ISSN 1883-7409 (print)
ISSN 1884-0183 (on-line)

Outside the Box: The Tsukuba Multi-Lingual Forum

Volume 3, Issue 1

Autumn, 2010

Foreign Language Center

Tsukuba University

Japan

Theory and Other Dangerous Things

Teaching Tips & Techniques

Around the World

Creative Writing

Tidbits from the Corpus

John P. Racine

Dokkyō University

Abstract: This article provides language teachers with examples of the kinds of authentic language that can be uncovered through corpus linguistics. Specifically, facts about vocabulary in context, frequency of collocations, and grammar as it is actually used in English text are revealed. Corpus research projects, both for language teachers and their students, are suggested.

In recent years, an increasing number of teacher training courses and language-related graduate programs have incorporated Corpus Linguistics (CL) components. The increased popularity of CL is due, in part, to a greater emphasis placed on the use of authentic materials in TEF/SL classes. Increased availability of corpus technology online has also contributed to CL's popularity. Perhaps the most obvious reason for the popularity of CL, however, is simply the inherent interest that corpus research findings stimulate in those who take the time to examine these linguistic nuggets.

Despite teachers' growing interest in corpus research, some say (e.g., Groom, 2009) that the use of corpora has yet to be successfully integrated into classroom practice. In this article, I hope to offer some examples of the kinds of information that can be retrieved from linguistic corpora and to suggest how such data may be useful to you as a language teacher.

For those of you who are not familiar with this burgeoning field, we should first address a simple question: *What is corpus linguistics?* Corpus linguistics is an approach to the investigation of language which utilizes large databases, or *corpora*. These corpora contain samples of naturally occurring language extracted from such diverse sources as books, magazines, newspapers, scientific reports, and even brochures and menus. Spoken language corpora incorporate samples of language drawn from public speeches, television and radio broadcasts, as well as spontaneously recorded speech.

The utility of corpora as language research tools stems from their sheer size (the British National Corpus, for example, now contains over a 100 million words) and from the fact that they contain *naturally occurring language*. Thus, they provide a cross-section of the English language as it occurs in *actual use*. Note that this differs from the *prescriptive use* of language described in traditional grammar textbooks and teaching materials.

How then can corpus data be applied to teaching practice? Well, for one, this data[1] allows us to examine vocabulary in context.

Vocabulary and Collocations

Since corpus data is stored in computers, it can be manipulated easily, sorted and counted. One of the important offshoots of this is that we can learn how often words appear in spoken vs. written contexts. Thus, we know that *really*, for example, is in the top 50 spoken words. And we are now certain, as you may have expected, that the adverbs *really* and *pretty* are in much greater use orally, than in written form (nine and seven times more often, respectively).

You can see how these findings might influence teaching/learning practice. Increased emphasis can be placed on teaching vocabulary that learners are most likely to encounter in real life. Speaking classes could focus on words that appear most often in conversation while reading and writing classes could focus on the most popular written English.

[1] I am sure some of you prescriptive grammarians out there noticed that I did not refer to *data* in plural form (i.e., *these* data). A quick check of the corpus reveals that approximately two-thirds of all contemporary uses of *data* are in the singular collective form I have used here. Latinists may scoff, but this is authentic English. As we'll see throughout this article, this is only one example of the kinds of English usage that are revealed through corpus research.

Word frequency is just one example of the potentially useful information that can be gleaned from corpus data. We can also discover which collocations (combinations of words) occur most frequently. For example: What kind of adjectives typically follow *That would be…* in conversation? The Top 5: *nice*, *good*, *great*, *fun*, and *cool*. We also know that 83% of the uses of *yet* are in negative statements (e.g., *No, not yet.*) while only 10% of uses of *must* and *might* are negative (*You must not do that.*). This kind of collocational information can deepen learners' knowledge of word use.

Frequencies of collocations can also be useful in disambiguating word meanings. I have demonstrated elsewhere (Racine, 2009) how traditional dictionary definitions may not always be useful in distinguishing meanings of related words such as *sensual* and *sensuous*, or *comprised* and *composed*. Indeed, corpus data may allow teachers and learners to identify differences in word usage that are not apparent in traditional dictionary definitions.

More vocabulary tidbits

- Everyone likes *like*: *Like* is in the top 15 English words.
- *Today* and *tomorrow* appear in spoken English more often (two and three times, respectively) than in written English. *Yesterday* appears slightly more often in writing.
- When asking people to repeat themselves, *I'm sorry?* is more common than *Excuse me?*

Grammar

Besides vocabulary, there are also many facts about English grammatical structures to be discovered in corpus data. For example, the passive tense is more common in oral news reports than in regular conversation. The passive is even more common in written news reports where it is five times more likely to appear than in regular conversation. Wouldn't the passive voice be an important point of review in Reading or Current Events classes?

To return to the concept of actual vs. prescriptive usage, perhaps nothing provides a more salient example than the frequency of grammar "mistakes" in native English. For example, people in North America say *I wish I was…* and *If I was…* three times as often as they say *I wish I were…* and *If I were ...* Prescriptive grammarians will surely be appalled, but if we are to prepare our students for interaction in English as a foreign language, shouldn't we spend more class time familiarizing them with these "incorrect" spoken forms?

More Grammatical Morsels:

- The present perfect (*I've studied.*) is 10 times more frequent than the present perfect continuous (*I've been studying.*).
- The ambiguous present continuous: *He's not coming.* and *They're not having fun.* are less common than *He isn't coming.* and *They aren't having fun.*

Research Projects

Finally, there are a potentially unlimited number of corpus findings that could provide a starting point for further student (or teacher!) research. As an example: North Americans say *It's cold.* ten times more often than they say *It's hot.* Is this a geographical phenomenon? A linguistic one? Are the occurrences of *It's hot.* rising along with world temperatures?

What does it mean that the "positive" halves of most antonymous adjective pairs are more frequent than their negative counterparts? For example, *good* and *full* are used five times more frequently than *bad* and *empty*. Is this merely due to linguistic marking? Why is it that *easy* is less frequent than *difficult*?

Popular words, possible projects:

- *Mother* is the head of the nuclear family. *Uncle* is the most popular member of the extended family.
- In the animal world, *horse* beats *dog* by a nose.
- Sorry, Britney: *John* and *Mary* are still the most popular names.
- When mining the British National Corpus, you are seven times more likely to find *gold* than *aluminum*.

- *Sunday* is the most popular day of the week.

Hopefully, I have introduced at least a few ways in which corpus findings can be relevant to you as a language teacher. But don't take my word for it. Take a look at the many corpora and corpus linguistics resources that are now available online. Here are just a few:

BNC Web – http://bncweb.info/

MICASE (Michigan Corpus of Academic Spoken English) – http://www.elicorpora.info/

David Lee's *Bookmarks for Corpus-based Linguistics* – http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/CBLLinks.htm

Mark Davies's Links – http://davies-linguistics.byu.edu/personal/

Laurence Anthony's *AntConc Homepage* – http://www.antlab.sci.waseda.ac.jp/antconc_index.html

Of course corpus-based reference materials are not limited to online resources. A variety of academic journals related to corpus linguistics have been established in recent years, including *The International Journal of Corpus Linguistics* in 1995, *Corpus Linguistics and Linguistic Theory* in 2005, and *Corpora* in 2006. The number of corpus-inspired books is also growing rapidly. More and more corpus-informed reference materials (e.g., Carter & McCarthy, 2006; Sinclair, 2003) and textbooks (e.g., McCarthy, McCarten, & Sandiford, 2005; McEnery, Xiao, & Tono, 2006) are now available to language teachers. *Their prescription? Authentic language!*

References Cited

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English*. Cambridge: Cambridge University Press.

Groom, N. (2009). Introducing corpora into the language classroom. *The Language Teacher, 33*(7), 26-28.

McCarthy, M., McCarten, J., & Sandiford, H. (2005). *Touchstone: Student's book 1*. Cambridge: Cambridge University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies*. New York: Routledge.

Racine, J. P. (2009). Corpus linguistics as teacher tool: A *sexy* study of four adjectives. *Studies in Humanities and Communication, 6*, 263-288.

Sinclair, J. (Ed.). (2003). *Collins COBUILD advanced learner's dictionary* (4th ed.). Glasgow: HarperCollins.

About the author: John P. Racine teaches in the Interdepartmental English Program at Dokkyō University.

.