Analyzing genre-specific words of American TV dramas

Analyzing genre-specific words of American TV dramas

Jared Baierschmidt
Junko Noguchi

Analyzing genre-specific words of American TV dramas

Jared Baierschmidt
Junko Noguchi

Literature review

Benefits of using television programs for English vocabulary learning

TV programs are a popular form of effective learning material for many English learners (Lin & Siyanova, 2014; Webb and Rodgers, 2009). One of the most prominent benefits of using TV programs for English learning is incidental vocabulary learning (Lin & Siyanova, 2014; Webb & Rodgers, 2009). In fact, in d'Ydewalle and Van de Poel's study (1999), superior incidental vocabulary learning was reported compared with syntax and grammar. Also, repeated and continuous viewing helps learners because, according to Schmidt and Carter (2000), "due to the incremental nature of vocabulary acquisition, repeated exposures are necessary to consolidate a new word in the learner's mind" (p. 4). To gain the benefits of repeated exposure, regular watching of at least an hour of television a day is recommended by Webb and Rodgers (2009). Narrow viewing is also helpful (Schmidt et al, 2000; Webb & Rodgers, 2009). Narrow viewing refers to watching videos on the same topic over the course of several videos. There are many advantages from this type of viewing. First, narrow viewing helps viewers to become familiar with the topic and gradually build background knowledge for watching future episodes. Also, topic-related key words tend to reoccur, which facilitates vocabulary learning. The enhanced familiarity and acquisition of frequently used words eases the lexical burden on viewers and therefore

frees up cognitive capacity for understanding the content or learning other words (Hwang & Nation, 1989).

In addition, in Rott's (1999) study on the role of reading in language learners' incidental vocabulary acquisition, he found that six encounters resulted in a significant effect on "more receptive as well as productive word knowledge (p. 604)." Furthermore, when combined with other contextual factors such as audio or visual cues that enhance the learning condition, the vocabulary meanings may be acquired even faster (Webb, 2008; Webb, 2010). These results may carry over to learning through watching TV shows (Rodgers & Webb, 2011). In order to conduct narrow viewing, learners need to select TV shows to watch based on their related topics or genres.

Genre plays a key role in effective learning from TV shows

Genre is the most important element that affects vocabulary learning from TV shows because the vocabulary demands of television programs vary depending on the genre (Webb & Rodgers, 2009). Webb and Rodgers (2009) investigated vocabulary coverage and the number of encounters of low frequency vocabulary in 88 television programs including TV dramas. They found that when watching TV dramas, knowledge of the most frequent 3,000 word families as well as proper nouns and marginal nouns (PNAMW) provided 95% coverage and a vocabulary of the most frequent 6,000 word families and PNAMW reached 98% coverage. However, their definition of "drama" was arbitrarily broad, including four quite different types of shows. The TV programs used in the study were *The West Wing*, *The Sopranos*, *C.S.I.*, and *24*, whose topics range from politics, gang life, investigative crime work and terrorism.

On the other hand, Webb (2010) used a sub-genre of the drama domain (i.e. medicine) in order to look at two TV shows that are situated in a very similar context, which enabled the researcher to investigate the transcripts more closely. The TV programs selected were

House and *Grey's Anatomy*, both being medical dramas that depict doctors' lives revolving around what goes on in hospitals. In the study, glossaries consisting of setting-specific word families, which are low-frequency word families which occurred 10 or more times in the studied episodes, were created. Changes in the 95% and 98% coverages by adding the words provided by the glossaries were also analyzed and compared. The results showed that there were many low-frequency word families that appeared in both of the TV shows and indicated that it is likely that related TV shows share many genre-specific words. Thus, Webb (2010) recommended that it might be useful for learners to have such a list of genre specific words when watching these programs.

The question raised then is whether or not only certain genres share a large portion of such words, making it worthwhile to conduct narrow viewing or create glossaries. Because it requires a lot of time and energy to research and select related TV shows and make glossaries, this question of cost-efficiency should be answered. Therefore, inspired by the studies that have studied vocabulary coverage in TV genres (i.e. Webb & Rodgers, 2009; Webb, 2010), the present study addresses the following questions:

1. Are there any differences among American TV dramas in terms of the vocabulary coverage?

2. Are there any differences among American TV dramas in terms of genre-specific words?

Method

Materials

The transcripts of 30 episodes of 6 different American television shows were analyzed in total. The details of the shows are presented in Table 1. Two TV shows that share similar settings (location, time, age of characters, etc.) and are considered to belong to the same television genre (i.e. drama) and sub-genre (i.e. legal drama, school drama, family drama,

supernatural drama) were compared. It was assumed that they contain similar vocabulary due to their similarity in the topics that they are dealing with (Rodgers & Webb 2011), which is the rationale behind narrow viewing (Webb, 2010).

They were also selected according to their availability, degree of similarity, running time, and date when first aired. Transcripts of the first 5 episodes of each TV show were downloaded from the Internet for analysis. Words that were not spoken such as stage commands, storyline, and speakers' names were removed from the transcripts. Hyphenation was removed from hyphenated words and these words were kept as separate since it is highly likely that the presence of hyphenation in the orthography does not influence comprehension (Grant & Bauer, 2004).

Analysis

Analysis of the transcripts was done using the Range program (Heatley, Nation, & Coxhead, 2002), which was downloaded from Paul Nation's website (www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx). There are three different versions of the program based on what word lists will be used for frequency analysis and the one that uses Nation's (2004) fourteen 1,000-word lists was chosen for this study. One of the reasons is the ease of comparison because other similar studies used this version, such as Webb and Rodgers (2009) and Webb (2010). The program shows the number of times each word occurred, and the 1,000 word level (1,000–14,000) at which the words occurred according to Nation's (2004) fourteen 1,000-word lists, which are based on the frequency and range of occurrence of word families in the British National Corpus (BNC). In the list, level 6 word families, according to Bauer and Nation's (1993) word family classification, are used, which include inflections and over 80 derivational affixes All word stems were free forms not bound forms (Webb, 2010).

Table 1

Six TV dramas used in this study categorized by sub-genre.

Sub-genre	Legal	Supernatural	Family
Title	Boston Legal	Supernatural	Fuller House
	Ally McBeal	Shadow hunters	Modern Family

Procedure

The transcripts of 10 episodes in each sub-genre were analyzed separately to calculate the word family level at which the cumulative coverage reaches 95% coverage, which Webb and Rodgers (2009) suggest is sufficient for comprehension of television programs. The results for each genre were compared to see if there were any differences across sub-genres. As in Webb and Rodgers (2009), the coverage of the proper nouns and marginal words (e.g., *ah*, *oh*, *huh*) were included in the cumulative coverage because such words are regarded to be more easily learned than typical word families (Nation, 2006) and should be recognizable for most learners with the knowledge of the 3,000 most frequent words (Webb & Rodgers, 2009). The results produced by Range were examined to find the most frequent word families from the 4,000 to 14,000 word levels and *not in the lists* (less frequent than the 14,000 word level), the number of times those word families were encountered, and the coverage of those words.

Also, the number of setting-specific word families for each genre that are less frequent than the 4,000 word level and were encountered 4 or more times in a set of 10 episodes for each genre was counted. These criteria were created in reference to Webb (2010) as well as the number of occurrences needed for incidental learning to occur. Webb included in his glossaries those words that reoccurred approximately once every 2.5 episodes. Thus, with 10 episodes for each genre, we adopted a criteria of "4 or more times in 10." Also, the number of those words that occurred in both of the TV shows in the same sub-genre

was counted to see how many of these genre-specific words appear across the two TV shows in the same sub-genre. Furthermore, in order to examine how many of those genre-specific words would recur frequently enough to be incidentally acquired, the number of words that appeared more than 6 times was counted for each genre based on the findings by Rott (1999) that more than 6 times of exposure to unknown words will likely contribute to incidental learning. The number of words that occurred more than 5 times was also counted because it may be possible for learners to learn the meaning of unknown genre-specific words more quickly after exposure from enhanced input with contextual cues as well as audio-visual support from TV shows (Webb, 2008; Webb, 2010).

Results

Research question 1: Vocabulary coverage

　　Legal dramas: *Boston Legal* and *Ally McBeal*. Table 2 shows tokens, types, word families and cumulative coverage, with proper nouns and marginal words at each word level for legal dramas (*Boston Legal* and *Ally McBeal*). Looking at the third column of Table 2, row 17 shows that the percentages of words not in the lists was 0.9%, accounting for the sixth highest percentage of tokens. This means that a relatively large chunk of data were not categorized into the fourteen 1,000-word families and implies that extended lists based on a larger corpus with a wider variety might be appropriate for more sophisticated and precise analysis. For such analysis, Range which is based on the BNC/COCA word family lists may be useful since it consists of 29 word family lists, which is larger than the number of word families of the Range used in this study. Proper nouns and marginal words (PNAMW) were added to the text coverage needed to know each 1,000-word list, assuming that it is highly likely that they are already known or have a minimal learning burden (Nation, 2006).

Table 2

Tokens, types, word families and cumulative coverage, with proper nouns and marginal words at each word level for legal dramas (Boston Legal and Ally McBeal)

	Tokens		Types		Word	Coverage including proper nouns and marginal words
Word list	Raw	%	Raw	%	Families	
1,000	42,397	86.54	1,846	39.5	895	88.71
2,000	2,236	4.56	887	18.98	611	93.27
3,000	904	1.85	453	9.69	356	95.12[a]
4,000	727	1.48	303	6.48	247	96.6
5,000	386	0.79	202	4.32	165	97.39
6,000	234	0.48	135	2.89	119	97.87
7,000	170	0.35	113	2.42	100	98.22[b]
8,000	150	0.31	86	1.84	76	98.53
9,000	71	0.14	56	1.2	52	98.67
10,000	57	0.12	38	0.81	38	98.79
11,000	57	0.12	38	0.81	37	98.91
12,000	31	0.06	28	0.6	24	98.97
13,000	29	0.06	19	0.41	19	99.03
14,000	35	0.07	24	0.51	22	99.1
Proper nouns	896	1.83	150	3.21	150	
Marginal words	166	0.34	12	0.26	3	
Not in the lists	443	0.9	283	6.06		
Total	48,989	2.17	4,673	9.53	2,914	100

a Reaching 95 % coverage, b Reaching 98 % coverage

Supernatural dramas: *Supernatural* and *Shadowhunters*. Table 3 shows tokens, types, word families and cumulative coverage, with proper nouns and marginal words at each word level for supernatural dramas (*Supernatural* and *Shadowhunter*). In answer to the first research question, with knowledge of PNAMW the vocabulary necessary to reach 95% coverage when watching the supernatural thrillers is 3,000 word families while a vocabulary of 10,000 word families plus PNAMW is necessary to reach 98% coverage.

Table 3

Tokens, types, word families and cumulative coverage, with proper nouns and marginal words at each word level for supernatural dramas (Supernatural and Shadowhunter)

Wordlist	Tokens		Types		Word Families	Coverage including proper nouns and marginal words
	Raw	%	Raw	%		
1,000	35585	86.48	1452	41.19	775	89.57
2,000	1588	3.86	634	17.99	456	93.43
3,000	734	1.78	353	10.01	277	95.21[a]
4,000	407	0.99	166	4.71	143	96.2
5,000	209	0.51	113	3.21	97	96.71
6,000	131	0.32	79	2.24	67	97.03
7,000	170	0.41	59	1.67	55	97.44
8,000	50	0.12	39	1.11	35	97.56
9,000	87	0.21	41	1.16	37	97.77
10,000	106	0.26	36	1.02	33	98.03[b]
11,000	36	0.09	29	0.82	29	98.12
12,000	35	0.09	17	0.48	15	98.21
13,000	18	0.04	15	0.43	15	98.25
14,000	37	0.09	10	0.28	9	98.34
Proper nouns	1016	2.47	137	3.89	137	
Marginal words	255	0.62	18	0.51	4	
Not in the lists	685	1.66	327	9.28		
Total	41149	100	3525	100	2184	100

a Reaching 95 % coverage, b Reaching 98 % coverage

Family dramas: *Fuller House* and *Modern Family*. Table 4 shows tokens, types, word families and cumulative coverage, with proper nouns and marginal words at each word level for family dramas (*Fuller House* and *Modern Family*). In answer to the first research question, with knowledge of PNAMW the vocabulary necessary to reach 95% coverage when watching the family dramas is 4,000 word families. However, 98% coverage was never reached, even after including PNAMW.

Table 4

Tokens, types, word families and cumulative coverage, with proper nouns and marginal words at each word level for family dramas (Fuller House and Modern Family)

Word list	Tokens		Types		Word Families	Coverage including proper nouns and marginal words
	Raw	%	Raw	%		
1,000	27150	85.29	1410	40.34	772	88.25
2,000	1285	4.04	558	15.97	425	92.29
3,000	545	1.71	300	8.58	251	94.00
4,000	314	0.99	180	5.15	158	94.99[a]
5,000	232	0.73	120	3.43	103	95.72
6,000	196	0.62	92	2.63	79	96.34
7,000	98	0.31	68	1.95	60	96.65
8,000	69	0.22	40	1.14	37	96.87
9,000	41	0.13	29	0.83	28	97.00
10,000	66	0.21	27	0.77	26	97.21
11,000	22	0.07	18	0.52	17	97.28
12,000	15	0.05	12	0.34	11	97.33
13,000	24	0.08	17	0.49	15	97.41
14,000	10	0.03	10	0.29	9	97.44
Proper nouns	579	1.82	115	3.29	115	
Marginal words	363	1.14	23	0.66	4	
Not in the lists	823	2.59	476	13.62		
Total	31832	100	3495	100	2110	100

a Reaching 95 % coverage, b Reaching 98 % coverage

Table 5 shows a summary of tokens, word family, cumulative coverage, and percentage of words in "Not in the lists" for all the TV shows analyzed in this study. Judging from the total number of tokens, it seems that learners can get the largest amount of exposure to English from legal dramas followed by supernatural dramas and family dramas, in that order, by watching the same number of episodes. The total number of word families indicates that legal dramas require a wider variety of vocabulary followed by supernatural dramas and family dramas in that order. On the other hand, cumulative coverage indicates that family dramas required the highest word family level for both 95% and 98 %. However, it should be noted that the percentage of words in the category of "Not in the lists" is 2.59% for family dramas whereas that of legal dramas and supernatural dramas are 0.9% and 1.66% respectively.

Table 5

Summary of cumulative coverage, with proper nouns and marginal words for all the TV shows analyzed in this study

	Legal dramas	Supernatural dramas	Family dramas
Total token	48,989	41149	31832
Total word families	2,914	2184	2110
95% coverage	3000	3000	4000
98% coverage	7000	10000	N/A
Percentage of words in "Not in the lists" (%)	0.9	1.66	2.59

Research question 2: Genre-specific words

Table 6 shows the number of words less frequent than the 4,000 word level that occurred 4, 5, and 6 times in the 10 episodes for each genre, which also appeared at least once in each of the sub-genre TV dramas. The number in the parenthesis is the number of such words that occurred 4, 5, and 6 times in the 10 episodes for each genre regardless of whether it appeared in one or both TV dramas. The results indicate that legal dramas have the largest likelihood of using genre-specific words in the same family repeatedly across the genre followed by family dramas and supernatural dramas in that order. This implies that supernatural dramas use the same genre-specific words in the same family repeatedly within a TV show but not necessarily across the genre. The supernatural drama category seems to have the least number of genre-specific words that learners may encounter frequently within the genre.

Table 6

Number of low frequency words that occurred and appeared in both TV shows for each genre (one of the shows)

Number of occurrences	Legal	Family	Supernatural
4	41 (51)	21 (25)	16 (22)
5	28 (34)	15 (19)	14 (18)
6	19 (22)	14 (18)	10 (13)

Discussion

For the first research question, the results indicated that there were large differences between reaching 95% and 98% coverage, as was also found by Webb and Rodgers (2009). Also, the vocabulary demands of American television dramas do not vary much depending on the sub-genre in the word family levels to reach 95% coverage. The results showed that knowledge of the most frequent 3,000-4000 word families plus PNAMW provided more than 95% coverage of all the sub-genres, similar to Webb and Rodgers (2009). However, there were some differences between sub-genres in vocabulary size required to reach 98% coverage. The results show that, for legal dramas, knowledge of the most frequent 7,000 word families and PNAMW provided 98% coverage, which resonates with the findings in Webb and Rodgers (2009) whereas for supernatural dramas, knowledge of the most frequent 10,000 word families and PNAMW provided 98% coverage. On the other hand, family dramas demand the highest level of word family to reach 98% coverage and even with 14,000 word family level it reaches only 97.44%. This seems to be counter-intuitive considering that legal dramas, which contain specialized vocabulary, should include more vocabulary with high frequency than the other genres. One thing that may explain this reversed phenomenon might be that the percentage of words in the category of "Not in the lists" is 2.59% for family dramas, which accounts for the third highest percentage of tokens whereas those of legal dramas and supernatural dramas are 0.9% and 1.66% respectively. Many of the words in this category are heavily colloquial (ex. snazzy, gettable and loopy), cultural, recent (ex. peerenting, texting and wowing) or creative. Some of them are abbreviation such as BFF (Best Friends Forever), OJ (Orange Juice), and PJ (Pajamas). Also, some of them are predominately U.S. phrases. For instance, "diapers" was not included in the list because they are called "nappies" in British English and not included in the BNC word family lists (Nation, 2004) which is based on British English and a mostly written corpus. As a result, the data implies that it is not possible to reach 98% coverage

with vocabulary knowledge of the 14,000-word family level because more than 2% of the coverage is out of the range of 1-14 1000 word families by default. Hence, it is likely that whether learners can reach 98% coverage or not is closely related to how much the learners are familiar with American colloquial expressions and especially pop culture, which family dramas seem to reflect to a large degree.

Conversely, it might be that legal dramas have the least percentage of words "Not in the list" because the words used in the genre have many overlaps with the kind of words included in the BNC fourteen 1000-word lists. In other words, accuracy of the Range output may depend on how similar the text to be analyzed is to the texts used to create the BNC corpus. If there is a large mismatch between the kinds of words used in the texts to be analyzed and the BNC word family lists, a large portion of the texts will not be categorized into frequency families, making the results imprecise and requiring caution when interpreting them. This especially applies when examining the vocabulary size necessary to gain 98% coverage. This is because from the 4th 1000-word family level, which usually accounts for less than 1% of the vocabulary, small percentages make a significance impact. In other words, the percentage of words in "Not in the lists" influences whether the 98% coverage level can be reached.

Reaching 98% coverage may be a difficult goal for many average learners to begin with (Webb & Rodgers, 2009), especially those learners in an EFL context with scant L2 input outside of the classroom, considering the fact that it requires at least knowledge of the most frequent 7,000 word families. Due to this as well as the possibility of inaccuracy in determining the vocabulary size necessary for adequate comprehension, the vocabulary size necessary for adequate comprehension with 95% coverage may be a better indicator to use to investigate vocabulary demands for learners to understand American TV shows.

Also, when analyzing American TV shows it is better to use an analysis tool based on vocabulary lists that use a corpus that includes mainly American spoken forms from different contexts, such as the version of Range based on BNC and COCA.

As for the second research question, the results indicate that legal dramas have more established technical terms that are commonly used across the TV shows across the genre compared to supernatural dramas or family dramas. This makes sense because technical terms are used in a relatively strict and consistent way in the legal industry. Thus glossaries for technical words that are infrequent in general but commonly used in such genres, similar to the ones created by Webb (2010) for medical TV shows, will be useful when narrow viewing TV series of the legal sub-genre.

On the other hand, the results indicates that supernatural dramas seem to use the same genre-specific words repeatedly within the same TV show but not necessarily across the genre. Among all three sub-genres reported in this study, supernatural dramas have the most creative elements and thus there are many uniquely used words for the particular shows in addition to genre-specific terms. *Shadowhunters* especially included words unique to the show such as "shadowhunters," "parabatai," "downworlder," and "mundanes." Supernatural dramas seem to have the least number of genre-specific words that learners may encounter across the genre. In addition, there are a few words that occur across more than one TV show that are likely strongly represented in all American dramas such as "awesome," "buddy," "cute," "dude," "girlfriend," "hug," and "kiss".

Conclusion

TV programs are popular English learning materials (Lin, 2014) and research studies show that it is important to have a vocabulary size of at least the most frequent 3,000 word family as well as knowledge of relevant proper nouns and marginal words in order to have a general understanding of the contents of TV programs (Webb & Rodgers, 2009). As this

study has shown, language learners may require additional vocabulary support if required or encouraged by a teacher to attempt narrow viewing of particular sub-genres. Medical and law-related shows seem to require the least amount of scaffolding, with genre-specific terms frequently appearing across TV series. Supernatural and family dramas, on the other hand, may require show-specific glossaries. In particular, teachers need to be aware that family dramas use many colloquial expressions and references to popular culture, which students may need a lot of help with understanding.

References

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, *6*, 253–279.

d'Ydewalle, G., & Van de Poel, M. (1999). Incidental foreign-language acquisition by children watching subtitled television programs. *Journal of Psycholinguistic Research*, *28*, 227–244.

Grant, L. & Bauer, L. (2004). Criteria for re-defining idioms: Are we barking up the wrong tree?". *Applied Linguistics*, *25*(1), 38–61.

Heatley, A., Nation, P. & Coxhead, A. (2002). RANGE [Computer software].Retrieved from http:// www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Hwang, K. & Nation, I. S. P. (1989). Reducing the vocabulary load and encouraging vocabulary learning through reading newspapers. *Reading in a Foreign Language,* *6*(1), 323–335.

Lin, P. M. S., (2014). Investigating the validity of internet television as a resource for acquiring 2 formulaic sequences. *System, 42,* 164–176.

Lin, P. M. S., & Siyanova, A. (2014). Internet television for L2 vocabulary learning. In D. Nunan & J. C. Richards (Eds.), *Language learning beyond the classroom* (pp. 149-158). London: Routledge.

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 3–13). Amsterdam: John Benjamins.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening?. *The Canadian Modern Language Review*, *63*, 59–82.

Rodgers, M. P. H. & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly 45*(4), 689-717.

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition through reading. *Studies in Second Language Acquisition*, *21*, 589–619.

Schmitt, N., & Carter, R. (2000). The lexical advantages of narrow reading for second language learners. *TESOL Journal, 9*(1), 4-9.

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language, 20*(2), 232–245.

Webb, S., & Rodgers, M. P. H. (2009). The vocabulary demands of television programs. *Language Learning*, *59*, 335–366.