English Podcasts: A Corpus Linguistics Study

Joe LAUER
Institute of Foreign Language Research and Education
Hiroshima University

This study investigates what types of vocabulary items and grammatical patterns are found in some English podcasts. By utilizing corpus linguistics software and compiling frequency lists, it was found that grammatical articles (*the* and *a*), prepositions, and pronouns are among the most frequent words in podcasts. Interestingly, the grammatical article *the* could be categorized into nine types of usage. Such findings, it is hoped, will help teachers develop better podcasts and vocabulary teaching materials.

BACKGROUND

Podcasts are an exciting way for students to improve their English skills. For virtually no money, students can download interesting and educational programs from the Internet into personal computers and mobile devices such as i-Pods and cell phones. The audio and/or visual programs automatically come to students, often on a weekly or even on a daily basis. Learners can utilize those materials whenever they want, wherever they want, and as often as they want — again, for practically no money. Lauer (2011) identified some of the best English-learning podcasts available for no charge through the iTunes Store.

In recent years, some studies have told us about Japanese students' attitudes and habits concerning English-learning podcasts, and even the effectiveness of those podcasts. For example, Lauer (2009, 2008) found that students generally say they enjoy such podcasts, but in reality they do not listen to them very much. Lauer and Enokida (2010), in an important group longitudinal study, found that some students make progress when studying podcasts, but other students do not. Gromik (2008, p. 58), also in a longitudinal study, found that podcasts stimulate students "to be responsible for their own learning."

But a key question remains: When students study English-learning podcasts, exactly what types of words are they being exposed to? Fortunately, corpus linguistics software allows teachers and linguists to picture the language contained in podcasts.

Indeed, understanding the frequency of words and expressions in a conversation or a text is invaluable information. One of the first people who stressed the importance of studying word collocations was Firth (1957). More recently, one of the most authoritative advocates of corpus linguistic studies is Sinclair (2004), who argues that knowing lexical patterns allows for a more powerful understanding of language than does traditional grammar. According to Sinclair, when we understand phraseology and frequency phenomena better, we will be able to create "the ultimate dictionary."

Today, the 450-million-word Bank of English corpus, created by COBUILD at the University

of Birmingham, is one of the most comprehensive groups of English words ever analyzed. It is 71% British English, 21% North American, and 8% Australian, with 86% of the text written and 14% transcribed spoken data.

According to studies (summarized in Moon, 2010), the 10 most frequent lemmas (the canonical forms of words) are *the, be, of, and, a, in, to* (infinitive particle), *have, to* (preposition), and *it*. Among the top 100 words, the top noun lemmas are *year, time, person/people, day, man,* and *way*, while the most frequent verb lemmas are *say, go, make, get, take, know, see, come, think,* and *give*. The top adjectives are *new* and *good*.

Interestingly, the 10 most-frequent words occupy about one-fourth of all words in a typical utterance or text, and the top 100 words occupy 45% of most things we say or write. Thus, if students fully understand just 100 words, they should be able to understand almost half of all English spoken and written language!

What are some differences between spoken and written vocabulary? Studying the Bank of English's twenty-million-word subcorpus of British conversation and local radio broadcasts, it is noticed that the words *mean, sort, thing*, and *want* are all much more common in oral language than in written language. Also, adverbs such as *very, quite, perfectly,* and *really* are found in both spoken and written texts, but a large number of adverbs, such as *blissfully, deliriously, gloriously, infectiously, radiantly,* and *serenely*, are much more common in written language. Also, conversation is marked by many discourse markers and phatics, such as *yeah, right, well, OK*, and *oh*; all of them have important pragmatic functions.

In spoken language, words tend to use their base forms and primary definitions, while words have broader usages and meanings in written language (Moon, 2010). For example, the common verb *know*, in conversation, usually is used like *I don't know how,* or *Well, you know, we should…* But in written language, *know* is often used in passive grammatical constructions or with deeper meanings, such as *The painting has become well known to…* or *It was the stare of a man who knew he was going to…*

The study of multi-word units is also important (Greaves and Warren, 2010; Carter and McCarthy, 2006). The most common multi-word units are preposition + article, subject + verb, subject + verb with complement items, and noun phrases + *of*. There are many two-word units such as *you know, in the, in fact, for example*, and *there is*, while high-frequency three-word units include *a lot of, be able to*, and *it was a*. Of course, there are few word units with lengths of five or more, and there are almost no nine or ten-word units, except for set phrases such as proverbs.

Another high-frequency multi-word unit is *one of the* (Scott and Tribble, 2006). Using the British National Corpus, it was found that the expression is most often followed by the word *most*. Other post-position words, in decreasing order of frequency, are *main, major, first, reasons, parties, earliest*, and *problems*.

Importantly, multi-word units can be categorized according to their discourse functions, and Biber et al. (1999) say that there are four main categories: stance bundles, interactional bundles, referential bundles, and text organizers. Stance bundles convey attitudes, such as *I*

don't know why and *are more likely to*. Interactional bundles include, for instance, politeness (e.g., *thank you very much*), and reported speech (e.g., *and I said to him*). Referential bundles involve time, place, and text markers, such as *at the beginning of, the end of the*, and *or at the same time*. Text organizers refer to contrast (e.g., *on the other hand*), inference (e.g., *as a result of*), or focus (e.g., *it is important to*). They say that the first two types of discourse markers are more commonly found in daily conversation, while the latter two types are more commonly found in academic discourse.

Similarly, Carter and McCarthy (2006) say that oral communication contains a lot of expressions which reflect interpersonal meanings, as in *you know, I think*, and *I know what you mean*. On the other hand, formal written language has more linking expressions, such *as at the same time, in the first place*, and *as a result of*.

Some corpus linguistics studies have tried to identify language which can clearly be called "idioms." O'Keeffe et. al. (2007) say that researchers can identify idioms using corpus software if they search using "basic cognitive metaphors", such as parts of the body. Thus, for example, when investigating the word *face*, expressions such as *face to face, on the face of it*, and *let's face it* appear. Using the Cambridge and Nottingham Corpus of Discourse in English, it was found that the five most frequent idioms in British spoken language are *fair enough, at the end of the day, there you go, make sense*, and *turn round and say*.

Finally, corpus linguistics software can be used to study gender stereotyping (see Moon, 2010, p. 208, for some previous research findings). For example, *husband* often collocates with *abusive, unfaithful, hardworking*, and *drunken*, while *wife* often collocates with *good, perfect, battered, pregnant* and *beautiful*.

THE PRESENT STUDY

This study asks the following questions:

1) Do Hiroshima University's English Podcast dialogs use the same kinds of expressions as:
   a) Other popular podcasts?
   b) Movies and TV programs?
   c) The British National Corpus's lists of the most frequent spoken and written words?
2) How are the notorious grammatical article *the* and the very-frequent word *to* used in the Hiroshima University corpus?
3) What are some other interesting characteristics of the words in the podcasts?

To answer these questions, first, 42 Hiroshima University English Podcast dialogs were analyzed using AntConc3.2.1 software (available free at http://www.antlab.sci. waseda.ac.jp/antconc_index.html). The dialogs were written by university students (both foreign and Japanese) and were edited by a native English speaker (this researcher). The dialogs mostly involved a man talking with a woman, and the topics varied extremely, from Australia to Hiroshima, and from gossiping to fishing. The dialogs appeared online from Nov. 16, 2010 to Nov. 15, 2011. Hiroshima University's English Podcast has about 5,000 listeners per week. Each dialog was about 300 words in length, or one A4 page, so this corpus totaled about 13,000 words.

For comparison purposes, two other widely-known podcasts were analyzed using the same software: ECC Podcasts and the British Broadcasting Company's (BBC) 6-Minute English programs. The ECC Podcasts are similar to Hiroshima University's English Podcast in that two people (a Japanese woman and a native English-speaking man) analyze dialogs for 15 to 20 minutes. But while the university podcasts are each based on one long dialog, the ECC podcasts each have two very short dialogs — each dialog being only about 50 words — and each dialog focuses on one conversational idiom, such as *to be on fire, to rake it in*, or *to rain on (one's) parade*. A total of 26 of these podcasts (52 dialogs), posted on iTunes from Aug. 21, 2011 to Nov. 11, 2011, were analyzed. They totaled about 2,500 words.

The BBC 6-Minute English podcasts were different from the above two podcasts in that the BBC podcasts were not based on written dialogs; rather, they had two native speakers talking about a topic, such as the night sky, political gaffes, handwriting, or stress, and its associated vocabulary. They each included a bit of audio from BBC news or feature reports, and the speakers spoke relatively quickly. So, it can be said that they were upper-intermediate or lower-advanced level podcasts. Also, it possibly can be said that these conversations were "more natural" than the other podcasts, in that they were not based around written dialogs. But this conclusion cannot really be made because, undoubtedly, many of the BBC conversations, too, were "prepared" to some degree beforehand. In any case, nine podcasts, which had been posted onto iTunes from Sep. 22, 2011 to Nov. 17, 2011, along with complete transcripts, were analyzed using the AntConc software. They totaled about 8,400 words.

The final two spoken corpora utilized in this study were very comprehensive ones. The first one, compiled by Wiktionary (2006), a free online dictionary service, ranks 29 million words which appear in movie and TV scripts available on Internet. And the other corpus is probably the most reputed one in existence: The British National Corpus of Spoken and Written Language (available at Leech et al., 2001). It ranks millions of words collected from audio and written sources.

RESULTS AND DISCUSSION

Word frequency comparisons can be made by analyzing Table 1. The most striking finding is that all five of the spoken corpus texts featured amazing similarity. For example, the items *I, you, the*, and *to* were among the most frequent spoken items in all the corpuses. Also, grammatical articles (*the* and *a*), prepositions, and pronouns — the so-called functional words — were very prevalent. The pronouns *I* and *you* were less frequent in the BBC podcasts than in the other two podcasts, undoubtedly because the BBC ones were not based on dialogs, while the other two were.

A second important finding was that the three corpus texts which clearly involved "prepared" scripts — Hiroshima University English Podcasts, ECC Podcasting, and the Movie and TV scripts — had very similar rankings to the "natural conversation" British National Corpus. This implies that the podcast writers, as well as the screen writers, are writing dialogs which are quite natural in terms of word frequency.

Finally, the spoken corpora are interesting in that they are so similar to the written corpus, and yet have a couple of key differences. One interesting difference is that the word *was* never appears in the top 20 of any or the spoken corpora, but it appears at the high Number 9 position in the written corpus; this implies that people write more about the past than they speak about it. Also interestingly, *I* is high in the spoken corpora, but it is as low as Number 17 in the written corpus.

Table 1. The Most Frequent Words

(The numbers in parentheses mean the numbers of occurrences in the corpora studied)

	Hiroshima University's English Podcast		ECC Podcast		BBC 6-minute English		Movies and TV[1]	BritishNational Corpus Spoken/Written[2]
1	I	(641)	the	(101)	the	(448)	you	the / the
2	you	(461)	I	(94)	to	(222)	I	I / of
3	the	(400)	to	(90)	a	(208)	to	you / and
4	to	(309)	you	(73)	of	(205)	the	and / a
5	a	(298)	a	(61)	and	(193)	a	it / in
6	it	(257)	and	(40)	it	(173)	and	a / to
7	and	(219)	it	(37)	in	(137)	that	to / is
8	that	(167)	he	(37)	is	(136)	it	of / to
9	of	(147)	in	(35)	you	(133)	of	that / was
10	in	(137)	on	(30)	that	(131)	me	in / it
11	is	(130)	is	(29)	I	(125)	what	we / for
12	just	(124)	of	(28)	we	(90)	is	is / that
13	we	(123)	this	(21)	for	(81)	in	do / with
14	what	(119)	be	(20)	this	(78)	this	they / he
15	for	(114)	no	(19)	so	(73)	know	was / be
16	can	(107)	was	(19)	are	(65)	for	yeah / on
17	this	(106)	that	(18)	but	(65)	no	have / I
18	me	(99)	for	(18)	they	(53)	have	what / by
19	are	(97)	me	(18)	about	(52)	my	he / at
20	have	(96)	about	(17)	can	(47)	just	that / you

1) Movies and TV: Based on Wiktionary's (2006) list, analyzing millions of words in movie and TV scripts available on Internet.
2) British National Corpus: Found at Leech et al. (2001), analyzing millions of words.

But word frequency is only a part of the descriptive linguistics picture; word collocations also tell an important story. Since grammatical articles are so frequent in the corpora and pose such a big obstacle for students learning English in Japan, this study examined closely the distribution of the grammatical article *the* in the Hiroshima University Podcast dialogs.

Adapting the University of Toronto's Writing Homepage (2011) description of grammatical

article usage to the corpus data, it was found that the grammatical article *the* could be classified into nine usage categories. (See Table 2.) Idiomatic expressions have to be learned by memorization, but the other eight categories are relatively easy to understand. Thus, if students can learn these eight simple patterns, involving 78% of all occurrences of *the*, they will go a long way toward mastering the usage of *the*!

Table 2. Distribution of the Grammatical Article *the* in Hiroshima University Podcasts

	Categorization (～ means "noun")	Percentage of Total *the* Occurrences (Actual number in corpus)	Some Examples
1	Idiomatic Expressions	22% (90)	*make the best of it, in the mornings, at the moment, all the way, slap you round the ear, the Carp, the same*
2	Things in the Neighborhood or Home	15% (60)	*the store, in the park, in the bushes, on the couch, the dishes, the flashlight*
3	Reference for 2nd Time in Talk	11% (44)	*the service, the noise, the exam*
4	Superlatives	7% (28)	*the biggest, the worst, the slightest, the man* (meaning *the best*)
5	*the* ～ (relative clause)	12% (47)	*The man who, the trouble* (which I have), *the song* (which is entitled) *"The Nightingale", the mood* (I'm in)
6	*the* ～ (*of* ～ , where the *of* clause is not clearly stated)	10% (41)	*the government* (of Britain), *the entrance* (of the building), *the teacher* (of the class), *the history* (of Halloween)
7	*the* ～ *of* ～ (where the *of* clause IS clear)	6% (24)	*the price of, the bottom of, on the eve of, the number of*
8	the (adjective + noun)	9% (36)	*the Japanese Embassy, the 21st century, the speed limit*
9	*the* ～ (prepositional phrase)	8% (30)	*the batteries for, the way to, the queue at*

Also, this study looked closely at the Hiroshima University collocations containing the lexical item *to*, which was the fourth most common word in the corpus. First, it appeared in 76% of cases (234 out of 309 total usages) as an infinitive, in expressions such as *It's OK to sit outside*, *They need to hire*, and *People are having to wait*. Of the remaining usages, *to* was used as a preposition in 17% of total occurrences; specifically, *to* was followed by a place in 12% of total occurrences (e.g., *go to Japan*) and a living person or animal in 5% of occurrences (e.g., *talk to me*). The remaining 6% of usages can be called idiomatic usages, and they are all listed in Table 3.

Table 3. Idiomatic Usages of the Word *To* in the Hiroshima University Data

have tickets to ～	wake up to ～	sink to an all-time low
What have you been up to?	from head to tail	get back to (-ing verb)
sentence ～ to ～ (2 occurrences)	be allergic to ～	refer to ～ by ～
According to ～ (2 occurrences)	cling to ～	do ～ to ～
consign ～ to ～	get used to ～	similar to ～
be up to ～ (meaning "depend on")	prefer ～ to ～	prove ～ to ～

Note: ～ means "noun phrase"

Finally, this study looked a bit at gender references in all three podcasts (total about 24,000 words). For example, it was found that the word *he* appeared 153 times, but *she* only 56 times. The word *husband* appeared twice, in *eat with my husband* and *I love my husband*. The word *wife* appeared a total of four times, in the expressions *moved into a new house with his new wife, my wife is cooking, He was caught cheating on his wife*, and (*the cat's*) *with my wife all the time*.

CONCLUSION

This study has shown that corpus linguistics software can successfully be used to identify which types of vocabulary are found in English podcasts. By looking at the frequency lists, it can be seen that students of English crucially need to master grammatical articles (*the* and *a*), prepositions, and pronouns. Interestingly, specific uses of *the* were identified. Also, it was found that podcasts which contain dialogs have higher frequencies of the words *I* and *you* than does a podcast which does not contain dialogs. Also, some uses of vocabulary related to gender were noted. In the future, researchers should study a larger number of podcasts, and should try to identify language patterns so that better podcasts and teaching materials can be made.

REFERENCES

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *The Longman Grammar of Spoken and Written English*. Harlow, England: Pearson Education.

British National Corpus of Spoken and Written Language (2001). Retrieved Dec. 5, 2011 at http://ucrel.lancs.ac.uk/bncfreq/flists.html

Carter, R. A. and McCarthy, M. J. (2006). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Firth, J. R. (1957). *Papers in Linguistics 1934–1951*. London: Oxford University Press.

Greaves, C. and Warren, M. (2010). What can a corpus tell us about multi-word units? In *The Routledge Handbook of Corpus Linguistics*, edited by A. O. Keeffe and M. McCarthy. New York: Routledge, 212–226.

Gromik, N. (2008). EFL learner use of podcasting resources: A pilot study. *JALT CALL Journal*, 4, 47–60.

Lauer, J. and Enokida, K. (2010). A longitudinal study: The effectiveness of podcasts for learning

English. *Hiroshima Studies in Language and Language Education*, 13, 75–92.
Lauer, J. (2009). Podcast Power: Hiroshima University's new English listening materials. *Hiroshima Studies in Language and Language Education*, 12, 85–94.
Lauer, J. (2008). High-quality podcasts for learning English. *Hiroshima Studies in Language and Language Education*, 11, 95–106.
Leech, G., Rayson, P. and Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Retrieved Dec. 5, 2011 at http://ucrel.lancs.ac.uk/bncfreq/flists.html
Moon, R. (2010). What can a corpus tell us about lexis? In *The Routledge Handbook of Corpus Linguistics*, edited by A. O. Keeffe and M. McCarthy. New York: Routledge, 197–211.
O'Keeffe, A., McCarthy, M. J. and Carter, R. A. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.
Scott, M. and Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
Sinclair, J., Jones, S. and Daley, R. (2004). *English Collocation Studies: The OSTI Report*. London: Continuum.
University of Toronto's Writing Homepage (2011). Retrieved Dec. 5, 2011 at: http://www.writing.utoronto.ca/advice/english-as-a-second-language/articles
Wiktionary (2006). Retrieved Dec. 5, 2011 at http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#TV_and_movie_scripts

要　約

英語学習用ポッドキャスト教材
― コーパス言語学からの分析研究 ―

ジョー・ラウアー
外国語教育研究センター

本論文では，いくつかの英語学習用ポッドキャストをとりあげ，その中でどのようなタイプの語彙項目や文法の諸パータンが取りあげられているかを調査した。コーパス言語学において用いられるソフトを使い，頻度リストを作成することにより，ポッドキャストにおける最頻出語が，冠詞（the と a），前置詞および代名詞等であることが判明した。興味深いことに，冠詞 the の語法は，9つのタイプへ分類することができた。これらの新たな知見は，教員がよりよいポッドキャスト教材および語彙教育用教材を開発する際の一助になると期待される。