has been scarcely utilized in the fields of linguistics and applied linguistics. Word2vec employs a shallow neural network model to efficiently learn the vector representations of words, although a detailed explanation is beyond the scope of this paper (see Goldberg & Levy, 2014 for a detailed explanation). If we use the same

previously explained approach with the entire COCA, the list of collocations continues endlessly up to the number of the types of words. The word2vec application automatically groups the collocations into certain dimensions, typically 200 to 500, to define the "meaning" of the words with a list of numbers.

Using the full-text data from the COCA (1990–2015; approximately 600 million words), we created a word2vec model with a *gensim* library in Python. To vectorize per sentence, the *sent_tokenize* function in the *nltk* library was utilized to separate the text into sentences, which were then converted to lowercase with the *LineSentence* function used for modeling.[1]

For example, this model generates a vector for "orange" of [-0.782, 1.389, -1.732, …, -0.234], or 300 numbers in total; if we calculate the top five similar words in terms of the cosine similarity score, we get "yellow" (0.682), "tangerine" (0.620), "peach" (0.606), "blue" (0.602), and "red" (0.600). As anticipated, these words relate to either colors or fruits. It should be noted that this pseudo-representation of word meanings reflects both semantic and syntactic characters. In other words, low similarity scores indicate semantic as well as syntactic differences between the word pairs.

One note to be added here is that the list may contain what are considered antonyms. For example, the closest word for "increase" in our model is "decrease" (0.878). This is a natural result considering that these two words can appear in extremely similar contexts. These can even be used interchangeably, in such sentences as "The number of students in the university *increased* [decreased] by 5% last year." This is at times perceived as a disadvantage of word embedding but is rather advantageous in this study. Antonyms typically belong to the same word family, including "like" and "dislike," "known" and "unknown," and "useful" and "useless." On the one hand, if the contexts are similar enough, word2vec will assign high similarity scores to the antonym, which may imply that there is no need for special treatment in the classroom. On the other hand, if the score for the relationship between the base word and its antonym is low, this suggests the need for special attention.

3.2 Word family list

One of the most extensive word family lists is Paul Nation's BNC/COCA family lists, a modified version of which is available as "BNC/COCA family lists + extras" (ver. 2.00)[2] from AntLab (managed by Laurence Anthony). This list contains

headwords with word family members with the following examples:

A	ABOUT
AN	ABOVE
ABLE	ABSOLUTE
ABILITIES	ABSOLUTELY
ABILITY	ABSOLUTES
ABLER	ABSOLUTISM
ABLEST	ABSOLUTIST
ABLY	ABSOLUTISTS
INABILITY	
UNABLE	

The list includes inflections, such as “abilities” as the plural form of “ability,” in addition to derivations; for instance, “ability,” “inability,” and “unable” are derived from “able.” It has 50,890 headwords and 105,476 forms, including each base form.

3.3 The CEFR-J wordlist

CEFR-J wordlist[3] (ver. 1.6), which is widely used in Japan, is employed as a testing vocabulary set to focus on words that are useful for English learners. The advantage of this list is that it classifies words into the CEFR levels (A1, A2, B1, and B2), which comprise a common structure to assess language ability. It indicates which word is more difficult for the learners to master, and thus, is a possible indicator of the difficulty of the word forms in a word family. Although the CEFR-J wordlist has different levels of the same surface form (e.g., sentence [noun]): A1, sentence [verb]: B2), for the sake of simplicity, the level of the highest one on the wordlist was chosen here.

The CEFR-J wordlist contains 7,801 words; we excluded compounds (e.g., “bus stop” and “each other”) and words without any inflections or derivations (e.g., “about” and “above”). Furthermore, we ignored the low-frequency British forms (e.g., “industrialise” and “familiarise”) as well as the words that appear less than 16 times in our dataset (e.g., “narcissistic” and “aubergine”) to ultimately yield a list of 6,290 words.

3.4 Dataset

We used the word family list to assign family members to each headword in the selected list (e.g., "ability," "abilities," and "inability," among others). This resulted in a list of 17,206 pairs, such as "able–ability," "able–inability," and "absolute–absolutist," with an average of 3.95 pairs per headword. As these pairs include rare word forms that can be considered unimportant to learners, the top 20,000 most frequent word forms in the COCA are chosen. This process excludes such word pairs as "absolute–absolutist," "bottom–bottoming," and "computer–computationally." Finally, 7,540 unique pairs were incorporated into our target dataset; Figure 1 illustrates the process of creating our dataset.

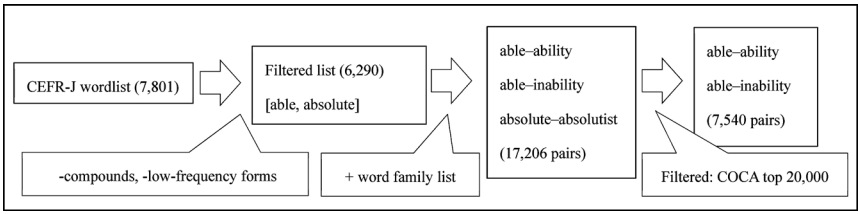

Figure 1. Creating the Target Dataset

4. Finding word pairs that deserve special attention

We hypothesized that words with a low similarity score are difficult to learn and hence require special attention in teaching and learning. To consider the CEFR levels of each word form, we selected 1,543 word pairs out of the 7,540 pairs in which both words were assigned a CEFR level for the current analysis. Table 3 displays some random examples of word pairs with low average similarity scores whose similarity score (represented as cos. in the table) is equal to or below 0.1.

The lowest-scoring word pair in this table is "detect" and "detective." These two forms seem to require special attention in the classroom. The word "detective" refers to a person or a police officer who investigates crimes; this word is typically taught on a different occasion from the verb "detect," which is primarily used in the context of experiments or machines (sensors). Thus, learners may consider these as different vocabulary words and may impose extra cognitive costs, although linking these words

Table 3. Examples of Word Pairs with a Low Similarity Score

word form 1	word form 2	cos.	word form 1	word form 2	cos.
detect (B2)	detective (B1)	-0.061	conduct (B1)	conductor (B2)	0.039
expect (A2)	unexpectedly (B1)	-0.035	suppose (B1)	supposedly (B1)	0.043
time (A1)	timeless (B2)	-0.025	exhaust (B2)	exhausting (B2)	0.044
vary (B1)	invariably (B2)	-0.025	remark (B2)	remarkably (B2)	0.053
double (A2)	doubly (B2)	-0.025	late (A1)	lately (B1)	0.054
over (A2)	overly (B2)	-0.014	character (A1)	characterize (B1)	0.063
count (B1)	countless (B1)	-0.001	end (A1)	endlessly (B2)	0.067
drama (A1)	dramatically (B2)	0.009	point (A1)	pointless (B2)	0.077
last (B1)	lastly (B2)	0.011	govern (B1)	governor (B1)	0.080
suit (A2)	suitable (A2)	0.018	ready (A1)	readily (B2)	0.080
time (A1)	timely (B1)	0.019	second (A1)	secondly (B2)	0.085
lightly (B1)	lighten (B2)	0.025	special (A1)	specialize (B1)	0.093
belong (A2)	belongings (B2)	0.028	mean (A2)	meaningful (B1)	0.094
total (B1)	totally (B1)	0.038	bear (A1)	unbearable (B2)	0.096
pave (B2)	pavement (B2)	0.039	remark (B2)	remarkable (B1)	0.100

using the word family concept would help them efficiently understand and remember the word meanings. Some other word pairs in Table 3 are also remarkable. For example, “last” and “lastly” as well as “suit” and “suitable” are clear cases of caution for teachers and learners. Specifically, the base forms are polysemous, and some do not explicitly relate to the meanings of the derived forms—and hence, deserve special attention. Other cases to note are “belong” (A2) and “belongings” (B2), “total” (B1) and “totally” (B1), and “character” (A1) and “characterize” (B1). These cases all demonstrate the usefulness of the word2vec similarity scores in discovering the word families that demand special care in an English educational context. It should be noted that the results reflect not only the semantic differences of each word but also their syntactic differences. For instance, there is a clear syntactic difference between “last” and “lastly,” especially since the latter functions as a sentential adverb, and the environment in which they appear is substantially distinct.

If low-scoring word pairs require special attention, it can be assumed that one of

the word forms is at a higher level. In other words, it can be predicted that the word pairs with different CEFR levels (e.g., A2–B1) will have lower similarity scores than the word pairs with the same CEFR level (e.g., B1–B1). Around 537 pairs (approx. 35%) out of 1,543 have the same CEFR level [e.g., "manage" (A2) and "manager" (A2)] whereas others (1,006 pairs) have different levels [e.g., "know" (A1) and "unknown" (A2)].

We conducted the Welch two-sample *t*-test (R ver. 4.1.0) with the similarity scores of the two groups. The findings revealed that the words with the same CEFR level [$N = 537$, $M = 0.42$, $SD = 0.18$] have significantly higher similarity score ($t(1034.7) = 6.10$, $p < .001$, $d = 0.33$ [0.23, 0.44]) than the words with different levels ($N = 1{,}006$, $M = 0.36$, $SD = 0.17$). In other words, the difference in the CEFR levels significantly contributes to the variance in the scores of each group. It should be noted that there are cases where each word needs special attention even when the words share the same CEFR level, and these cautious words can be located by using the similarity score. For example, in Table 3, both "suit" and "suitable" are labeled as A2, but they may be worth extra attention presumably due to the polysemy of the verb "suit." Thus, it can be concluded that the word2vec scores serve as a suitable indicator of word difficulty within a word family, particularly in terms of the CEFR level, and are useful in finding words of caution for teachers and learners.

5. Applications of the research results

5.1 Finding additional forms to be listed in the wordlist

The previous section has proved the usefulness of word2vec in identifying words that need special attention. However, we have confined our scope to cases of word forms listed in the CEFR-J wordlist. As an application of our methodology, this subsection attempts to discover word forms that are not currently included there but are noteworthy and could be added in a future edition of the wordlist. We are aware that each wordlist has its own policy for the selection of the words. For example, a wordlist may only include those words with basic derivations assuming that the complex forms (e.g., "unkindness") are rare, and their meanings could be deduced from its more basic form. However, even certain simple derivations could be noteworthy, and the following section is an attempt to find those word forms.

We used the CEFR-J wordlist as a base list to investigate each headword's family members from the BNC/COCA family lists. Subsequently, the similarity scores were calculated against its base form for the derived forms that are not included in the CEFR-J wordlist. For example, "acceptability" and "accepted" are not part of the CEFR-J wordlist, and therefore, we calculated their similarity scores with the base form "accept" (0.128 and 0.520, respectively). Given the 5,997 pairs (7,540–1,543) employed for this experiment, low similarity scores suggest that the word form behaves differently from the base form, hence requiring special attention. Table 4 lists the word forms with the lowest similarity scores against each base form.

Table 4. The Lowest Five Word Forms in the Similarity Score Against the Base Forms

Rank	Base	CEFR	COCA Rank	Form	CEFR	COCA Rank	cos.
1	decide	A2	1,663	decidedly	#N/A	11,057	-0.152
2	center	A2	8,194	centrist	#N/A	18,208	-0.110
3	correct	A1	1,876	correctional	#N/A	15,699	-0.107
4	mark	A2	3,728	markedly	#N/A	13,559	-0.101
5	offend	B2	14,929	offenders	#N/A	7,520	-0.091

This list includes the rare word forms (e.g., "centrist" and "correctional") and an inflected form (e.g., "offenders") but has simultaneously derived forms that deserve attention (e.g., "decidedly" and "markedly") with meanings and syntactic behaviors that clearly differ from the base form. Table 5 lists the words that were manually selected from our experimental results.

These word forms, which have low similarity scores and are not currently included in the CEFR-J wordlist, can be considered as irregular members within a word family, and hence, deserve special treatment in classrooms. The results indicate that certain base forms (e.g., "elevate" and "refine") are not included in the list. Furthermore, some frequent and important word forms (e.g., "notably" and "namely") are missing. Therefore, our approach is successful in finding the word forms that deserve consideration for future addition to the list.

Table 5. Examples of Word Forms with Low Similarity Scores

Base	CEFR	COCA Rank	Forms	CEFR	COCA Rank	cos.
elevate	#N/A	16,855	elevator	A2	5,211	-0.075
strike	A2	3,849	strikingly	#N/A	15,493	-0.029
note	A1	1,604	notably	#N/A	6,434	-0.017
react	B1	5,118	reactor	#N/A	9,807	-0.004
flat	B1	2,087	flatly	#N/A	15,426	-0.004
period	A1	669	periodically	#N/A	9,565	0.005
name	A1	376	namely	#N/A	6,827	0.008
man	A1	136	unmanned	#N/A	18,315	0.011
refine	#N/A	15,701	refinery	B2	17,599	0.012
man	A1	136	manned	#N/A	19,537	0.015

5.2 Difficulties of affixes

The similarity score can be calculated for each word form against its base, and it is possible to estimate the difficulties of affixes if we compute the average scores of the words with a specific affix. For instance, by comparing the average score of "kind–unkind," "conscious–unconscious," and "aware–unaware" against the average of "take–mistake," "understand–misunderstand," and "use–misuse", it would be clarified which ("un" or "mis") affix should receive more attention in teaching and learning.

For this purpose, we use the 1,543 word pairs with CEFR levels employed in the previous section. Based on the affix levels proposed by Bauer and Nation (1993), each pair is grouped into major affix patterns, such as "-ly," "un-," and "-ment." For example, the pairs "total–totally" and "kind–kindly" can be grouped as "-ly" pairs because both their derivation forms have the "-ly"' ending. Table 6 displays the affix levels, which are based on the affixes' productivity, predictability, and regularity, among other traits. It is assumed that elementary learners are only aware of affixes at the lower levels, but advanced learners know the higher-level affixes as well.

Consequently, 1,333 pairs out of 1,543 (86.4%) are given an affix classification while such irregular pairs as "we–ourselves," "who–whoever," and "that–those" are not classified. Each pair's scores were calculated based on their derivation and base forms; then, they were averaged within each group. For example, the "-ment" group has 43 pairs—such as "announce–announcement" (0.403), "measure–measurement" (0.585),

Table 6. Affix Levels (Bauer & Nation, 1993)

Level	Description / Affixes
1	Each form is a different word
2	Inflectional suffixes
3	The most frequent and regular derivational affixes
	-able, -er, -ish, -less, -ly [adv.][1], -ness, -th [ordinal number], -y [adj.], non-, un- [antonym] (all with restricted uses)
4	Frequent, orthographically regular affixes
	-al [adj.], -ation, -ess, -ful, -ism, -ist, -ity, -ize, -ment, -ous, in- (all with restricted uses)
5	Regular but infrequent affixes
	-age, -al [noun], -ally, -an, -ance, -ant, -ary [adj.], -atory, -dom, -eer, -en [adj.], -en [verb], -ence, -ent, -ery, -ese, -esque, -ette, -hood, -i, -ian, -ite, -let, -ling, -ly [adj.], -most, -ory, -ship, -ward, -ways, -wise, ante-, anti-, arch-, bi-, circum-, counter-, en-, ex-, fore-, hyper-, inter-, mid-, mis-, neo-, post-, pro-, semi-, sub-, un- [reverse]
6	Frequent but irregular affixes
	-(ate+)able[2], -ee, -ic, -ify, -ion, -ist [adding to unexplained consonant], -ition, -ive, -th, -y [noun], pre-, re-
7	Classical roots and affixes

Note. [1]Brackets [ ] after some affixes include the part of speech, meaning, or other information produced by attaching the affix. [2](*-ate*) before *-able* means that *-ate* is typically omitted, such as *permeable*.

and "pay–payment" (0.506)—with an average score of the pairs in this group being 0.313 (figures in parentheses indicate each similarity score). Some words were grouped into multiple categories, such as "will–willingness" ("-ing," "-ness"), "create–creativity" ("-tive," "-ity"), and "vary–invariably" ("in-," "-able," "-ly"). However, these are excluded from the calculation to avoid mixed effects. It is assumed that the average score represents the affix's difficulty. Table 7 displays the highest 10 and lowest 10 groups of affixes with more than four words.

As the table indicates, the following inflectional affixes have relatively high scores: "-s," such as "sport–sports" (0.702) and "thank–thanks" (0.695); "-ing," such as "wrap–wrapping" (0.709) and "feel–feeling" (0.656); and "-ed," such as "attach–attached" (0.695) and "worry–worried" (0.689). Additionally, various morphemes that

Table 7. Average Scores of Each Affix Group with More Than Four Words

No.	Highest 10			Lowest 10		
	Morpheme (Level)	Average	Count	Morpheme (Level)	Average	Count
1	-s (2)	0.509	15	-less (3)	0.141	18
2	in- (4)	0.498	15	-or (NA)	0.213	23
3	-ing (2)	0.484	127	-able (3)	0.272	22
4	-an (5)	0.463	6	-ence (5)	0.307	25
5	-ism (4)	0.460	6	-ment (4)	0.313	43
6	-ic (6)	0.442	22	-ive (6)	0.315	25
7	un- (3)	0.431	29	-er (3)	0.318	105
8	-ed (2)	0.429	80	-age (5)	0.344	7
9	-ist (6)	0.429	17	-ness (3)	0.349	23
10	-ship (5)	0.426	6	-ary (5)	0.358	5

create antonyms have high scores: “in-,” such as “accurate–inaccurate” (0.672), “expensive–inexpensive” (0.642), and “effective–ineffective” (0.624); and “un-,” such as “aware–unaware” (0.729), “familiar–unfamiliar” (0.679), and “happy–unhappy” (0.632). This is likely because they do not change the part of speech and are used in similar contexts. As morphemes with high scores are relatively easy to master, teachers may assume that learners do not need morphological instructions for them. However, even when an affix group can be considered as being straightforward for learners, some word forms deserve special attention. The word2vec results make it possible to list such cases. For example, “-s,” “-ing,” and “-ed” are simple inflectional endings, but such low-scoring word pairs as “mean–means” (0.392), “custom–customs” (0.285), “miss–missing” (0.119), “concern–concerning” (0.184), “puzzle–puzzled” (0.147), and “mark–marked” (0.173) should be carefully addressed in the classroom. Moreover, it should be noted that a few of the level 3 affixes—such as “-less,” “-able,” “-er,” and “-ness”—are contained in the lowest 10 list. Although these affixes are productive and predictable, their low scores imply the necessity for extra instruction. For instance, such word pairs as “point–pointless” (0.077), “need–needless” (0.095), and “hope–hopeless” (0.136) have low similarity scores, which implies that their usage and meaning of derivation highly differ from their base forms. Additionally, “-er,” which is one of the simplest suffixes, is not free from irregularity; its various word pairs—such

as “begin–beginner” (0.023), “train–trainer” (0.140), “stick–sticker” (0.150), “deal–dealer” (0.184), and “hold–holder” (0.211)—may require special treatment.

6. Conclusion

This study examined the usefulness of an NLP application called word2vec in analyzing word families. We employed the similarity scores generated by the application to assess the similarities among the word forms in each word family, uncovering cases that deserve special attention in teaching and learning English. In other words, the low similarity scores between the derivation and base forms clearly indicated the necessity for supplemental instruction on the word pair in the classroom context. Additionally, the results identified some word forms that could be included in a vocabulary list. Reorganizing the findings based on the affix groups revealed each affix’s difficulty level as well as the information on remarkable words within the group. These results reveal the usefulness of word2vec and potential possibilities for collaborations between the NLP and the English education fields.

The limitation of this study is that it discussed the similarity scores in relation to word difficulty but did not actually experiment with English language learners as subjects. Although there were statistically significant differences based on the CEFR levels, a more detailed analysis would be needed to objectively determine the difficulty level of each word form using the word pairs presented in this study. Furthermore, the focus of this study was on finding words in a word family that need special consideration in class and not on how to teach them. Future research might include a more qualitative analysis of the words with low similarity scores, including how they should be treated for different levels of learners.

Notes

1. The parameters for the model were: size = 300, window = 5, min_count = 3, and iter = 5, and the other parameters were set to default.
2. http://www.laurenceanthony.net/resources/wordlists/bnc_coca_cleaned_ver_002_20141015.zip
3. The CEFR-J Wordlist Version 1.6. Compiled by Yukio Tono, Tokyo University of Foreign Studies. Retrieved from http://www.cefr-j.org/download.html in June, 2020.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP22H00677.

References

Anglin, J. M., Miller, G. A., & Wakefield, P. C. (1993). Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development, 58*(10, serial no. 238). https://doi.org/10.2307/1166112

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

Brown, D., Stoeckel, T., McLean, S., & Stewart, J. (2020). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, *43*(3) 1-7. https://doi.org/10.1093/applin/amaa061

Cobb, T., & Laufer, B. (2021). The Nuclear Word Family List: A list of the most frequent family members, including base and affixed words. *Language Learning*, *71*(3), 834–871. https://doi.org/10.1111/lang.12452

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

Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.

Goodwin, A. P. (2016). Effectiveness of word solving: Integrating morphological problem-solving within comprehension instruction for middle school students. *Reading and Writing*, *29*(1), 91–116. https://doi.org/10.1007/s11145-015-9581-0

Goodwin, A. P., & Ahn, S. (2010). A meta-analysis of morphological interventions: Effects on literacy achievement of children with literacy difficulties. *Annals of Dyslexia*, *60*(2), 183–208. https://doi.org/10.1007/s11881-010-0041-x

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Kirby, J. R., & Bowers, P. N. (2017). Morphological instruction and literacy: Binding phonological, orthographic, and semantic features of words. In. *Studies in Written Language and Literacy* D. L. C. Cain & R. K. Parrila (Eds.), (437–462). https://doi.org/10.1075/swll.15.24kir.

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

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary Test Scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, *13*(4), 377–392. https://doi.org/10.1080/15434303.2016.1237516

Laufer, B. (2021). Lemmas, flemmas, word families and common sense. *Studies in Second*

Language Acquisition, *43*(5), 965–968. https://doi.org/10.1017/S0272263121000656

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

Lin, M.-F. (2019). Developing EFL learners' morphological awareness: Instructional effect, teachability of affixes, and learners' perception. *International Review of Applied Linguistics in Language Teaching*, *57*(3), 289–325. https://doi.org/10.1515/iral-2015-0081

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

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

Morita, M., Uchida, S., & Takahashi, Y. (2019). The frequency of affixes and affixed words in Japanese junior high school textbooks: A corpus study. *ARELE: Annual Review of English Language Education in Japan*, 30, 129-143.

Morita, M., Uchida, S., & Takahashi, Y. (2021). The frequency of affixes and affixed words in Japanese senior high school English textbooks: A corpus Study. *ARELE: Annual Review of English Language Education in Japan*, 32, 81-95.

Nagy, W. E., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly*, *19*(3), 304–330. https://doi.org/10.2307/747823

Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, P. (1983). Teaching and testing vocabulary, *Guidelines*, *5*(1), 12–25.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *Language Teacher*, *31*, 9–13. https://jalt-publications.org/tlt/issues/2007-07_31.7.

Nurmukhamedov, U., & Webb, S. (2019). Lexical coverage and profiling. *Language Teaching*, *52*(2), 188–200. https://doi.org/10.1017/S0261444819000028

Reynolds, B. L. (2013). Comments on Stuart Webb and J. Macalister's "Is text written for children useful for L2 extensive reading?" *TESOL Quarterly*, *47*(4), 849–852. https://doi.org/10.1002/tesq.145

Ross, S., & Berwick, R. (1991). The acquisition of English affixes through general and specific instructional strategies. *JALT Journal, 13*(2), 131-140. Retrieved from https://jalt-publications.org/sites/default/files/pdf-article/jj-13.2-art2.pdf

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, *20*, 33–53.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework. *Studies in Second Language Acquisition*, *19*(1), 17–36. https://doi.org/10.1017/S0272263197001022

Sritulanon, A. (2013). The Effects of Morphological Instruction on Reading Abilities of Low Proficiency Adult EFL Learners at a University in Thailand. *LEARN Journal: Language Education and Acquisition Research Network*, *6*(1), 49–65. Retrieved from https://so04.tci-thaijo.org/index.php/LEARN/article/view/102721

Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, 1–23. https://doi.org/10.1017/S027226312000025X

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

（内田　　諭　九州大学）
（森田　光宏　広島市立大学）