An Etymologically-Categorized Vocabulary Order Generator for Increasing and Expanding Vocabulary Learning: A Sequential and Personalized Approach

Yi-Hung Liao, Meng-Hsuan Tsai, Po-Chang Lee and Yi-Lin Tsai

Abstract

Vocabulary learning is recognized as a pivotal process to acquire proficiency and competence in second/foreign language acquisition. Word power not only substantiates the learners' perception of knowledge for accuracy but also facilitates the learners' production of the target language for fluency. This paper suggests an etymologically-categorized approach for students to increase and expand their vocabulary size based upon their own prerequisite vocabulary learning experiences and cognitive level. Applying such sequential etymological method, vocabulary order is computed based on the individual learner's past vocabulary navigation history as well as the comprehensive personalized learning patterns and profiles in terms of word structure analysis (prefix, suffix, and roots). This recommended vocabulary order generator (VOG) has important pedagogical implications vis-à-vis the need for the inclusion of etymological cues and the optimal mode for personalized learning to enhance second language vocabulary gains.

Keywords—vocabulary learning; prefix; root; suffix; etymology, content-based filtering; vocabulary order generator

I. Introduction

Vocabulary learning and teaching has been extensively researched in decades. It is irrefutable that vocabulary plays a pivotal role in communication as the word power facilitates and substantiates both the L1 and L2 learners' acquisition of knowledge and production of knowledge. There are many prominent studies and rich theoretical orientations show that vocabulary knowledge is highly correlated with reading comprehension across the age span: in primary grades (Baker, Simmon & Kame'enui, 1998) in the intermediate grades (Beck, Perfetti, & McKeown, 1982), in high school (Cunningham & Stanovich, 1997), and with adults (Beers, 2007; Stahl & Fairbanks, 1986). The importance of vocabulary is even recognized in significant ways within the Common Core State Standards (CCSS), as well as the Next Generation Science Standards. For example, the CCSS for English Language Arts call for students to determine the meanings of unknown and multiple-meaning words and phrases using context clues (Language Anchor Standard 4); analyze meaningful word parts (Language Anchor Standard 4); and use a range of general academic and domain-specific words and phrases (Language Anchor Standard 6; NGA & CCSSO, 2010). However, increasing and expanding vocabulary has been a tedious work for both teachers of different levels and students of all ages. Owing to the fact that vocabulary knowledge is a precondition of learners' discourse comprehension and vocabulary learning is an indispensable process particularly for ESL/EFL learners, this paper aims to be in search for an adaptive and generative approach for students to increase and expand their vocabulary size based upon their own prerequisite learning experiences and cognitive level.

II. Literature Review: Vocabulary Learning & Instruction Strategies

As early in the early 20th century, John Dewey (1910) stated that vocabulary is critically important because a word is an instrument for thinking about the meanings which it expresses. Since then, there has been an "ebb and flow of concern for vocabulary" (Manzo, Manzo, & Thomas, 2006, p. 612; see also Blachowicz & Fisher, 2000). At times, interest in vocabulary has been high and intense, and at other times low and neglected, alternating back and forth over time (Berne & Blachowicz, 2008). Therefore, learning vocabulary is an important instructional aim for teachers in all content areas in schools. Recent research, however, indicates that vocabulary instruction may be problematic because many teachers are not "confident about best practice in vocabulary instruction and at times don't know where to begin to form an instructional emphasis on word learning" (Berne & Blachowicz, 2008, p. 315).

Theories of reading comprehension, such as the Verbal Efficiency Theory (Perfetti, 1985), and the Lexical Quality Hypothesis (Perfetti, 2007; Perfetti & Hart, 2001, 2002) propose that difficulties in reading comprehension partly arise from poor word identification skills and weak representations of a word's form (orthography and phonology) and its meaning. These theories are supported by correlational data that indicate word knowledge to be a critical predictor of reading ability (Dixon, LeFevre, & Twilley, 1988; Ouellette, 2006; Patterson & Hodges, 1992). Given the increased emphasis on vocabulary learning and instruction, commonly pedagogical approaches on vocabulary acquisition can be roughly divided into four major modes: incidental vocabulary learning; alphabetical sequence vocabulary learning; level of frequency in vocabulary learning; and level of difficulties in vocabulary learning

1. Incidental vocabulary learning

Incidental vocabulary acquisition refers to the absence of the conscious intention to commit a word to memory (Rott, 2012). It is generally described as the "picking up" of new words randomly when learners are engaged in an extensive tasks in reading, listening, speaking, or writing. In other words, word knowledge is thought to be accumulated and developed gradually through continued exposures in various contexts. Informal and incidental vocabulary learning is quite efficient and effective. In particular, Nagy et al,'s (1985) notion of incremental learning through repeated exposures has consequential in the realm of L2 vocabulary pedagogy; "Twenty-five to fifty percent of annual vocabulary growth can be attributed to incidental learning from meaningful context while reading (p. 134)" become a strong impetus for avid of proponents of reading including Krashen's naturalistic acquisition (1987) for promoting L2 vocabulary learning.

2. Alphabetical sequence vocabulary learning

Take a quick scan of the self-learning vocabulary reference materials displayed on the bookshelves in the bookstores, many of them are organized and presented with lists of words from A to Z like a dictionary. Some learners indeed benefit from this kind of vocabulary input and practice using a dictionary-like alphabetical sequence of word list, provided with pronunciation, bilingual translation, part of speech, sample sentences, and so on. Popular instructional or self-study references on vocabulary learning are like TOEIC Word list, TOEFL Vocabulary List.

3. Level of frequency in vocabulary learning

Word Lists by frequence, mostly developed in academic institutions, are lists of words grouped by

frequency of occurrence within certain given academic areas or text corpus, either by levels or as ranked list. This kind of frequency word lists are made for lexicographical purpose of vocabulary acquisition, serving as a sort of learners' checklist to ensure intelligibility and comprehensibility. Coxhead (2000), from Victoria University of Wellington, New Zealand, created the Academic Word List (AWL), which is a list of 570 word families that represent the general academic vocabulary from college textbooks, professional journals, and other academic writing. Gardner and Davies (2013) further warrantee another list of core academic word list titled A New Academic Vocabulary List (AVL), intending to improve the learning, teaching and research of English academic vocabulary in its many contexts.

4. Level of difficulties in vocabulary learning

In many language learning textbooks or readers for EFL learners, the target vocabulary list presented on the back for each chapter or book mostly conforms to the level of difficulties; in turn, reflecting the criteria of different level in the general English language assessment exams, such as GEPT (General ENglish Proficiency Test) from the Language Training & Testing Center in Taiwan, Main Suite test (KET/PET/FCE) from Cambridge English Language Assessment, and so on. Another common vocabulary learning strategies used by language learners are tier-like techniques. As Beck, Mckeown & Kucan, 2002, 2008, 2013) created a three-tiered system for learning vocabulary: Tier-1 words: basic words that appear in most children's vocabulary such as *party, walk, swim, house, ugly,* and so on; Tier -2 words: frequently occurring words that have "high utility for mature language users and across a variety of domains… like *contradict, circumspect, precede, auspicious, fervent, contrast*" which are great for explicit and targeted vocabulary instruction; Tier-3 words: consisting of low-frequency words that are often limited to specific fields or content areas (such as science or social studies) with example words like *"filibuster, pantheon, and epidermis"* (Beck, et al,. 2013, p. 9)

5. Morphological awareness instruction

Proficient readers use morphemic analysis in several ways. They begin by noting a word's use in context ("Distances among the stars are just incredible!"). They break the word into parts (i*n* + *cred* + *ible*) and assign meaning to each part (*in = not, cred = believe, ible = can be done*). Then they use the word-part meanings to put the word together again ("*cannot be believed*") to see if this meaning makes sense in the selection. Proficient readers also use morphemic analysis to identify words that are derived from a common base word (e.g., *night* as in midnight, nightly, nightshirt) or root (e.g., *cred* as in credo, credential, incredible) to determine word meanings. Understanding of morphemes in word formation and lexical processing constitutes significant values for the learners' vocabulary size as well as the literacy development for reading comprehension. Once the language teachers recognize this given fact and are able to provide a series of systematic vocabulary instruction within the classroom contexts, ESL/EFL learners' vocabulary knowledge and capacity will be increased and enhanced to some degree.

Morphemes are the smallest units of meaning in a language--units that can serve as independent words (e.g., *deep*) or that are added to such word (e.g., *-en* in *deepen*). Morphemes are combined in different ways to express particular meanings or to make different grammatical patterns or parts of speech (e.g., *heal, health, healthy*).

III .Vocabulary Learning & Instruction Recommendation Model

1. Scenario:

To apply for the graduate studies in the United States, Michael (a non-native English speaker from Taiwan) is required to take the Test of English as a Foreign Language (TOEFL) and Graduate Record Examination (GRE). Like most international applicants to prepare these intimidating graduate school preparation tests, Michael is working on vocabulary practice book from the well-known publishers like *Barron's, Princeton Review,* or *Kaplan Test Prep* on his own. A hypothesis of generative and adaptive learning sequence to increase and expand Michael's vocabulary size are presented as below:

T: the **T**otal number of suggested academic words to be learned in the practice books
K: a set of already **K**nown words from the student
R: a set of **R**emaining new words to be learned
T = **K** + **R**
n: next recommend word

Can Michael's current existed vocabulary acquisition profile and learning record (K: known and acquired vocabulary) help himself and the language instructor(s) to elicit an effective and meaningful vocabulary learning strategy to acquire the new words to be learned (n) from the remaining set (R: remaining vocabulary to be learned)? If we were to able to recommend a new word out of the remaining word set to a student, can we filter our criteria or choices based on students' acquired known vocabulary and past learning experiences?

2.Common approaches on vocabulary instruction and acquisition:

(A) decoding morphological segmentation approach

Given the values of morphemes in word recognition, common vocabulary instruction approach is to develop the learners' morphological awareness. In other words, morphological awareness teaching techniques lie in guiding learners to experiment with ways that morphemes are/can be combined. Take the word "*contradiction*" for instance, as they encounter different words and sentence contexts from their past learning experiences, learners can analyze and decode the unfamiliar word "*contradiction*" by breaking it down into units of meaning *ex* + *tract* + *ion* based on the morphological awareness (see Table 1.1). Moreover, they gradually learn to decipher and come to understand the use of prefix of "*ex*" referring "out of" (as the known word "***ex***it"), the root of "*tract*" meaning "to pull or to drag" (as the known word "dis***tract***"), and the suffix of "*ion*" denoting action or condition to form a noun (as the known word "ac***tion***"). Thus, when encountering a new word, the learner is able to construct the word meaning by applying his/her prerequisite morphological awareness and known linguist information cues to infer the meaning of the new target word and further to facilitate his/her reading comprehension

Table 1.1: morphological segmentation strategy [extraction]

extraction [ex-tract-ion]	**affixes**	**meaning**	**other sample word**
ex	prefix	out	**ex**it

tract	roo t	pull	dis**tract**
ion	suffix	noun agent	act**ion**

With sufficient access to morphemes and the richness of linguistic information cues (e.g., semantic features, grammatical roles) when encountering an unfamiliar words, learners indeed benefit from the morphological segmentation strategy of lexical representations which contributes to increasing vocabulary readiness and reading comprehension (Reichle & Perfetti, 2003).

(B) extended morphological series approach

Another frequently-used pedagogy applied by English language teachers to develop the learner's vocabulary growth is to provide direct instruction of extended series of words based on etymology and morphological cues. In either reading or listening classes, for instance, the English instructor will usually explicitly explain the chosen target word "*extraction"* from the context: the meaning, the pronunciation, the part of speech, the usage and the sentence sample, etc. In addition to the above typical pedagogy approach of direct vocabulary instruction, the English instructor are likely to introduce the other series of correlated words which share the same prefix "ex" or "e" (e.g., ***ex****tract, **ex**pel, **e**ject, **e**gress, **ef**fuse*, as in column (a); share the same word "extract" (e.g., ***extract****ion, **extract**ive, **extract**or*, as in column (b); share the same root"tract" (e.g., ***tract****ion, **tract**or, **tract**ive, **t ract**able, **tract**ile*, as in column (c); or share the same suffix"tion" or "ion" (e.g., *act**ion**, reaction, interaction, digression*, as in column (d). In other words, when encountering to the target word *"extraction,"* learners will be intentionally instructed and exposed to additional series of group words (with at least 4 types and 17 meaningful words) that are correlated to "*extraction,*" and thus they will not only increase their sensitivity about the morphological cues and word recognition but also expand their vocabulary size and growth.

Table 1-2: extended morphological series [extraction]

***extraction* (Target word to be taught)**	**Correlated Word Examples**
(a) Same prefix with different roots	***ex****traction, **ex**pel, **e**ject, **e**gress, **ef**fuse*
(b) Same word with different suffixes	***extract****ion, **extract**ive, **extract**or*
(c) Same root with different prefixes	*ex**tract**, re**tract**, pro**tract**, at**tract**, con**tract***
(d) Same suffix with different words	*extract**ion**, retract**ion**, protract**ion**, attract**ion,** contrac**tion***

(C) generative morphological plane mode

The popularity of pedagogical strategy of correlated morphological series extension in the English language classroom indicates that morphological awareness or analysis plays a crucial role in learners' vocabulary acquisition based upon their prior and prerequisite morphemic knowledge. Therefore, once the learner has acquired more morphological segmentation units (e.g. *ex. re, pro, ad,*

con, tract, pel, ject, gress, fus… see Table 1-3) and accumulated more series of correlated morphological words (series one of prefix extensions: *extract, expel, eject, egress, effuse*; series two of root extension: *retract, protract, attract, contract...*), he or she is gradually building up the consciousness to infer the meaning of any unfamiliar words by using the morphological segmentation and correlation series strategies, such as *repel* [*"re" + "pel"*], *project* [*"pro" + "ject"*], *aggressive* [*"ag" + "ress"*], *confuse* [*"con" + "fus"*] and so on based upon his/her prior acquired morphemic analysis strategies. With this deductive teaching and learning pattern, the learner's assets of vocabulary knowledge will be increasing from one word (*extract*) or one series of words (*extract, expel, eject, egress, effuse...; retract, protract, attract, contract…*) into a more complex plane mode of vocabulary clusters (at least 16 more words: *repel, propel, appeal, compel, reject, project, adjacent, conjecture, regress, progress, aggressive, congress, refuse, profuse, affusion, and confuse,* see Table 1-3) in terms of word consciousness and recognition, leading to the vocabulary growth and lifelong reading success.

Table 1-3 generative morphological plane

	ex	re	pro	ad	con
tract	extract	retract	protract	attract	contract
pel	expel	repel	propel	appeal	compel
ject	eject	reject	project	adjacent	conjecture
gress	egress	regress	progress	aggressive	congress
fus	effuse	refuse	profuse	affusion	confuse

III.A SIMPLIFIED CASE

To begin with, we compare our etymological approach with two other frequently-used learning order: random and alphabetical. In this simplified case, we have five words total (T)

T: {extraction, extractive, tractable, excavate, profusion}

The only known word (K) is "extraction" and we need to recommend one word (x) out of the rest of the four words (extractive, tractable, excavate, profusion).

K:{extraction}

R: {extractive, tractable, excavate, profusion}

1. Random order

If we choose each recommend word randomly for R rounds, the first round we have R choices and the second round we have R-1 choices until the Rth round which we only have one choice left. Therefore, we have R*(R-1)*(R-2)*(R-3)...2*1 different paths.

In simplified case, set R has 4 words and therefore R!=24 possible paths.

2. Alphabetical order

If we choose each recommend word alphabetically for R rounds, all we need to do is sorting set R from A to Z and the result is the alphabetical order of set R. In comparison to the random method, this is one deterministic path out of the R! paths.

In simplified case, the alphabetically path is as followed,

excavate, extractive, profusion, tractable

3. Etymological order

Of all the R! possible paths of acquisition, is the learning outcomes for each of the round equally effective, or some of them are better than the others? Our goal is to optimize the best path for the learner based on his/her past learning experiences. Can the information of a student's past learning record (**K**) help us to generate a better target vocabulary to be learned from the remaining set (**R**) for the next word (**n**)? If we were to recommend a new word out of the remaining word set to a student, can we select our choices based on students' known words.

To this end, we pick the word based on the closeness in meaning and also the etymological relationship. The purposely-designed scenario has the intrinsic characteristics: each of the remaining words are etymologically related to the known word by (word, root, prefix and suffix).

Pair	etymological relation
extraction / **extract**ive	share the word “extract”
ex**tract**ion / **tract**able	share the root “tract”, which means “pull” in Latin
extraction / **ex**cavate	share the prefix “ex”, which means “out”
extract**ion** / profus**ion**	share the suffix “ion” which is a noun agent

By inspection, we reaffirm our intuition that in terms of closeness between meanings

A learner’s awareness of

word > root > prefix > suffix

Hence, we develop each round of pick as following

Round	K: Known	R: Remaining words to be learned	n: suggested words	suggestion rationale
Round 1	extraction	extractive, tractable, excavate, profusion	extractive	extractive is the adjective form of **extract**, while extraction is the noun form.
Round 2	extraction, extractive	tractable, excavate, profusion	tractable	"**tract**able" shares the same root of spelling **"tract"** meaning "to pull" as in "ex**tract**ion" and also the meaning.
Round 3	extraction, extractive, tractable	excavate, profusion	excavate	**ex**cavate, meaning to "dig out" shares the same prefix "**ex**," with **ex**tract (to pull out),
Round 4	extraction, extractive, tractable, excavate	profusion	profusion	Profus**ion**, the last choice left to be learned share the same suffix **-ion (**which indicate the noun agent attribute) with extract**ion**, but with little correlation in meaning.

Therefore, this new method has the recommendation order : extractive, tractable, excavate, profusion.

In summary, we can get a taste of the advantage of etymological order in this simplified case: Instead of choosing the order randomly or alphabetically, it exploits the relationship of each word in terms of the meaning. Therefore, learners can pick up new word step by step with least amount of effort, which means the next word is most closely related to the known words etymologically and, most of the time, partially meaning (prefix, root or suffix) of the next chosen word is already seen for learners from the previous learnt words. Now, we will specifically explain how to extend this proposed simplified case to general case in the next section.

IV. The general case of content-based filtering in vocabulary

In order to build a general recommender system of learning new vocabulary, a database with morphological segmentation information of every word is needed. In this paper, every derivative English word is modeled as the concatenation of five distinctive parts: prefix, root, front word, hind word and suffix. Few examples are shown to further explain this proposed model.

A compound word like "basketball" will be a concatenation of front word and hind word.

basketball = basket (front word) + ball (hind word)

In a similar fashion, prefix and suffix can also be placed at the front and end of a root.

consequence = con (prefix) + sequ (root) + ence (suffix)

To maintain the consistency of this model, words with multiple prefixes or suffixes are defined as followings:

reinvention = re (prefix) + invention

nationalism = national + ism (suffix)

Last but not least, if a word is not derivative from another word or root, then this word is denoted as "singular".

In other words,

WORD	Prefix	Root	Head Word	Tail Word	Suffix
basketball	-	-	V	V	-
nationalism	-	-	V	-	V
reinvention	V	-	V	-	-
consequence	V	V	-	-	V
elephant	-	-	-	-	-

Also, for simplicity consideration, some words not applicable for this model are considered as rare cases. For example, maintain (two roots) or improvise (two prefixes and one root).

From the aforementioned simplified case, we want to extend the idea into a more generally-applicable, automatic, and systematic method in which a computer or any teacher can replicate and follow. The intuition in the simplified case is that the closer the etymological relation, the higher the recommending priority. Our objective here is to create a method to select a word that is closest to the set K in terms of etymological relationship.

To establish the etymological relationship between a word and a set of words, we need to start from the simplified case when there is only one known word. Based on the aforementioned model, we define the closeness between two different words k and x as a function of matched components in both words.

$$f_v(k, x) = w'w_w + p'w_p + r'w_r + s'w_s + s'w_s$$

where

w'=1 if k and x share the same base word, otherwise w'=0

p'=1 if prefix of k is the same as the prefix of x, otherwise p'=0

r’=1 if root of k is the same as the root of x, otherwise r’=0

s’=1 if suffix of k is the same as the suffix of x, otherwise s’=0

The intuitive idea is to assign the weights according to the importance of word, root, prefix and suffix to the etymological relation and therefore the relative weights among word, root, prefix and suffix should have the general property as

$$w_w > w_p > w_r > w_s$$

When the number of known word is greater than 1, we simply sum up the score of a word choice to each of the known words. That is, we generate a of the vicinity between our next word choice and all our known words.

$$F_v(K, x) = \sum f_v(k_i, x)$$

While finding the highest score of the group scores among the choices, we come to the conclusion that this choice has the closest etymology relation with a group of unknown words. Therefore, we make it the highest recommending priority for next word.

$$x_{next} = \max F_v(K, x_j)$$

V. Discussion

In this section, we explore the scalability and optimization aspects of this etymological recommend approach.

1. Scalability

The beauty of our etymological order is that it is applicable to any set of words. In other words, the total number of words in the given set is scalable. The following are the applications based on the total number of words.

- (10-100 words) unknown words in an article or a paper
- (2000-3000 words) frequent word list for tests like SAT, TOEFL or GRE
- (3000 or more words) a professional field like medicine or mathematics

2. Optimization

The “weight” attribute (w_w, w_p, w_r, w_s) in the proposed etymological approach is suggested but undetermined for scale of this paper. However, this opens up the new search objective for the optimization of the weights (the relative importance of base word, prefix, root and suffix in terms of the closeness of any two words). To optimize the weight attribute, we need to collect learners’ learning statistics. Due to the nature of this objective, we foresee plausible AI application to this weight optimization.

I. Conclusion

In the paper, we show how exploiting the etymological relation among words could help learners find the better learning order. In the future, we would further pursue other methods for adjusting the weighting parameters to optimize this approach.

References

1. Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing Words to Life: Robust Vocabulary Instruction.* New York: The Guilford Press.

2. Beck, I. L., Perfetti, C., & McKeown, M. G. (1982). The effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology, 74,* 506–521.

3. Beers, K. (2007). The measure of our success. In Beers, K., Probst, R. E., & Rief, L. *Adolescent literacy: Turning promise into practice,* pp. 1-14. Portsmouth. NH: Meinemann.

4. Berne, J. I., & Blachowich, C. L. Z. (2008). What reading teachers say about vocabulary instruction: Voices from the classroom. *The Reading Teacher, 62*(4), 314–323

5. Blachowicz, C. L. Z., & Fisher, P. (2000). *Teaching vocabulary in all classrooms.* Englewood Cliffs, NJ: Prentice Hall.

6. Coxhead, Averil (2000) A New Academic Word List. *TESOL Quarterly, 34*(2), 213-238.

7. Krashen, S. (1987). *Principles and practice in second language acquisition.* New York: PrenticeHall International.

8. Manzo, A. V., Manzo, U. C., & Thomas, M. M. (2006). Rationale for systematic vocabulary development: Antidote for state mandates. *Journal of Adolescent & Adult Literacy, 48*(7), 610–619.

9. Reichle, E.D. and C.A. Perfetti. 2003. Morphology in word identification: A word-experience model that accounts for morpheme frequency effects. *Scientific Studies of Reading 7*(3): 219–237.

10. Rott, S. (2012). Incidental Vocabulary Acquisition. *The Encyclopedia of Applied Linguistics.*

11. Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based metaanalysis. *Review of Educational Research, 56,* 72–110