Extensive Reading through Watching English-subtitled K-dramas

Wenhua Hsu

whh@isu.edu.tw

I-Shou University, Taiwan

Abstract

This research was inspired by the phenomenon of binge-watching Korean television drama series (K-dramas) among Taiwanese university students. K-drama fever may provide an impetus for sustained English-subtitled K-dramas viewing, which may serve as an alternative to extensive reading of graded readers that English teachers often encourage their students to do. The researcher compiled a corpus of 25+ million English-subtitled words from 240 different K-dramas, totaling 5,398 episodes and measured the vocabulary levels thereof along the BNC/COCA word-frequency scale. Results show that K-drama English subtitles reached the 2000–3500 word-family levels at 95% text coverage and extended to the 4000–5500 levels at 98% coverage subject to genres. EFL K-drama fans can encounter most of the first 5000 word families often enough for potential learning to occur through their continually watching up to 48 English-subtitled K-dramas. The results may serve as a reference for extensive reading practitioners and learners who are concerned with a certain vocabulary goal within the first 5000 word families.

Keywords: lexical text coverage, English-subtitled K-dramas, word-frequency levels, BNC/COCA

Introduction

The researcher-teacher at times overhears her students talking about which K-drama they have been binge-watching lately. The binge-watching phenomenon started in the late 2000s as a result of the rapid rise of OTT (over-the-top) streaming platforms, which can be accessed anywhere anytime and bypass traditional appointment viewing in a once-daily or once-weekly broadcast format. In a survey conducted by Netflix (2013) on nearly 1500 TV streamers, the respondents delineated their binge-watching behavior as viewing between 2–6 episodes of the same program in one sitting.

Korean Wave, termed by Beijing journalists to describe the surge of Korean entertainment, began in the mid-1990s and has had a history of international followers for over two decades (Jin & Yoon, 2017; Kim, 2013; Lee & Nornes, 2015). On streaming platforms such as Rakuten Viki, English is often the first language into which K-dramas are translated, and then as a pivot translation, English subtitles are translated into many other languages (Dwyer, 2017; Pedersen, 2019). With the prevalence of mobile devices with Internet connectivity, K-dramas are popular on college campuses in Taiwan, where English as a foreign language (EFL) is taught as a compulsory course. As such, the researcher-teacher is concerned with vocabulary growth if the viewing habit shifts from Chinese to English

subtitles. According to Markham and Peter (2003), subtitles are on-screen text in the viewer's first language (L1) with L2 audio as opposed to L1 videos with L1 captions. Different from the two situations, the present context refers to subtitles in the viewer's L2 (English) with L3 (Korean) audio, which may compel EFL non-Korean viewers like Taiwanese students to rely upon on-screen text in English. For this reason, watching English-subtitled K-dramas may serve as an alternative to reading graded readers extensively. The former and the latter have some common purposes such as building reading fluency, reading for pleasure as well as sustaining vocabulary growth (Day & Bamford, 2004).

This research addressed two research questions (RQ) regarding K-drama English subtitles as learning input and the vocabulary learning opportunities they afford.

RQ1: What are the vocabulary levels of K-drama English subtitles at 95% and 98% lexical text coverage?

RQ2: By watching English-subtitled K-dramas for one to four years during college, how many words beyond the first 2000 word families may EFL K-drama fans encounter at least 12 times to have a chance of learning them?

Literature Review

Extensive Reading

Extensive reading involves reading for pleasure over time and has long been proposed to complementing intensive reading for language learning (Day & Bamford, 1998, 2004; Horst, 2005). Learners freely select reading materials of their interest that are easier than or fit their current proficiency level without feeling overwhelmed (Waring & McLean, 2015). For extensive reading to be effective, learners need to read in quantities large enough to have sufficient encounters with unfamiliar words so as to eventually internalize how they are used and how they collocate with the words around them (Nation, 2015).

Extensive reading programs often resort to graded readers, which are books of simplified fiction and non-fiction classics for EFL learners. Publishers offer graded reader series with 6 to 7 levels, ranging from 250 to 3800 headwords (Claridge, 2012). The highest levels from Oxford Bookworms and Cambridge Readers top out at 4000–5000 headwords. In the recent decade, mid-frequency readers up to 8000+ headwords (Nation & Anthony, 2013) have been created to better graded reader resources. One may visit the Extensive Reading Foundation at http://erfoundation.org/wordpress/ for a wealth of graded readers.

As aforementioned, when Taiwanese college students choose English (L2) subtitles instead of Chinese (L1) while watching a Korean-speaking (L3) drama, they need to read subtitles in order to understand character dialogues, since they know nothing or a little bit about the Korean language. As with graded readers in extensive reading programs, English subtitles in this case may turn to be a valuable resource for lexical learning. Therefore, the vocabulary levels involved in K-drama English subtitles are an important issue that needs to be examined first (see RQ1).

Lexical Text Coverage and Repetitions for Vocabulary Learning

Lexical text coverage refers to "the percentage of running words in the text known by the reader" (Nation, 2006, p. 61). For adequate comprehension, Nation (2006) advocated that 98% lexical text coverage, two unknown words per 100 words, is ideal for guessing words from context and may provide good conditions for incidental vocabulary learning. Previous studies have disaccorded in the percentage of lexical coverage in a text that is needed before comprehension is interrupted. In the literature, two putative coverage points have been advanced for setting lexical thresholds for different purposes: 95% for reasonable comprehension (Laufer & Ravenhorst-Kalovski, 2010) as well as 98% for pleasure reading (Hirsh & Nation, 1992) and for independent/unassisted reading (Hu & Nation, 2000). As Laufer (1989) pointed out, the two coverage points suggest probabilistic thresholds over which learners have a chance of gaining a certain level of comprehension.

There is some reason to believe that in the current context of English-subtitled K-drama viewing, a lower criterion such as 95% coverage may be a likely condition for guessing successfully. In her recent research, Laufer (2020) found that 95% coverage could be reached with an initial knowledge of 90% words in the text and inferring an additional 5%. Guessing correctly, in part, consists in the amount of contextual information to help infer meanings (Webb, 2008). Video affords viewers extra-linguistic support (e.g., character actions, gestures and facial expressions) to gain understanding (van Zeeland & Schmitt, 2013). Visual clues may make it less difficult for EFL learners to guess meanings from context, even though their lexical knowledge falls short of 98% coverage. Webb and Rodgers (2009a) even surmised that the necessary coverage for adequate comprehension of television programs is likely to be as low as 90%. Taken together, this research adopted both 95% and 98% coverage as a gauge for measuring vocabulary thresholds.

As to the number of repetitions for vocabulary learning, research on vocabulary learning has documented that a single encounter with a novel word seldom supports robust learning of it (Horst, 2013; Horst et al., 2011). Referring to past studies, Nation (2014) inferred that it may take from five to 16 exposures to a word or a chunk for uptake to happen and conjectured that 12 encounters with a new word in a variety of contexts would just be enough to develop knowledge of that word. Following Nation (2014), the researcher adopted 12 repetitions as a cutoff frequency to address RQ2.

Vocabulary Learning from Television Programs and Movies

To promote vocabulary growth, Webb and Rodgers (2009a) proffered the idea that television (TV) programs can be engaging to L2 learners. They collected the transcripts of 88 American and British TV programs of various genres and used the ranked fourteen 1000 word-family lists derived from British National Corpus to measure the vocabulary levels. They targeted 95% and 98% coverage as the lower and upper boundaries to assess vocabulary demands. Results showed that the vocabulary sizes for 95% coverage ranged from the first 2000 to 4000 word families subject to genres, while the thresholds for 98% coverage spanned from the first 5000 to 9000 word families. Including proper nouns

and marginal words, knowledge of the first 3,000 word families provided 95.45% text coverage, while knowledge of the first 7,000 word families accounted for 98.27% coverage.

In a similar vein, Webb and Rodgers (2009b) measured the vocabulary demands of British and American movies. Like TV programs, movies entailed knowledge of the first 3000 word families plus proper nouns to reach 95% coverage. When set at 98% coverage, movies involving 6000 word families were less vocabulary-demanding than TV programs stretching to 7000 word families. The results have some implications for vocabulary learning. In recent years, a plethora of streaming platforms have brought moviegoers from a big screen to a personal screen so movies may be watched as frequently as TV programs. In their subsequent study on how TV programs can be utilized effectively for language learning, Rodgers and Webb (2011) proposed narrow viewing by watching related programs or by watching a complete season of a program instead of an equivalent number of random episodes from different programs. They provided some numerical evidence that the number of word families occurring once was smaller and the number of low-frequency word families occurring 10 times or more was larger in related programs than in unrelated programs. This suggests that programs with a similar topic or a story arc linking them in a season are likely to have a lower vocabulary load than a set of unrelated programs. Related TV programs may use fewer word families but learners may have a greater number of encounters with members of those word families, increasing the potential for vocabulary learning.

It is important to note that the above TV program corpora comprised the scripts of English-speaking programs, whereas the present corpus contained the English subtitles of Korean-speaking dramas. The researcher admits that it is more advisable for EFL learners to view English-speaking programs than non-English speaking programs with English subtitles. Nevertheless, avid fans can marathon a K-drama in one sitting on holidays or engage in serial viewing over days or weeks during their spare time (Rubenking & Bracken, 2021). The researcher argues that it is the impetus that makes possible sustained reading of English subtitles while binge-watching K-dramas.

Research Methodology

The Corpora

A typical weekday K-drama consists of 16 one-hour episodes, while a weekend family-oriented K-drama has 50 or more 40-minute episodes. The present corpus comprised 25,967,735 English subtitled words from 240 different K-dramas with high viewership ratings based on Nielsen Korea and OTT services, totaling 5,398 episodes with nearly 5,000 hours. High viewership means that these K-dramas are likely to have been viewed by a large number of fans. As of 2022, the five highest-rated miniseries on cable TV were *The World of the Married, Reborn Rich, Sky Castle, Crash Landing on You* and *Reply 1988*. Although this corpus contained more the recently popular miniseries, earlier national television networks' high-rated dramas were also included. For example, *Jewel in the Palace* with 60.8% nationwide viewership in 2003 have been broadcasting in 150 countries thus far.

To build a corpus with a good representation of K-dramas, a number of genres were taken into account (see Table 1). K-drama storylines often involve some societal issues to which global audience can relate, such as bullying, corruption, wealth inequality, sexual harassment as well as wide-ranging topics like family value, reincarnation, friendship, marriage, conflicts between classism and individualism. The diversity of subject matter in K-dramas implies that English subtitles thereof may be a rich language learning resource.

K-drama English subtitles in SubRip Text (SRT) files were downloaded from the Internet for research purposes. They are the plain text files, including the text of the subtitles in sequence along with the start and end timecodes. There are one or two lines of subtitles per scene and at most 7 to 8 words per line. Below is the excerpt in SRT format from Episode 16 of *Twenty-five Twenty-one*.

69

00:04:41,379 --> 00:04:44,507

I applied for a correspondent position

in New York and I got it.

…

101

00:07:21,122 --> 00:07:22,540

Do you have some time today?

102

00:07:22,623 --> 00:07:25,084

Of course. My time is all yours.

It should be highlighted that the English subtitles of a K-drama provided by different OTT services may not be exactly the same or may not have total accuracy. However, the present corpus was large and should provide a reliable assessment of vocabulary levels. Over 25 million subtitled words from 240 K-dramas with 5,398 episodes made up the corpus.

Ten genre sub-corpora with 12 K-dramas in each were created for comparison. Inclusion of 12 K-dramas in each sub-corpus was based upon the assumption that it is feasible for K-drama fans to watch one 16-episode drama per month, 12 K-dramas per year. However, a viewer's preference is changeable. Many fans follow their favorite k-drama stars (e.g., Hyun Bin 현빈, Song Joong-ki 송중기) and view their dramas regardless of genres. Also, a large number of audience watch K-dramas based on netizens' recommendations. In view of these two binge-watching choices, another two genre-mixed sub-corpora with 12 K-dramas in each were established (see the bottom two rows of Table 1 for popular K-drama stars and netizens' recommendation).

Instrument and Procedure

Drawing upon the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), Nation (2016, 2017) compiled twenty-five 1000-word-family lists based on frequency

and dispersion. He also gathered the ever-growing word lists of proper nouns, marginal words (e.g., spoken interjections), transparent compounds and acronyms. These word lists were installed into the AntWordProfiler program (Anthony, 2022) to profile the vocabulary levels of K-drama English subtitles.

Table 1

Sizes of the K-drama Sub-corpora

Sub-corpora	Number of K-dramas	K-dramas included	Tokens
Coming of age	12	*All of Us are Dead, Fight for My Way, The Heirs,* etc.	1,402,357
Historical	12	*Under the Queen's Umbrella, The Red Sleeve, Moon Embracing the Sun,* etc.	1,398,896
Time travel	12	*Mr. Queen, Go Back Couple,* etc.	1,323,344
Thrillers	12	*Kingdom, Penthouse,* etc.	1,389,778
Crimes	12	*My Name, Voice, Stranger, Mouse, Tunnel,* etc.	1,399,873
Medical	12	*Hospital Playlist, Doctor Romantic,* etc.	1,357,213
Legal	12	*One Dollar Lawyer, Innocent Defendant, Hyena,* etc.	1,466,432
Fantasy	12	*Guardian-The Lonely and Great God, My Love from the Star,* etc.	1,395,774
Action	12	*Again My Life, Vincenzo, Vagabond, City Hunter,* etc.	1,289,347
Romance	12	*Business Proposal, It's Okay to Not Be Okay, What's Wrong with Secretary Kim,* etc.	1,333,276
Popular K-drama stars	12	*Reborn Rich, Descendants of the Sun, Crash Landing on You,* etc.	1,377,456
Netizens' recommendations	12	*Signal, Dazzling, My Mister, Reply 1988,* etc.	1,376,499

Dang and Webb (2016) carried out a series of tests on a number of word lists and discovered that the BNC/COCA word-family lists performed consistently better on a variety of written and spoken texts than the other word lists in terms of greater text coverage. In subsequent research, Dang et al. (2022) recommended that the BNC/COCA2000, providing the greatest coverage of all kinds of texts and thus giving a good return for learning, is a very helpful high-frequency word list for L2 learners. The BNC/COCA word-family lists were therefore employed in this research.

Targeting 95% and 98% coverage, the researcher first sized up the vocabulary demand of K-drama English subtitles by counting the number of the ranked twenty-five 1000-word-family lists that was needed until the cumulative coverage reached 95% and 98%. Meanwhile, the vocabulary level was

extrapolated according to which 1000-word-family list was the last one being added. Then the number of words beyond the first 2000 word families appearing 12 times or more (hereafter 12+ times) was tallied. Before the calculation, the feasibility study for viewing English-subtitled K-dramas and for 12 repetitions as a cutoff point was undertaken, followed by an interview (see *Feasibility Study* later).

Word Families Used for Frequency Count

This research involves reading English subtitles while viewing a K-drama. For reading purposes, word recognition is essential for comprehension. Under the educational system that requires English as a college entrance exam subject, Taiwanese students already have some knowledge of root words, prefixes and suffixes as well as parts of speech, which contributes to word recognition when any member of a word family appears in a range of contexts (Nagy et al., 1989). In consideration of their English abilities, word types and tokens were not considered in the calculation of recognition vocabulary.

According to Webb (2021), word families as a counting unit may be preferable to lemmas in avoiding an overestimate of the recognition vocabulary amount necessary for adequate comprehension. A lemma comprises a word's base form and its inflected forms that are all the same part of speech, while a word family additionally includes word forms made with derivational affixes (Nation & Meara, 2010). For example, *age*, *ages*, *aged* and *aging* (all in verb forms) belong to the same lemma, while *age* and *ages* as nouns would be a different lemma. When it comes to lexical units, the above instance would be counted as two lemmas and one word family. There are thousands of word families and tens of thousands of lemmas (Webb, 2021).

Since the focus of this research was not to compile a word list for specific purposes, 'word family' was used as a counting unit to measure recognition vocabulary amounts. The frequency of a word stem plus its inflected and derived forms was summed up so that the total frequency would indicate the number of repetitions of the word in English subtitles.

Data Processing

The subtitles in SRT format were saved as text files in UTF-8, which is a standard system for encoding text on the web and encoding all possible characters in human languages. When running the AntWordProfiler program, the numeric time codes were eliminated automatically. Before the corpus was inputted into the AntWordProfiler program, contractions and connected speech were changed with the aid of TextMate 2.0 to conform with the spellings of the words in the BNC/COCA word lists. For instance, *I'll, wanna,* and *gotcha* were changed to *I will, want to*, and *got you* respectively. Otherwise they would be classified by AntWordProfiler as 'Words Not Found in Base Lists' (hereafter referred to as off-list).

If off-list words were personal or geographical names (e.g., Young-Hee for a Korean name, Mapo Gu for a Korean place), they were supplemented to the existing proper noun list. For hyphenated compounds in the off-list (e.g., *work-related*, *stir-fry*), TextMate 2.0 was utilized to replace these words with a space. The hyphens were summarily removed. Likewise, closed compounds in the off-list (e.g.,

bathwater, *birdhouse*) were supplemented to the existing compound list to avoid double counting if their component words are already in the BNC/COCA word lists. The acronyms in the off-list were added to the existing acronym list.

Proper nouns, marginal words, transparent compounds and acronyms do not make reading effortful. Excluding their text coverage would inflate the vocabulary level and overrate the vocabulary demand necessary for adequate comprehension, since the reading burden they present is manageable. Proper nouns are recognizable because of capitalization of the first letter. Marginal words (*geez*, *ouch*) can be ignored due to little influence on comprehension. It is not difficult to infer the meaning of transparent/semantically-compositional compounds if learners are already familiar with their individual words. Though being infrequent, acronyms do not pose a hurdle in reading. The meaning of an acronym may be talked about in subsequent dialogues, or it may be a familiar one in daily life. To avoid an overestimate, the researcher counted them in until 95% and 98% coverage were achieved.

Table 2

Vocabulary Levels of Three Hit K-dramas

K-drama	*Twenty-five Twenty-one*		*Attorney Woo*		*Squid Game*	
Genre	Coming of age		Legal drama		Survival thriller	
Word list	Tokens/ coverage	Cumulative coverage	Tokens/ coverage	Cumulative coverage	Tokens/ coverage	Cumulative e coverage
Proper nouns	2,715/ 2.68%	2.68%	3,671/ 3.12%	3.12%	561/ 1.44%	1.44%
Marginal words	313/ 0.31%	2.99%	235/ 0.20%	3.32%	132/ 0.34%	1.78%
Compounds	290/ 0.29%	3.28%	345/ 0.29%	3.61%	128/ 0.33%	2.11%
Acronyms	163/ 0.16%	3.44%	281/ 0.24%	3.85%	23/ 0.06%	2.17%
1st 1000	90609/ 89.33%	92.77%	98449/ 83.57%	87.42%	35287/ 90.62%	92.79%
2nd 1000	3,738/ 3.69%	***96.46%***	6,028/ 5.12%	92.54%	1,304/ 3.35%	***96.14%***
3rd 1000	1,407/ 1.39%	97.85%	3,945/ 3.35%	***95.89%***	592/ 1.52%	97.66%
4th 1000	841/ 0.83%	***98.68%***	2146/ 1.82%	97.71%	378/ 0.97%	***98.63%***
5th 1000	562/ 0.55%	99.23%	590/ 0.50%	***98.21%***	121/ 0.31%	98.94%
6th 1000	234/ 0.23%	99.46%	790/ 0.67%	98.88%	101/ 0.26%	99.20%
7th 1000	105/ 0.10%	99.56%	215/ 0.18%	99.06%	78/ 0.20%	99.40%
8th 1000	150/ 0.15%	99.71%	449/ 0.38%	99.44%	109/ 0.28%	99.68%
9th 1000	85/ 0.08%	99.79%	169/ 0.14%	99.58%	39/ 0.10%	99.78%
10th–25th 1000	217/ 0.21%	100%	489/ 0.42%	100%	86/ 0.22%	100%
Total	101,429		117,802		38,939	

Feasibility Study

Three K-dramas *Squid Game, Twenty-five Twenty-one* and *Extraordinary Attorney Woo* (hereafter referred to as *Attorney Woo*) were selected for feasibility study because they were among the big hits on Netflix in 2021 and 2022. Table 2 displays their vocabulary profiles.

Both *Twenty-five Twenty-one* and *Attorney Woo* had 16 episodes and contained 101,429 and 117,802 subtitled words respectively, equivalent to one novel in terms of word count, while the 9-episode *Squid Game* had only 38,939 words. Among the three K-dramas, *Attorney Woo* was the most vocabulary-demanding, reaching the 3000 word-family level at 95.89% coverage and the 5000 word-family level at 98.21% coverage. In contrast, knowledge of the first 2000 word families alone would be enough to reach over 95% coverage for *Twenty-five Twenty-one* (96.46%) and *Squid Game* (96.14%), and mastery of the first 4000 word families would suffice to grip over 98% coverage for these two K-dramas.

Three vocational high school graduates participated in this feasibility study. At that time, the three volunteers just finished the technical college entrance exam and were free during the summer vacation before entry to college. The Vocabulary Levels Test (VLT) was used to measure their vocabulary size. As per Webb et al. (2017), the VLT consists of 150 test items with 30 from each of the BNC/COCA 1st to 5th 1000 word families. If test takers correctly answer 29 or more of the 30 test items from the first 1000 word families, they have achieved mastery of that vocabulary level. They recommended a cutoff point of 29/30 for mastery of the 1000–3000 levels and 24/30 for the 4000–5000 levels.

Table 3

Student Score Profiles on the VLT

Level	1st 1000	2nd 1000	3rd 1000	4th 1000	5th 1000
Student 1	30/30	29/30	29/30	26/30	20/30
Student 2	29/30	29/30	21/30	15/30	9/30
Student 3	29/30	29/30	24/30	16/30	11/30

Table 3 demonstrates that Student 1 has reached the 4th 1000 level, because he scored 26/30 items correct at the 4th 1000 (> 24/30). Student 2 has acquired a base vocabulary of the most frequent 2000 words, while Student 3 had recognition knowledge of over the first 2000 word families but still needed to learn words at the 3rd 1000. According to their vocabulary capacities, Student 1 was asked to watch *Attorney Woo*, while Students 2 and Student 3 were assigned to view the survival thriller *Squid Game* and the coming of age drama *Twenty-five Twenty-one* respectively. When *Twenty-five Twenty-one* and *Attorney Woo* were first aired in 2022, they were all busy preparing for the college entrance exam and none of them had ever watched these two K-dramas before the preliminary study. However, Student 1 and Student 3 had watched Chinese-subtitled *Squid Game*, so Student 2 was requested to view it at her spare time. Although the researcher did not specify the deadline in the hope of incidental

word learning under real-life conditions, all of them finished watching the designated K-dramas within two weeks. Later in the interview, they mentioned that they used to binge-watch quite a few Chinese-subtitled K-dramas on long holidays. It was the first time for them to view an English-subtitled K-drama.

Three weeks prior to viewing, the three students were given a vocabulary meaning-recognition pre-test to assess knowledge of the target words. The three-week interval was intended to wash away the memory of the target words. They were asked to choose the correct Chinese meaning from a set of options for each word. To avoid guessing, 'I don't know' was added to the options for each test item (as below).

Test item #23.　perjury　(A) 偽證　(B) 傷害　(C) 陪審團　(D) 竊盜　(E) 不知道

From each of the three K-dramas, six to seven target words appearing fewer than 12 times were chosen, totaling 17 target words with three words appearing in two of the three K-dramas. The 17 target words are the words beyond the 5th 1000 level, which were likely to be unknown to the three students. Moreover, to prevent them from being alert to the target words, the 17 words were intermixed with 33 words as distractors, which did not occur in the three K-dramas. After finishing watching, the three students were given the post-test containing the same test items but in different order, followed by an interview (see *Interview Results* later).

Table 4

Number of Repetitions of the Target Words and the Test Results

K-drama	Target word	vocabulary level	Number of repetitions	Pre-test answer	Post-test answer
Attorney Woo	revoke	7th 1000	10	S1, S2, S3: All wrong	S1: Correct S2: Wrong S3: Wrong
Attorney Woo	overrule	8th 1000	8	All wrong	S1: Correct S2: Wrong S3: Wrong
Attorney Woo	perjury	9th 1000	10	All wrong	S1: Correct S2: Wrong S3: Wrong
Attorney Woo	expropriate	10th 1000	7	All wrong	All wrong
Attorney Woo	tactician	14th 1000	11	All wrong	S1: Correct S2: Wrong S3: Wrong
Squid Game	naive	5th 1000	3	S1: Correct S2: Wrong S3: Wrong	S1: Correct S2: Wrong S3: Wrong
Squid Game	elastic	5th 1000	3	All wrong	All wrong

Squid Game	thug	7th 1000	9	S1: Correct	S1: Correct
				S2: Wrong	S2: Correct
				S3: Wrong	S3: Wrong
Squid Game	mannequin	11th 1000	2	All wrong	S1: Wrong
					S2: Correct
					S3: Wrong
Squid Game/ *Twenty-five* *Twenty-one*	moron	10th 1000	3/1	All wrong	S1: Wrong
					S2: Correct
					S3: Wrong
Attorney Woo/ *Squid Game*	dementia	6th 1000	10/ 3	All wrong	S1: Correct
					S2: Correct
					S3: Wrong
Attorney Woo/ *Squid Game*	deceive	6th 1000	7/3	All wrong	S1: Correct
					S2: Wrong
					S3: Wrong
Twenty-five *Twenty-one*	cushion	5th 1000	6	S1: Correct	S1: Correct
				S2: Wrong	S2: Wrong
				S3: Wrong	S3: Correct
Twenty-five *Twenty-one*	pathetic	5th 1000	7	All wrong	S1: Wrong
					S2: Wrong
					S3: Correct
Twenty-five *Twenty-one*	traitor	7th 1000	10	All wrong	S1: Wrong
					S2: Wrong
					S3: Correct
Twenty-five *Twenty-one*	prodigy	9th 1000	4	All wrong	S1: Wrong
					S2: Wrong
					S3: Correct
Twenty-five *Twenty-one*	surety	11th 1000	7	All wrong	All wrong

Note: S1, S2, S3=Student 1, 2, 3; S1 viewed *Attorney Woo*, S2 *Squid Game*, S3 *Twenty-five Twenty-one*.

The post-test scores show that there is potential for incidental vocabulary learning through watching English-subtitled K-dramas. The three participants successfully guessed meanings from context and made gains in meaning recognition for most of the target words (see Table 4), even though they appeared fewer than 12 times. In particular, Student 2 chose the correct meaning of the word *mannequin* (the 11th 1000 level), which occurred merely twice in *Squid Game*. The reason was later revealed in the *Interview Results*. The results manifested that following Nation (2014), the cutoff point at 12 encounters for uptake to occur was not a too optimistic decision. Twelve encounters or more may

not result in full knowledge of a word, but the gain in partial knowledge of words at different encountered frequencies would incrementally build up learners' breadth and depth of lexical knowledge. This feasibility study brought a beacon of hope for low-intermediate learners like Student 2 with a base vocabulary only, who was able to view English-subtitled K-dramas like *Squid Game*.

Results and Discussion

Vocabulary Levels of K-drama English Subtitles

Concerning RQ1 'What are the vocabulary levels of K-drama English subtitles at 95% and 98% lexical text coverage?', Table 5 provides a snapshot of the vocabulary size necessary for good comprehension associated with 95% and 98% text coverage. The over-25-million-token corpus contained 240 K-dramas with 5,398 episodes and involved 11,835 word families.

Table 5

Coverage of the BNC/COCA Word Lists in the English Subtitles of 240 K-dramas

Word lists	Tokens	Text coverage	Cumulative coverage	Word families
Proper nouns	570,824	2.243%	2.243%	X
Marginal words	142,652	0.561%	2.804%	X
Compounds	60,476	0.238%	3.041%	X
Acronyms	61,696	0.242%	3.284%	X
1st 1000	22,237,994	**87.390%**	90.674%	1,000
2nd 1000	1,277,964	**5.022%**	**95.696%**	1,000
3rd 1000	389,372	1.530%	97.226%	999
4th 1000	221,212	0.869%	**98.096%**	991
5th 1000	137,380	0.540%	98.636%	960
6th 1000	85,738	0.337%	98.972%	904
7th 1000	48,752	0.192%	99.164%	841
8th 1000	52,514	0.206%	99.370%	755
9th 1000	37,754	0.148%	99.519%	709
10th—25th 1000	122,460	0.481%	100.000%	3,676
Total	25,446,788	100.000%		11,835

The 1st 1000 word families made up 87.39% of the total words in the corpus and the 2nd 1000 accounted for 5.022%. The combined coverage of the first 2000 word families (=92.412%) was higher than the average 86%–90% covered by the first BNC 2000 word families in texts of different genres (Nation, 2006). The above average coverage (92.412%>86%–90%) offers a glimpse of hope for

students with a base vocabulary only. As shown in Table 5, after the first 2000 word families, the text coverage of the 3rd 1000 dropped abruptly to below 2% and rapidly decreased to less than 1% at the 4th 1000.

This reveals that K-drama English subtitles used a small vocabulary, converging at the first 2000 level. Knowledge of the first 2000 word families plus proper nouns, transparent compounds and marginal words would suffice to provide 95% coverage for minimally acceptable comprehension (95.696%). At the level of the 4th 1000, the cumulative coverage reached 98.096%.

Table 6 demonstrates individual vocabulary demands of 12 sub-corpora. Along the BNC/COCA scale, K-dramas English subtitles involved varying vocabulary thresholds, ranging from the first 2000 to 5500 word families. The vocabulary necessary to reach 95% and 98% coverage across different genres was mostly consistent at the 2000 and 4000 word-family levels respectively, with a very few genres ranging from the 3000–3500 levels at 95% coverage to the 5000–5500 levels at 98%.

Table 6

Vocabulary Demands across Genres at 95% and 98% Text Coverage

Sub-corpus	Vocab demand at 95%	Vocab demand at 98%	Sub-corpus	Vocab demand at 95%	Vocab demand at 98%
Coming of age	2000	4000	Legal	3500	5500
Historical	3000	5000	Fantasy	2000	4000
Time travel	2000	4000	Action	2000	4000
Thrillers	2000	4000	Romance	2000	4000
Crimes	2000	4000	Popular stars	2000	4000
Medical	3500	5500	Netizens' recommendations	2000	4000

In Table 6, medical and legal dramas were the most vocabulary-demanding (knowledge of the first 3500 word families for 95% coverage and an extension to 5500 for 98%), followed by historical dramas (reaching the 3000 level at 95% coverage and the 5000 at 98%). It is not surprising to see that medical and legal dramas required the largest vocabulary size to reach 95% and 98% text coverage, since the content areas of these two genres often entail specialist knowledge. Low-intermediate learners may find some difficulty in viewing English-subtitled medical and legal K-dramas.

However, there was not much variation in vocabulary levels among the other genre sub-corpora either at 95% or 98% coverage. They attained 95% coverage at the 2000 word-family level plus proper nouns and marginal words and reached 98% coverage at the 4000 level. That is, watching K-dramas of any of these genres (coming of age, time travel, thrillers, crimes, fantasy, action, romance, popular stars and netizens' highly-recommended miniseries) did not demand a much higher vocabulary capacity than the others. The little difference in vocabulary demands between these sub-corpora may be explained by

most K-dramas being a fusion of several genres. For example, a historical K-drama *Mr. Queen* is intertwined with romance, time-travel fantasy and body-swap comedy, leading to many twists and turns.

The 95% coverage in most K-drama genres and slightly below 95% coverage in medical and legal genres that knowledge of the first 2000 word families plus proper nouns and marginal words would provide may still fall within the parameters where good comprehension takes place (Nation, 2006). In other words, five unknown words or slightly more in every 100 words may be tolerable before they disrupt comprehension. Different from extensive reading of graded readers, English subtitles are supported with visual imagery, which further lowers the vocabulary threshold. Therefore, EFL K-drama fans may be encouraged to give viewing English-subtitled K-dramas a try, because reading English subtitles may not be a formidable task as we have expected (see *Interview Results* later).

When learners' vocabulary has developed to the 5500 word-family level, they would feel at equal ease when viewing English-subtitled K-dramas of any genre, in terms of the frequency of guessing words. By and large, EFL fans planning to watch English-subtitled K-dramas may need to know a minimum of the first 2000 word families and optimally, 4000 word families, in order to read English subtitles (see Table 5).

Potential Vocabulary Growth beyond the First 2000 Word Families

Assuming fans view one K-drama of 16 to 20 episodes per month, watching 12 to 48 K-dramas over one to four years is equivalent to exposure to about 1.3 to 4.9 million English-subtitled words. To measure 4 different sized datasets, a random selection of 12, 24, 36 and 48 K-dramas from the 240-Kdrama corpus was repeated 10 times respectively and the results of each dataset were averaged by 10 times of measurement. If 12+ repetitions for vocabulary learning to occur is sensible, the bottom row in Table 7 answers RQ2 about how many words beyond the first 2000 word families EFL K-drama fans may encounter often enough for learning to occur by watching English-subtitled K-dramas during college. An average of 3346, 4725, 5617 and 6234 word families would be met sufficient times for potential vocabulary growth when they finish viewing 12, 24, 36 and 48 K-dramas respectively.

It is worth noting that the words appearing fewer than 12 times in a single K-drama may be met a few more times later when viewing another. When more and more K-dramas are watched, the number of word families beyond the first 2000 occurring 12+ times would gradually increase. As shown in Table 7, fans just need to watch 12 English-subtitled K-dramas (attainable within one year), they would encounter most of the 3rd 1000 word families 12+ times (878/1000 > 800/1000), if we consider 'most' as 'at least 80%'. Knowing 80% or more of the words at a certain 1000-word-family level indicates that one knows enough of the words at that level to focus on learning words from the next level (Nation, 2014). If fans continue to watch K-dramas for another year, making a total of 24 K-dramas, they would meet 842 of the 4th 1000 word families 12+ times on average. When they keep viewing English-subtitled K-dramas during 4 years of college, eventually they would meet most of the 5th 1000 word families (824) often enough to have a chance of acquiring them (see Table 7 for the column of 4 years).

However, even though avid fans can watch 240 K-dramas for over many years (reaching the 7th 1000 level), they would still not encounter most of the 1000 word families (>800) enough times from the 8th 1000 onwards. This gives some indication that from 48 to 240 K-dramas, the vocabulary recycling of K-drama English subtitles is strong within the first 5000 word families, but EFL learners may make slow progress beyond that level.

Table 7

Number of Word Families Occurring 12+ Times

Viewing amount	12 K-dramas	24 K-dramas	36 K-dramas	48 K-dramas	240 K-dramas
Attainable time frame	1 year	2 years	3 years	4 years	20 years
Subtitle amount (in words)	1.3 million	2.5 million	3.7 million	4.9 million	25+ million
1st 1000	979	980	990	1000	1000
2nd 1000	955	974	987	998	1000
3rd 1000	**878**(>800)	940	961	976	999
4th 1000	737	**842**(>800)	**895**(>800)	913	991
5th 1000	535	701	778	**824**(>800)	960
6th 1000	403	556	645	687	894
7th 1000	293	447	533	591	**831**(>800)
8th 1000	220	356	441	498	725
9th 1000	147	267	347	399	688
10th–25th 1000	133	616	1017	1346	2926
Total	5280	6679	7594	8232	11014
Subtotal (beyond 2000)	3346	4725	5617	6234	9014

Interview Results

During the interview, when asked how they felt about watching an English-subtitled K-drama, the three participants showed a level of agreement in their responses. They remarked that English subtitles were not so daunting as they had expected. They gave their reason that through imagery and cumulative background knowledge, it was not arduous to understand the meaning of characters' dialogues per scene even though there were some unfamiliar words in the subtitles. They admitted that although initially they had difficulty catching up with the subtitling speed and needed to press pause constantly to read subtitles, they picked up reading speed later on and became used to the broadcasting speed. They also commented that they had no choice but to rely on English subtitles because of little knowledge of the Korean language.

Overall, their responses hint that while 95%–98% coverage (percentage of known words in a text) may be ideal, in the current setting, it was likely that coverage lower than 95% may have provided

general comprehension of the plots. Because the three students mentioned the use of guessing strategies for unknown words, ample visual representations may have assisted them in effectively link English subtitles (word form) with Chinese meanings and have therefore facilitated comprehension as well as incidental vocabulary learning on occasion.

Very surprisingly, they even observed that some daily expressions with familiar words had never appeared in their General English textbooks but they could guess meanings effortlessly for the first time encounter. Examples include *loan sharks* (money lenders who charges extremely high rates of interest), *take a leak* (urinate), *pull yourself together* (recover control of emotions and behave normally again), *give me an earful* (give me a lengthy reprimand), *he is so worked up* (bring one to a state of excitement or anger), *get wasted* (drunk), *seeing each other* (dating), *you are dead meat*, *you are done for* (in big trouble), *work one's ass off* (work extremely hard), to name but a few. Student 2 even joked that she would call her sister a loan shark, because she always asks more in return. *Give me an earful* appeared only twice in *Twenty-five Twenty-one* Episode 6. However, Student 3 used this phrase without hesitation by complaining that his mother would give him an earful whenever he binge-watches K-dramas but she herself has been watching even more. The interview reveals that learners would still likely make gains in developing different aspects of lexical knowledge for familiar words, such as idioms and phrases.

Moreover, the three students pointed out a situation where they could not guess some difficult words but they could guess the meanings when they were used with other words. To verify their statements, the researcher played back some scenes in some episodes and asked them about the meanings of some expressions with unknown words. All of them guessed the meanings accurately. For instance, *served as a debt guarantor* was Student 3's Chinese answer to the phrase *stood surety for a loan* (the unknown word *surety* from the 11th 1000 level) occurring three times in *Twenty-five Twenty-one* Episode 9. Student 1 replied that money is awarded to owners for their lands taken away by the government for the phrase *land expropriation compensation* (the unknown word *expropriation* from the 10th 1000) appearing twice in *Attorney Woo* Episode 8. In *Squid Game* Episode 7, where the remaining 16 players are forced to cross a sky-high bridge consisting of 18 pairs of glass tiles with each made of one of the two kinds, tempered glass (which can withstand the weight of two people) and normal glass (which shatters easily), Student 2 blurted out *reinforced, safety glass* in Chinese for *tempered glass*. Even though she did not know the word *tempered*, she still could guess the meaning correctly based on the scene she witnessed.

Besides, the researcher thought the word *mannequin* (the 11th 1000 level) was more difficult than the word *tempered* (the 7th 1000) because the former has more letters and irregular spelling-sound correspondence. To her surprise, Student 2 (the least proficient among the three participants) could recognize the Chinese meaning of *mannequin* in the vocabulary post-test, even though the word appeared only twice in *Squid Game*. Student 2 told the researcher that she often helped her parents with clothing business. The scene in which the players are asked to choose one out of 16 mannequins numbered from 1 through 16 impressed her, because there are many mannequins in her parents' store.

She recollected that she chose the correct meaning of this word in the post-test through the association of the word part *man* in *mannequin* with the dummy imagery. Student 2's remarks imply that learners may be inclined to uptake the words that relate to their life experiences and relatable scenes inspire them to use some learning strategies to retain word meanings.

Conclusion and Implications

This research was a preliminary study on the impact of K-dramas (part of Korean Wave) on EFL learners. Because our students enjoy watching K-dramas and spend so much time binge-watching, this suggests that there is value in finding new ways to fuel English vocabulary growth. Among the sources of input that have the same motivating properties as graded readers in extensive reading programs is English-subtitled K-dramas.

K-drama English subtitles reached the 2000–3500 word-family levels at 95% text coverage and extended to the 4000–5500 level at 98% coverage subject to genres. This research adopted 12 repetitions as a threshold for incidental word learning. Results showed that 3346 to 6234 words beyond the first 2000 word families have met this criterion if watching 12 to 48 English-subtitled K-dramas. This may be a conservative estimate because viewers may gain partial knowledge for the words encountered fewer than 12 times. Generally, EFL K-drama fans can encounter most of the first 5000 word families often enough for learning to occur through their continually watching English-subtitled K-dramas during four college years. However, their vocabulary size may level off at the first 5000 word families just as they do extensive reading of graded readers. For further vocabulary growth, students can be encouraged to read all sorts of English texts extensively.

Low-intermediate learners may find watching English-subtitled K-dramas a challenge at the onset. Deficient comprehension should be expected when watching the first few episodes. But when learners get used to the subtitling speed, comprehension should improve with more and more episodes being watched. Moreover, visual imagery assists learners in creating the link between L2 form and L1 meaning and therefore makes up for insufficient lexical knowledge. There are still some other ways of overcoming fear of reading English subtitles. Gaining background knowledge through trailers and plot introduction aids comprehension. Watching the L1-subtitled K-dramas is a good support to watching the same K-drama again but in English subtitles. With background knowledge from mother-tongue subtitles, low-intermediate learners may find it less difficult in the second time viewing. Another way to increase the potential for lexical learning is to encourage them to watch the same English-subtitled K-drama multiple times if they enjoy it. This would increase the number of encounters with unfamiliar words to the point where they may be learned.

The goal of this research has been twofold: to measure the vocabulary thresholds of English subtitles for EFL K-drama fans to begin with and to estimate the potential vocabulary growth for them during college years. The results may serve as a reference for freshman English teachers who are concerned with sustained extensive reading.

Lastly, this research has been worked within a narrow focus on individual words. Despite the potential learning of the most frequent 5000 word families, concealed in the first 5000 word families are multiword expressions, which are worth investigation but beyond the present focus. Further research may follow the present study by switching the focus from single words to multiword expressions such as collocations, non-compositional phrasal expressions and idioms.

References

Anthony, L. (2022). *AntWordProfiler* (Version 2.0.1) [Computer software]. Waseda University. https://www.laurenceanthony.net/software

Claridge, G. (2012). Graded readers: How the publishers make the grade. *Reading in a Foreign Language, 24*(1), 106–119. http://hdl.handle.net/10125/66668

Dang, T. N. Y., & Webb, S. (2016). Evaluating lists of high-frequency words. *ITL International Journal of Applied Linguistics, 167*(2), 132–158. https://doi.org/10.1075/itl.167.2.02dan

Dang, T. N. Y., Webb, S., & Coxhead, A. (2022). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research, 26*(4), 617–641. https://doi.org/10.1177/1362168820911189

Day, R. R., & Bamford, J. (1998). *Extensive reading in the second language classroom*. Cambridge University Press.

Day, R. R., & Bamford, J. (2004). *Extensive reading activities for teaching language*. Cambridge University Press.

Dwyer, T. (2017). *Speaking in Subtitles. Revaluing Screen Translations*. Edinburgh University Press.

Hirsh, D., & Nation, I.S.P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8*(2), 689–696. http://hdl.handle.net/10125/67046

Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *The Canadian Modern Language Review, 61*(3), 355–382. https://doi.org/ 10.1353/cml.2005.0018

Horst, J. (2013). Context and repetition in word learning. *Frontiers in Psychology, 4*(149), 1–11. https://doi.org/10.3389/fpsyg.2013.00149

Horst, J. S., Parsons, K. L., & Bryan, N. M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology, 2*, 1–11. https://doi.org/10.3389/fpsyg.2011.00017

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430. http://hdl.handle.net/10125/66973

Jin, D., & Yoon, T. (2017). The Korean wave: Retrospect and prospect–Introduction. *International Journal of Communication,11*, 2241–2249. https://ijoc.org/index.php/ijoc/article/viewFile/6296/2047

Kim, Y. (2013). *The Korean wave: Korean media go global*. Routledge, Taylor & Francis Group.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M.

Nordman (Eds.), *Special language: From human thinking to thinking machines* (pp. 316–323). Multilingual Matters.

Laufer, B. (2020). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *TESOL Quarterly, 54*(4), 1076–1085. https://doi.org/10.1002/tesq.3004

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30. http://hdl.handle.net/10125/66648

Lee, S., & Nornes, M. (2015). (Eds.). *Hallyu 2.0: The Korean wave in the age of social media.* University of Michigan Press.

Markham, P., & Peter, L. (2003). The influence of English language and Spanish language captions on foreign language listening/reading comprehension. *Journal of Educational Technology Systems, 31*(3), 331–341. https://doi.org/10.2190/BHUH-420B-FE23-ALA0

Nagy, W., Anderson, R., Schommer, M., Scott, J. A., & Stallman, A. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24*(3), 262–282. https://doi.org/10.2307/747770

Nation, I. S. P. (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, I. S. P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language, 26*(2), 1–16. http://hdl.handle.net/10125/66881

Nation, I. S. P. (2015). Principles guiding vocabulary learning through extensive reading. *Reading in a Foreign Language, 27*(1), 136–145. http://hdl.handle.net/10125/66705

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing.* John Benjamins.

Nation, I. S. P. (2017). *The BNC/COCA Level 6 word family lists* (Version 1.0.0) [Data set]. http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Nation, I. S. P., & Anthony, L. (2013). Mid-frequency readers. *Journal of Extensive Reading, 1,* 5–16. https://lextutor.ca/conc/eng/nation_anthony_2013.pdf

Nation, I. S. P., & Meara, P. (2010). Vocabulary. In N. Schmitt (Ed.), *An introduction to applied linguistics* (2nd Edition) (pp. 34–52). Edward Arnold.

Netflix. (2013). *Netflix declares binge watching is the new normal.* https://www.prnewswire.com/news-releases/netflix-declares-binge-watching-is-the-new-normal-235713431.html

Pedersen, J. (2019). Fansubbing in subtitling land: An investigation into the nature of fansubs in Sweden. *Target, 31*(1), 50–76. https://doi.org/10.1075/target.18017.ped

Rodgers, M. P. H., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly, 45*(4), 689–717. https://doi.org/10.5054/tq.2011.268062

Rubenking, B., & Bracken, C. C. (2021). Binge watching and serial viewing: Comparing new media viewing habits in 2015 and 2020. *Addictive Behaviors Reports, 14,* 100356. https://doi.org/10.1016/j.abrep.2021.100356

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics, 34*(4), 457–479. http://dx.doi.org/10.1093/applin/ams074

Waring, R., & McLean, S. (2015). Exploration of the core and variable dimensions of extensive reading research and pedagogy. *Reading in a Foreign Language, 27*(1), 160–167. http://hdl.handle.net/10125/66708

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2), 130–163. http://hdl.handle.net/10125/66776

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language, 20*(2), 232–345. http://hdl.handle.net/10125/66826

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

Webb, S., & Rodgers, M. P. H. (2009a). Vocabulary demands of television programs. *Language Learning, 59*(2), 335–366. https://doi.org/10.1111/j.1467-9922.2009.00509.x

Webb, S., & Rodgers, M. P. H. (2009b). The lexical coverage of movies. *Applied Linguistics, 30*(3), 407–427. https://doi.org/10.1093/applin/amp010

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL-International Journal of Applied Linguistics, 168*(1), 34–70. https://doi.org/10.1075/itl.168.1.02web

About the Author

Wenhua Hsu is currently teaching at I-Shou University in Taiwan. She holds a PhD in the field of ESP from the University of Essex, UK and MBA from Kansas State University, USA. Her interest includes frequent academic/sub-technical and lay-technical vocabulary as well as lexical bundles in specialized fields.