Concordia Working Papers in Applied Linguistics, 2, 2009

What's in a Television Word List? A Corpus-Informed Investigation

Katie MacFadden
Concordia University

Katherine Barrett
Concordia University

Marlise Horst
Concordia University

Abstract

Watching TV can be a challenge for ESL students, because they do not know enough vocabulary to be able to understand the show. The goal of the study was to identify, analyze and compile a list of specialized vocabulary that occurred frequently in a 500,000-word TV corpus and to determine whether this list could potentially boost learners' comprehension of TV talk and thereby support the learning of new words. A 689-word Television Word List (TWL) was created following procedures used by Coxhead (2000). In comparing the TWL to the other lists, the TWL appears to be a useful specialized list that provides 1 to 2% coverage of TV talk depending on the particular show. The study also identifies some potentially problematic words that occur frequently on TV. Implications for research and pedagogy are discussed.

Teachers of second or foreign languages have always given their students homework but recently, watching television has become a popular assignment in many language classes. Teachers assume that television is a useful source of authentic language input, and that the viewer is likely to acquire new linguistic features. In this study we focus on the potential of watching TV for the development of new second language (L2)

vocabulary. The idea that new vocabulary can be learned incidentally through comprehension-focused processing of L2 input is well established (Nation, 2001); however, this evidence is based largely on studies of reading (e.g. Horst, 2005; Waring & Takaki, 2003; Webb, 2005). Little research that we know of has addressed the extent to which learners acquire new word knowledge through exposure to spoken input of media such as movies or television. However a number of studies (discussed below) have considered the vocabulary knowledge a L2 learner would need to have in order to be able to understand authentic language input well enough to be able to learn the unknown vocabulary items he or she encounters in it. Obviously, being able to understand a native speaker conversation or TV show in one's L2 can be a challenge if one does not have the required vocabulary for comprehension; adequate understanding of the input seems a logical pre-condition for learning any new items in it.

Listening comprehension is more than recognizing the meanings of the words that occur in spoken material such as the dialogue of a television show. Vocabulary knowledge is recognized as one dimension among many in a complex psycholinguistic process in which listener, language and input variables interact. In comprehending speech, the listener processes the speech signal, segments it into strings of words, and parses the strings syntactically. In her discussion of an integrated model of L2 reading comprehension, Bernhard (2005) points out that understanding L2 input differs from L1 comprehension in that L1-based familiarity factors also play a role (e.g. the availability of cognates) – in addition to the important effect of knowledge of L2 vocabulary and syntax. Listening comprehension is also affected by input variables such as the connectedness of the speech and social and regional variation in pronunciation. Such variation may make understanding L2 speech particularly difficult if it diverges substantially from the norms students are familiar with (Major, Fitzmaurice, Bunta, & Balasubramanian, 2005). We believe that vocabulary knowledge is important in understanding spoken input and that studying a principled list of frequently used television words will be useful in making it more comprehensible, but it is clear that other factors are also relevant.[1]

A study by Laufer (1989) set out to explore the connection between L2 vocabulary knowledge and comprehension of unsimplified input designed for native speakers. She found that in order to comprehend a written text adequately, the learners she investigated needed 95% coverage. This suggests that to be able to read and understand a text, the meanings of at least 19 in every 20 words must be known to the L2 reader.

The less coverage one has, the more difficult it is to understand the passage. The 95% figure has been confirmed in a series of reading comprehension studies in a variety of contexts; (see Nation, 2001 for an overview of this work). In a 2006 study, Nation set out to determine the number of word families a learner of English would need to know in order to meet this criterion. A word family is defined here as a root word and basic inflected and derived forms; thus it is assumed that a learner who knows the word family of happy would also know happily, unhappy and happier. Using frequency lists based on the British National Corpus, Nation determined that in the case of most written texts, the 95% coverage criterion can be reached with the knowledge of 8000 to 9000 frequent English word families. Coverage of spoken texts, which generally tend to use fewer unusual words, reaches the 95% known word criterion if the hypothetical learner has a vocabulary size of 6000 to 7000 families. Clearly the attainment of an L2 English vocabulary size on this order represents a considerable challenge. As Laufer (2000) has shown, vocabulary sizes of L2 learners beginning university studies in English are often well below these figures (on the order of 2000-3000 families) even after years of study. Television is arguably a particularly supportive type of oral text. It is possible that with the help of visuals, an intermediate level ESL learner viewer may be able to follow a story line of a TV show even if he or she does not know 95% of the words that are spoken. But with so much of the linguistic material incomprehensible or difficult to process, many features in it will likely not be noticed (or learned).

The challenge in the case of television has been delineated more closely in a recent study by Webb and Rodgers (2009). They collected a corpus of 88 transcripts of both British and American shows totaling 264,384 words. From this corpus they were able to show that an L2 learner needs to know a minimum of 3000 word families plus many proper nouns and marginal words in order to understand TV (marginal words are items characteristic of spoken interaction such as *um*, and *uh-huh*). However, without the knowledge of proper nouns (many of which may be difficult for learners unfamiliar with North American culture) and marginal words, they conclude that the learner would need to be able to understand at least 12,000 word families – a vocabulary size figure associated with very advanced learners (Nation, 2001). In sum, this research suggests that television may be a great deal more difficult to comprehend than most ESL teachers realize and that learners who have not yet attained advanced levels of proficiency may not be in a good position to reap many language learning benefits from watching it.

One solution to the problem of providing beginning and intermediate L2 learners with input that they can readily comprehend (and learn from) is the development of simplified materials. To this end, publishers such as Oxford, Cambridge and Longman have created large collections of graded readers for ESL learners at a range of proficiency levels. To our knowledge, this effort has not been matched by large scale production of systematically graded listening or video materials (though some graded readers are also available in audio format). An alternative approach to the problem of providing learners with input they can comprehend has been to help them attain the limited amounts of vocabulary needed to comprehend particular genres. For instance, Coxhead (2000) investigated the vocabulary that is needed to understand university textbooks. She realized that 95% coverage could be difficult for a student beginning university in a second or foreign language to attain. In her view, students entering university could be expected to already know the 2000 most frequent words of English on West's (1953) General Service List, but doubted that these were enough to equip learners of English to read university materials designed for native speakers. She set out to identify an additional set of words that recurred frequently in a large corpus of academic texts, which represented a wide variety of subject areas. This resulted in a list of 570 word families known as the AWL (*Academic Word List*). Coxhead's analyses showed that students could attain 95% coverage of academic texts with knowledge of just the 2000 most frequent English words and the 570 words of the AWL. This meant that the foreign students entering university in English would be able to understand and read the texts books necessary for their classes without too much of a struggle.

Some researchers (Hyland & Tse, 2007) have argued that Coxhead's coverage figures are optimistic and shown that some academic subjects (e.g. law) are better covered by the list than others (e.g, science); nonetheless, the AWL stands as a very useful list of words for university-bound learners of English to know. Attaining knowledge of the thousands of frequently used 'general' English words – 8000 or 9000, according to Nation (2006) – that enable effective L2 reading comprehension is likely to be a process of many years. By focusing on the vocabulary needed to understand a specific genre, Coxhead's 570-word AWL offers the academic learner an efficient shortcut through that lengthy process.

We decided to take a similar approach to investigating the language of television. We wondered whether it would be possible to identify a list of words that recur frequently in this genre that would offer learners a high

level of coverage, assuming as Coxhead (2000) did, that they already know the 2000 most frequent families. If such a list could be identified, it could offer a useful alternative to having to learn all of the 12,000 word families or the long list of marginal words and proper nouns that Webb and Rodgers (2009) identified as necessary to reach 95% coverage of the vocabulary used in TV shows. We were also interested in the kinds of words learners might hear frequently on TV. To discover what sort of vocabulary learners encounter while watching typical North American drama and sitcom television shows, and whether there is a potentially useful set of words that recur frequently, we gathered and examined a large corpus of television talk. We first describe the development and validation of the word list. Then we outline our test of its coverage powers and look more closely at some of the words that are very likely to be learned by TV viewers due to their frequent occurrence across many shows. The findings will hopefully provide useful insights into the vocabulary learning potential of watching TV that will be of use to course designers, teachers and learners.

RESEARCH QUESTIONS

The primary aim of this research was to determine whether a Television Word List (TWL) could be compiled. Assuming a TWL can be identified, we would then be interested in validating this list. In order to claim that the list is truly reflective of television content, we would need to demonstrate that it differs substantially from a list derived from a different genre (e.g. an academic corpus). By the same logic, it should bear demonstrable resemblances to a list derived from a different corpus of the same genre (i.e., another TV corpus). A valid television list should also represent various kinds of television content in equal measure. These validity issues are addressed in the second question. The final question puts the list to the crucial test, asking what it can deliver in terms of coverage of the TV corpus. The questions are as follows:

1. Can a list of words which occur frequently across many television shows be identified?
2. If so, does the list a) differ from a non-TV-based list, b) resemble another list that is TV-based, and c) represent both comedy and drama sub-genres?

3. What coverage does this list of words provide? Will knowing the words on the list enable students with knowledge of only the 2000 most frequent words of English to reach a high level of coverage of the vocabulary they encounter on TV?

METHODOLOGY

The television corpus used in this study was established through the combined efforts of Applied Linguistics graduate students at Concordia University. The corpus contains 10 popular TV shows – five comedies and five dramas – that the graduate students, in the corpus linguistics course, deemed to be typical of what learners might be asked to watch as part of their language enrichment homework. The five comedies were: *How I Met Your Mother*, *The Office*, *Seinfeld*, *Two and a Half Men* and *Frasier*. The five dramas were: *Alias*, *Desperate Housewives*, *Grey's Anatomy*, *Lost* and *Prison Break*. The corpus material is narrative; news, commentaries or talk shows were not included. The sub-corpora from the 10 shows were compiled by downloading transcripts freely available on the internet; stage prompts and other non-spoken material in the transcripts were deleted manually. Each of the 10 show corpora amounted to around 50,000 words; the number of episodes represented in each ranged from 11 to 18 (due to differences in show length and amounts of talk that occurred in them). In total the corpus contained approximately 500,000 words in roughly equal halves, i.e., the comedy and drama sub-corpora amounted to about 250,000 words each.

In order to create the proposed list of words it was necessary, as done previously by Coxhead (2000), to identify word families that were used often and across many of the various sub-corpora. Coxhead (2000) looked at texts from 28 academic subject areas and chose words that occurred in at least 15 of these 28 subject areas; in other words, she judged that a word had adequate 'range' if it occurred in 15 of 28 – or approximately 50% – of her sub-corpora. So to create a list of vocabulary words for the TV corpus, we used a similar criterion: a word was considered to be well represented across the TV corpus if it was mentioned in at least 5 of the 10 shows. Another important criterion for the list was that it not include items from West's (1953) GSL lists of the 2000 most frequent English families; we assume that the intermediate-level learners who might find the list useful already know these basic words. The third criterion was that the word must be frequent. In order to arrive at a list of a manageable size,

'frequent' was defined as occurring at least seven times the entire corpus. Once these three criteria were established, each of the show corpora was entered individually into the Range program available at Cobb's Lextutor website. This software program, which is specifically designed to extract lists from corpora, allows the user to specify range and frequency criteria as well as lists to be excluded from the analysis (e.g. the GSL 2000).

In answering the second research question, which addresses the validity of the list, we used corpus analysis tools available at Lextutor. To test whether our television list differed from an academic list, we used the site's VocabProfile program. This lexical frequency profiling tool determines the proportions of GSL 1000, GSL 2000, AWL, and other words in a submitted text. The TWL was then compared with another television list, based on the US TV Talk corpus (a 2-million-word collection of transcriptions of broadcasts dating from the 1980s and 1990s available online). Lextutor's concordancing tool was then used to identify frequencies of various words on the television lists in the two corpora.

In addressing the third research question about the proportion of the TV corpus that was covered by our TWL, we entered each of the 10 TV subcorpora into a customized version of the VocabProfile program. Due to the large size of the TV corpus it was not possible to enter it into the program as a whole. In order to arrive at figures that reflect coverage of the entire television corpus, we used the individual sub-corpus coverage percentages to calculate mean coverage percentages for the five comedies and the five dramas.

RESULTS

Following the procedures outlined above, we were able to identify a list of 689 word families that occurred frequently in the corpus as a whole (see the Appendix for the entire list). Therefore, the initial answer to the question of whether a television word list could be created based on the presented methodology appears to be positive. The presence of items such as *guy*, *okay*, *damn* and *dude* on the list seem to be an accurate reflection of television interactions and inspire confidence in the specialized character of the list.

The results of the first test of the TWL's validity are shown in Table 1. It was hypothesized that the television-based TWL would have little in common with a list based on a very different genre such as a corpus of university textbooks. The data in Table 1 confirm this; the overlap

between Coxhead's *Academic Word List* (AWL) and the TWL is rather small with fewer than 30% of the items shared. The finding that most (70%) of the TWL words do not occur frequently in written academic texts confirms its distinctive spoken character. However, one might expect the television list to have much less in common with the AWL. The fact that the two list overlap as much as they do may be explained by the fact that the some of the television shows revolve around topics that can be seen as academic, e.g. *Fraser* with its psychiatry theme or *The Office* with its business setting.

Table 1. Distribution of the Television Word List (TWL)

Word List	Number of Words	Percentage
AWL	203	29.46%
Other (not 1K, 2K or AWL)	486	70.54%

The second validity test involved comparing the TWL to a second list created using the same methodology but based on another comparable corpus, the US TV Talk corpus. In this case, it was expected that the two lists – both based on TV corpora – would be similar, thereby confirming the specialist character of the TWL. Table 2 shows the 20 most frequently occurring words on the lists in order of frequency. Both lists feature *okay* as a highly frequent word, and four other words (*guy, kid, Jack* and *couple*) are shared between them. Otherwise, there is not as much congruence as might be expected.

Table 2. TWL and US TV Talk Lists – Top 20 Word Families

Television Word List	US TV Talk
okay	anytime
guy	twin
alright	okay
hell	mama
kid	grab
Jack	tape
job	television
sex	guy
Jerry	major
damn	stare
crazy	counter
couple	Jack
surgery	apartment
dude	couch
code	booking
wed	peanut
ass	kid
whoa	jacket
honey	hallway
kidding	couple

As an additional check on validity, we used the concordancer at *Lextutor* to determine how often the 13 most frequent TWL words (names *Jack* and *Jerry* excluded) occurred in our TV corpus and the US TV Talk corpus. Although the two corpora differ greatly in size (US TV Talk is four times larger), we surmised that the pattern of results might be similar. These frequency findings are shown in Figure 1. The results indicate some similarities; *okay* is more frequent than *guy* in both corpora and both of these are more frequent than the other 11 words. However, most striking in this chart are the large differences in distributions in the two corpora. Words that occur hundreds of times in our TV corpus occur far less frequently in the much larger US TV Talk corpus.

This lack of similarity between the two lists may be explained at least in part by problems with the US TV Talk corpus. In our concordancing work with this corpus, we noticed that names of actors, show titles, and other extraneous material have not been stripped out. It also appears to contain transcripts of material that had been broadcast over a particular period of

time, rather than a selection of specific shows. We also discovered that it contained commercials as well as some British shows; these problems along with its age (none of the shows are current) suggest that the two corpora may not be truly comparable.

Another validity question concerns the extent to which the TWL represents words that occur in the types of show we investigated, dramas and comedies. As we have specified, in order to be included in the list, words had to be in at least five of 10 shows, repeated at least seven times, and not part of the GSL 2000, but the extent to which a word featured in either the drama or comedy sub-corpus was not taken into consideration in the selection process. This issue is important because if the list were found to over-represent one sub-genre at the expense of the other, the overall usefulness of the list would be compromised and there might be reason to divide the list in two. Table 3 indicates that the distribution was fairly even. Only one word (*slash*) occurred in all five comedy corpora but not in any of the dramas. Similarly, only one word (*assault*) occurred in all five drama corpora but not in any comedies. The counts of numbers of words with either a comedy or drama bias were determined following range criteria. That is, if a word appeared in six of 10 shows and four of six were comedies, that item was counted as featuring more often in comedies (even though the item may have actually appeared many times in a single drama show). Similarly, if a word was in nine out of the 10 shows and five of the nine were dramas, then it was counted as having a drama bias. Thus counts showing 302 comedy-biased and 231 drama-biased words disguise a great deal of use across the two genres. It is interesting that this methodology which favours the detection of differences still identifies 154 words as being evenly distributed.

We now turn to the crucial coverage question: Will knowing the words on the TWL enable students with knowledge of only the 2000 most frequent words of English to reach a high level of coverage of TV shows? In Table 4, coverage percentages for the GSL 1000, GSL 2000, the AWL, the TWL and Off-list (i.e., words not on any of these lists) are shown for each of the 10 show sub-corpora. As can be seen in the rows for the TWL, its coverage ranges from under 1 to almost 3%. The mean for dramas approaches 2%, while the mean coverage of comedies is higher at almost 2.5%. The bottom rows in the table show the coverage offered by the GSL 1000, 2000 and TWL combined. Most of these totals amount to about 90% and none reaches the 95% known word criterion that has been identified as necessary for adequate comprehension. Thus the usefulness of the TWL appears to be somewhat limited. Only in the case of *Lost*, where the

combined contribution of the GSL 1000, 2000 and TWL amounts to 93%, can the criterion be considered to be almost met. Nonetheless, the overall contribution of the TWL should not be considered negligible: With coverage figures fairly consistently at around 2%, this means that one word in every 50 the TV viewer encounters is a TWL item.

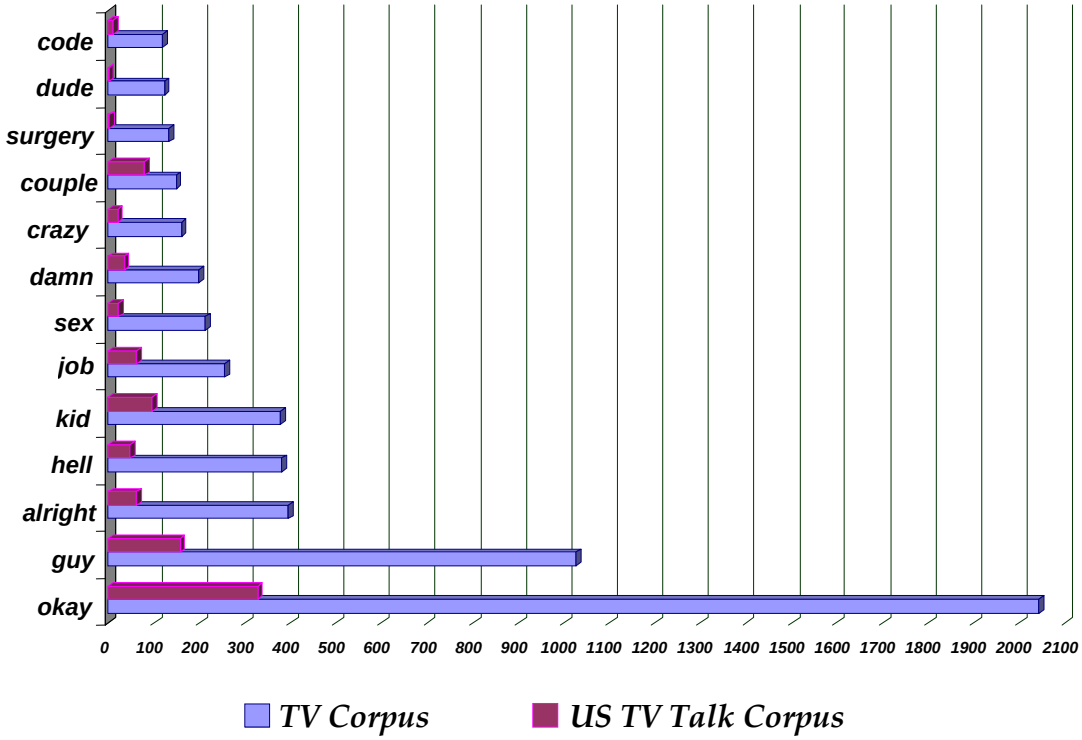

Figure 1. Comparison of frequencies of 13 TWL words in the TV and US TV Talk Corpora

Table 3. Distributions of TWL Words in TV Genres

	Number of Words	*Percentage*
Only Comedies	1	0.15
Only Dramas	1	0.15
Primarily Comedies	302	43.83
Primarily Dramas	231	33.53
Evenly Distributed	154	22.34
Total	689	100.00

Table 4. Word List Coverage of the TV Corpus Word List Corpus in %

Percent Coverage Drama						
	Alias	Desperate Housewives	Grey's Anatomy	Lost	Prison Break	Mean
1k	83.37%	84.71%	82.78%	86.04%	84.23%	84.23%
2k	4.20%	5.14%	5.09%	4.68%	4.70%	4.76%
AWL	2.15%	0.89%	1.36%	0.61%	1.14%	1.23%
TWL	1.70%	1.90%	2.07%	2.32%	1.85%	1.97%
Off-list	8.56%	7.35%	8.69%	6.35%	8.08%	7.81%
Total (1k, 2k, TWL)	89.27%	91.75%	89.94%	93.04%	90.78%	90.96%

Percent Coverage Comedy						
	Frasier	How I Met Your Mother	The Office	Seinfeld	Two and a Half Men	Mean
1k	84.15%	81.47%	80.75%	82.07%	82.70%	82.23%
2k	5.27%	5.11%	5.40%	5.40%	5.58%	5.35%
AWL	0.97%	0.90%	1.21%	0.85%	0.95%	0.98%
TWL	1.47%	2.64%	2.54%	2.40%	2.66%	2.34%
Off-list	8.15%	9.88%	10.10%	9.28%	8.10%	9.10%
Total (1k, 2k, TWL)	90.89%	89.22%	88.64%	89.87%	90.94%	89.91%

Finally, observant readers may have been surprised to see words like *hell* and *ass* among the most frequently occurring words on the TWL (see Table 2). Figure 1 shows that *hell* occurs almost 400 times in the corpus; this level of repetition amounts to a virtual guarantee that it will be noticed by viewers and remembered. This observation prompted us to take a closer look at other TWL words that ESL teachers might consider to be inappropriate for classroom use. The frequencies of *hell*, *ass* and six others are reported in Table 5. Clearly, these are words that show up regularly in interactions on TV. The presence of these words on the list is not really surprising given that it was created from popular North American TV shows. We recognize that learners need to know and use all kinds of words and that TV may play a useful role in teaching vocabulary that some teachers would rather not explain in class. Still we felt we should alert prospective users of the list to the presence of these potentially controversial items.

Table 5. Eight questionable TWL words and their frequencies in the TV corpus

TWL word	Frequency
hell	382
ass	110
bitch	68
idiot	42
piss	25
jerk	22
dumb	20
whore	8

DISCUSSION AND LIMITATIONS

In this study we identified the TWL, a list of 686 word families that occur frequently across a variety of current television shows that are popular with many viewers and plausible choices for ESL viewing homework. As expected, the television-based TWL proved to differ substantially from the AWL, a list of families that occur frequently in academic writing. Although it was difficult to identify strong similarities between our list and another TV-based list due to problems with the comparison corpus, we are confident that the TWL is a good reflection of the vocabulary used in current North American television shows. In other comparisons that we are not able to report in detail here, we found that the TWL accounted for 3% of the words used in another speech corpus (a 100,000-word collection

of ESL teacher talk) but only 1% of the words in a corpus of essays written by native-speaker college students (the LOCNESS corpus). We interpret the coverage advantage for the spoken corpus as an indication that the TWL is a credible reflection of the character of the spoken interaction on TV.

A pedagogical limitation of the study is that it focuses on single words. As corpus work by O'Keeffe, McCarthy and Carter (2007) shows, spoken interaction is characterized by the frequent use of multi-word lexical units. For instance, they found that the chunks *you know* and *I mean* were used very frequently in their corpus of conversations – even more frequently than the very basic single-word items *people* and *much* (p. 69). The case of *you know* illustrates the point that chunks may be made up of simple words but have less-than-simple meanings; understanding the meaning of the verb *to know* is of limited helpfulness in understanding the pragmatics of the chunk. It is very likely that the TV corpus we gathered contains thousands of expressions and idioms that would be useful for L2 learners to know; understanding them would probably make television shows a great deal more comprehensible. In this initial analysis, it was not feasible to also identify recurring strings, but we see this as an important goal for a future study.

Also, in terms of offering learners of English a manageable list of several hundred word families that can substantially boost comprehension of television, as the AWL appears to be able to do for university texts, the TWL proved to be weaker than the AWL. Coxhead's analyses (2000) show that knowledge of the words on the 570-word AWL can provide added known-word coverage that ranges from 5 to 10%. But in the best case scenario (see data for the *Two and Half Men* sub-corpus in Table 4), our 689-word TWL added 2.64% coverage. It is possible that a larger, more representative television corpus might have resulted in a more powerful list. We also recognize that in this initial foray into corpus-based list building, we may have made questionable design decisions that affected its coverage power. Nonetheless, we are convinced that the TWL is a pedagogically useful list. As mentioned, the figures indicate that the list accounts for over 2% of words on TV and while this may not sound like a great deal, it means that at least 1 word in every 50 is an TWL item. The data in Table 4 also show that if learners know the words on the TWL (in addition to the 2000 most frequent words), levels of known-word coverage hover around 90%. With this amount of coverage, TV comprehension may be difficult but perhaps not impossible given the visual support for meaning available. Thus there is reason to think that

study of the TWL can assist comprehension, and in doing so, increase the likelihood that new vocabulary in the input will be noticed, understood and learned.

In addition to the recommendation that learners who watch TV should familiarize themselves with the words on the TWL, the results have pedagogical implications. One of these relates to the differences in coverage percentages of various frequency lists in the shows that appear in Table 4; these indicate that some shows are more difficult than others in terms of their vocabulary content. The means in the leftmost column of Table 4 indicate that basic vocabulary at the 1000 most frequent level plays a larger role in dramas than in comedies, and that the contribution of the specialized TWL is larger in comedies than in drama. These are two ways of saying that the comedies are generally more challenging (at least in terms of vocabulary) than the dramas. Although these differences are small and in need of further confirmation, teachers may wish to take this into consideration when choosing shows to assign. The case of *Lost* is particularly interesting in this regard. The data in Table 4 show that this drama program stands out for having the highest proportion of 1000-level words (86%); this suggests that it is a good choice for teachers looking for a show that learners can readily understand. The figures also identify *The Office* as the most difficult show, with 81% of its words on the GSL 1000.

The finding that the TWL contained a number of off-colour words is also worthy of note. For this reason, teachers may wish to use the list with discretion, taking the age and cultural background of the learners into account, as well as the social setting in which any classroom TV viewing might occur. Teachers may also be interested to know that such items were found to occur frequently and that the likelihood that learners will encounter such words in TV watching homework is high.

There are also implications for further research. In working with this corpus and the Teacher Talk corpus mentioned above, we found evidence of shared vocabulary. It is to be expected that the two spoken corpora would have a great deal of basic English vocabulary in common, but we wondered about the extent to which more unusual words heard in classroom speech might reinforced by hearing them again on TV and the types of shows that would do this most effectively. There is also the problem of words that may be useful to know but are unavailable for incidental acquisition because they do not occur in either teacher talk or TV. These are interesting avenues for further exploration.

There are certain limitations to the research reported here. One became evident when we compared our corpus to the US TV Talk corpus and

found limited amounts of overlap in lists of frequently occurring vocabulary. As mentioned above, the comparison corpus was more varied in its contents than ours, which meant that the two corpora were probably not really comparable. This raises the question of the kinds of shows that might be included in a truly representative television corpus. For instance, we know that some ESL teachers encourage their learners to watch documentaries or news programs and that learners sometimes report having learned English through watching children's program such as *Sesame Street*. It seems likely that the exclusive use of dramas and comedies in our corpus constrains the usefulness of the word list we derived from it.

Secondly, although the 500,000-word corpus seemed large to us, many corpora used in current Applied Linguistics research are much larger; O'Keeffe and McCarthy (2007) warn that generalizations based on a corpus under 1 million words in size may not be reliable. We recognize this limitation on our work and suggest that it might be overcome by addressing the content problem outlined above. That is, the addition of a wider variety of programs – ideally chosen in consultation with ESL teachers and learners – could serve to make the corpus both larger and more representative.

In conclusion, we have been struck by the potential of a corpus-informed approach to reveal useful and unexpected discoveries about learners' language input. When we undertook this project, we were not sure that a list such as the TWL actually existed, and it is gratifying to see that its creation was possible. We certainly did not know that *okay* and *guy* would prove to be the most frequently and consistently used words across the 10 shows – by far. We look forward to continuing along this path, confident that insights derived from corpora will continue to inform and improve language teaching and sometimes also surprise us.

REFERENCES

Cobb, T. *Web Vocabprofile* [accessed April 2009 from http://www.lextutor.ca/vp], an adaptation of Heatley & Nation's (1994) *Range*.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *32*, 213-238.

Hyland, K., & Tse, P. (2007). Is there an "Academic Vocabulary"? *TESOL Quarterly, 41,* 235-253.

Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *Canadian Modern Language Review, 61*, 355-382.

Laufer, B. (1989) "What percentage of text-lexis is essential for comprehension?" In C.

Lauren & M. Nordman (Eds.), *Special Language: From humans thinking to thinking machines*, Clevedon: Multilingual Matters.

Laufer, B. (2000). Task effect on instructed vocabulary learning: The hypothesis of 'involvement.' In AILA '99 Organizing Committee (Eds.), *Selected Papers from AILA '99*. Tokyo (pp. 47-62). Tokyo: Waseda University Press.

LOCNESS Corpus, courtesy of Margaret Levy & Sylviane Granger.

Major, R., Fitzmaurice, S., Bunta, F., & Balasubranmanian, C. (2005). Testing the effects of regional, ethnic, and international dialects of English on listening comprehension. *Language Learning, 55,* 37-69.

Nation, I. (2001). *Learning vocabulary in another language*. United Kingdom: Cambridge University Press.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: language use and language teaching*. Cambridge: Cambridge University Press.

Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on computer* (pp.107-118). New York: Longman.

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*, 130-163.

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

West (1953). *A general service list of English words*. London: Longman.

APPENDIX

WORD LIST – 689 WORD FAMILIES

abuse	assign	bomb	candy	coach
academy	assist	bond	capable	cock
access	assume	boob	capture	cocktail
accomplish	assure	booking	career	code
accurate	attach	boom	carve	coincide
acquire	attitude	boot	cash	comment
addict	attorney	booze	casual	commit
adjust	authority	boring	celebrate	communicate
adore	available	boss	cell	community
adult	aware	bother	cereal	complex
affect	awesome	bounce	challenge	compromise
aggressive	awful	bout	champion	compute
airport	ay	boyfriend	channel	con
alarm	babe	brace	chart	concept
alcohol	babysit	brassiere	chase	conduct
alert	bachelor	breast	chat	confer
allergy	bail	brief	chef	confirm
alley	balloon	briefcase	chew	confront
alright	banana	Britain	chill	consequence
alternative	bang	bubble	chin	constant
amaze	baseball	buck	china	contact
America	basement	buddy	chip	contest
analyse	bastard	bug	chocolate	contract
angel	bat	bull	choke	converse
announce	bathroom	bullet	chop	convict
annual	beach	bump	Christ	convince
anti	beep	bunk	cigar	cookie
anytime	beer	busted	cigarette	cop
apartment	benefit	butt	civil	cord
apparent	bet	cab	classic	corporate
appreciate	betray	cable	click	costume
approach	bike	caffeine	client	couch
appropriate	biological	Canada	clinic	counter
area	bitch	cancel	closet	county
aspirin	blank	cancer	clothe	couple
ass	blend	candidate	clown	courtesy
crap	drip	final	grill	insist
crawl	drug	finance	guarantee	instinct

crazy
create
credit
crew
cruise
custody
cute
damn
darling
decorate
dedicate
definite
delicious
dentist
deny
deposit
depress
design
desperate
despite
dessert
detect
dial
diaper
diet
disaster
disorder
dispose
distract
divorce
dock
document
doll
donate
drama
laundry
le
league
leak
lease
legal
legend
lemon
dude
dumb
dump
earring
edit
elevate
embarrass
emergency
emotion
encounter
energy
engage
entitle
environment
equip
errand
escort
Europe
eventual
evidence
ex
executive
exhaust
exit
expert
expose
fabulous
facility
fake
fantastic
fart
fascinate
feature
feed
fiancée
mission
mistress
monster
mood
motive
mount
movie
mug
fist
flatter
flaw
flexible
flip
flu
fluid
flush
focus
folk
forever
foundation
France
frank
freak
frustrate
function
fund
garbage
gee
generate
genius
Germany
ghost
giant
girlfriend
glove
glow
goal
golf
gorgeous
grab
grade
grant
grape
pants
parade
paranoid
partner
passion
patch
pathetic
pea
gut
guy
gymnasium
hallway
hamburger
handcuff
handsome
headache
hell
hero
hockey
honey
honeymoon
hooker
hop
horrible
hug
huge
humiliate
humour
hysterical
id
identify
idiot
ignorant
image
imply
impress
incident
incredible
India
indicate
infect
injure
innocent
predict
pregnant
prescription
privacy
privilege
pro
proceed
process
institute
instruct
intelligence
intense
internal
interview
intimate
investigate
involve
irony
issue
Italy
item
jack
jackass
jacket
jail
jam
Japan
jar
jerk
jerry
jet
job
junior
junk
kid
kidding
kidney
label
laboratory
labour
lame
lane
launch
reside
resolve
resort
resource
respond
response
reveal
reverse

licence
lick
link
liquor
literal
locate
locker
loop
ma
magazine
magic
maid
major
mall
mama
manipulate
massage
massive
mate
mature
maximum
medical
medication
mental
mess
metaphor
Mexico
minor
miracle
sedate
seduce
seek
series
session
sex
shift
shotgun
shrink
significant
silly
site
sketch
ski
muscle
nah
naked
nap
nasty
Nazi
negative
nerve
nervous
nightmare
non
normal
obsess
obvious
occur
odd
offence
okay
onion
opera
option
ounce
outfit
oven
overhear
pacific
pal
panel
panic
strategy
stress
style
sub
subtle
sucker
sue
suicide
suitcase
sunset
super
supermarket
supervise
surgeon
peanut
percent
period
personality
petty
phase
phrase
physical
piano
pie
pill
pillow
pilot
piss
pitch
pizza
plastic
plug
plumb
plus
poke
policy
politic
pop
popcorn
positive
potato
potential
pre
tense
tequila
terrific
terrify
terror
theory
therapy
thou
thrill
tick
tiny
tissue
toast
toilet
professional
project
promote
psych
psychology
psychotic
punch
purchase
purse
quit
quote
radar
rage
range
react
recall
reception
recover
refrigerate
register
rehearse
reject
relax
release
relevant
remote
remove
require
research
vest
victim
video
virgin
visible
vision
volume
volunteer
vulnerable
wallet
warrior
web
wed
weird
rib
ridiculous
rig
rip
romance
romantic
route
routine
rum
rumour
sack
salad
sandwich
sane
Santa
satellite
scare
scenario
schedule
scheme
score
Scotland
scout
scream
scrub
seal
seatbelt
section
secure

skip	surgery	tone	whatsoever
slap	surgical	toothbrush	whoa
slash	surveillance	torture	whore
smart	survive	toss	withdraw
snack	sweater	tradition	woo
snap	sweetheart	traffic	wrinkle
sneak	switch	trail	yell
sniff	symbol	transfer	zip
soccer	tag	transmit	
soda	tale	transport	
someday	talent	trash	
source	tank	truck	
Spain	tape	tuck	
specific	target	turkey	
spy	tattoo	twin	
squeeze	team	ultimate	
stab	technical	underwear	
stable	technology	vacation	
stalk	teddy	van	
stare	teenage	vanilla	
starve	television	vehicle	
status	temporary	vent	
steak	tennis	version	

[1] We are grateful to an anonymous reviewer for useful perspectives on the L2 listening comprehension process.