USING CORPORA IN LANGUAGE TEACHING AND LEARNING

by **James Thomas**
Masaryk University,
Brno, Czech Republic
thomas@fi.muni.cz

Introduction

In June 2005, I attended Lexicom 2005[i] which was held at the Faculty of Informatics, Masaryk University (FI MU) in the Czech Republic. The workshop was run by Adam Kilgarriff, Sue Atkins and Michael Rundell, who together form the Lexicography MasterClass[ii]. Dictionaries for language learners was a recurring topic, in particular the criteria for deciding which lexical items to include, and how to present this distilled information to learners. Some of the corpus-based methodology employed by modern day lexicographers is similar to the approaches taken by language teachers and students using corpora for their own study of language.

It is with a statement from Michael Rundell's opening session that I would like to begin this article proper. In considering types of knowledge, he quoted the American Secretary of Defense, Donald Rumsfeld:

> "Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns, the ones we don't know we don't know. ...." *DoD Press Briefing*.[iii]

Michael Rundell pertinently adds that there are also *unknown knowns.* These are the things we do not know we know, i.e., things we know only subconsciously. For example, in the case of language, it can be quite difficult to account for how one chooses a particular word instead of one of its synonyms, or what difference word order makes, or the effect of pragmatic devices, or in English, the use of *for* in the sense of *because*, or *I'll think about it* vs. *I'll think it over* vs. *I'll give some thought to it* or *take a photo of something* vs. *photograph something*. These language choices are particularly puzzling to native speakers, who by and large use language subconsciously.

Starting with Language

To account for language phenomena, we need to examine a large sample of genuine, or *attested*, language not invented “possible” sentences. John Sinclair (1991: 6) effectively pruned the argument in favour of invented sentences when he wrote: "One does not study all of botany by making artificial flowers." Regardless, there are not enough artificial sentences to draw meaningful conclusions from and furthermore, they are created purely on the basis of intuition, to which he optimistically commented, "the stranglehold of intuition is being relaxed" (ibid. p.6).

As is well-known, the large samples of attested language come in the form of language corpora. These now exist for many languages and sub-languages, such as corpora of academic language, legal, medical, tourist and computer language. Using a concordancer, the type of program that searches corpora and presents the findings, the existence of *unknown knowns* can manifest and the constraints on particular language choices can be observed. From such data comes information which, given the necessary conditions, can become knowledge.

Here is an example. A post-graduate computer science student emailed me recently asking about the use of *against* after *robust.* Intuitively it sounded wrong and *robust against* was not found in the *Cobuild Dictionary* (1995) – this was not surprising as it does not appear in the 56 million words of the Cobuild’s *Corpus Concordance Sampler*[iv] - nor in the *Macmillan English Dictionary for Advanced Learners* (2002). In addition to these learner dictionaries, *The New Oxford Dictionary of English* (1998) was consulted with the same result. The student remained convinced that he had seen *robust against* often enough.

The concordancing program, *Word Sketch Engine*[v] (Kilgarriff and Rychlý, 2004), presents computationally intelligent summaries of corpus data in very straightforward formats. I used this program to consult the British National Corpus[vi] (BNC) with its 100 million words of naturally occurring English between 1960 and 1994 (94% between 1985-1993). It accorded with my intuition in finding no such bi-gram. A search of texts from the computer domain did, however, find that *robust against* did indeed occur in that domain.

From this example, a number of points can be observed. A corpus of general English demonstrated that *robust against* is not core English, while consulting an appropriate corpus showed that it exists in a specific domain. From a pedagogical point of view, the student consulted the teacher who consulted the resources. With a little training, the student can now consult the resources himself.

This leads us to ask who uses corpora in language pedagogy: on the one hand, teachers, teacher trainees and students of language and translation, on the other, resource

writers ranging from teachers producing ephemera to textbook authors, grammarians and lexicographers.

Before describing some of the activities these applied linguists undertake, I would like to make a point about vocabulary study. It seems that while students of English acquire a sophisticated range of concepts and metalanguage relating to grammar and syntax, lexical and semantic concepts do not figure to nearly the same extent. And this is despite the oft repeated cry that vocabulary teaching has finally assumed its rightful place alongside grammar. See, for example, *The Lexical Approach*, (Lewis: 1993), *How to Teach Vocabulary* (Thornbury: 2002) and *Vocabulary, Semantics, and Language Education* (Hatch & Brown, 1995). On another level, the fuzzy border between vocabulary and grammar, and the interdependence of them, seem to be under continual investigation.

Some of the concepts that language students are partly, rarely or never acquainted with include:

- synonymy, antonymy, polysemy;
- hyperonym, hyponym, troponym;
- metonym, meronym, synecdoche;
- collocation, semantic prosody, lexical support;
- colligation, complementation, valency, frames;
- denotation, connotation, metaphor;
- lexeme, chunk, phrase, lexical unit;
- homonym, homophone, homograph;
- affixation.

Being unaware of these concepts renders it improbable that the student can make the vocabulary choices that depend on them. There are several practical examples below. Corpora also yield a wealth of data that reveal some of the *unknown knowns* of grammar. *The Longman Grammar of Spoken and Written English* (Biber, et al: 1999) is perhaps the most graphic example of this as the authors present their statistical findings about grammar using graphs and charts. Here are two examples from pp 488-9, which present some of the findings concerning the frequency of modal verbs.

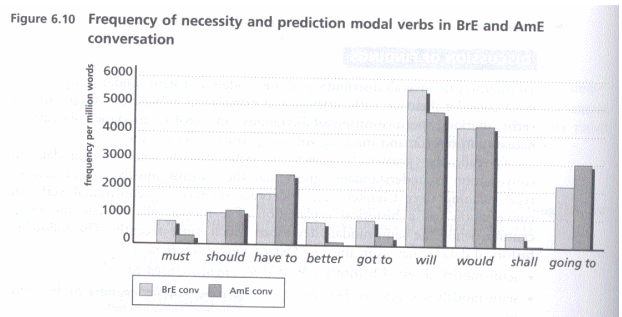


Figure 6.10 Frequency of necessity and prediction modal verbs in BrE and AmE conversation

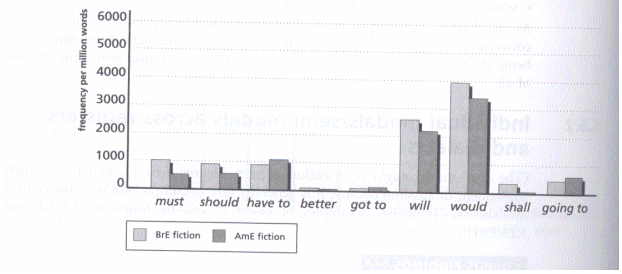
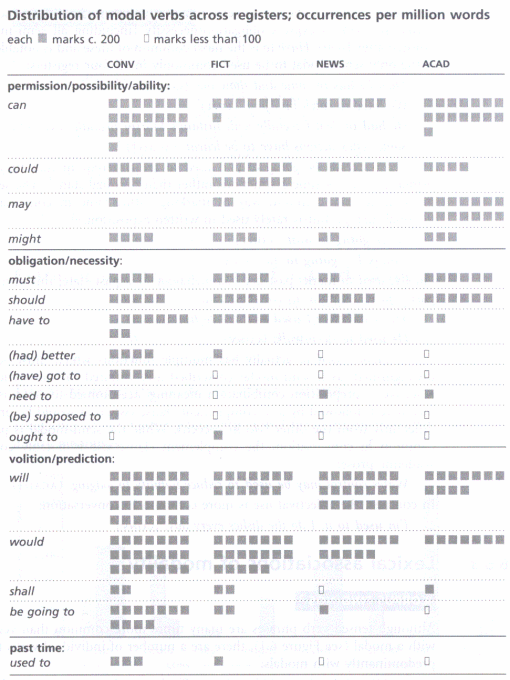


Figure 6.11 Frequency of necessity and prediction modal verbs in BrE and AmE fiction



Pedagogical Applications and Implications

The teacher's practical application of corpora can be divided into in-class use and out-of-class use. Illustrative sentences are used widely in language teaching and testing, and a corpus is an excellent source of them. Concordancers efficiently find very specific language phenomena. A practical example is the issue of how to avoid using the same word repeatedly. While synonyms are often seen as a remedy to this, synonyms are often mutually exclusive because of the very features that distinguish them from each other, i.e., constraints. Hypernyms are often a better option, and the corpus can exemplify this: vehicle → car

1. … upon her getting out of the car, they manoeuvred the vehicle so as to ….
2. whether it be ratings out of 10, defects per vehicle or warranty costs on each car leaving the factory gate .
3. Heron -- which builds houses, owns petrol stations and imports Suzuki vehicles as well as selling other cars including Rolls-Royce

Another example: *Ready for First Certificate* (Norris 2001: 45), the textbook I am currently using with a class, introduces some uses of *take.* As a supplementary activity, I created a pairwork questionnaire using some of the commonly occurring instances. The WSE displays a table[vii] of the grammar patterns (colligation) that the search word engages in. And under each grammar pattern, the statistically significant words (collocates) are listed. From that data, questions such as the following were written.

- Did the aftermath of Hurricane Katrina take you by surprise?
- Does your family take precedence over your friends?

- Have you ever been taken for a ride?
- Have you ever taken in a lodger?
- Who in your family do you take after?
- Were you surprised by eBay's buyout of Skype?
- What do you take off when you enter a house in winter?

From the same textbook comes the instruction: "Write down three more adjectives to go with the noun *device*". Students can think of three and then the WSE can show them the full gamut, either in real time using a data projector, or by passing around some printouts or displaying as overheads. In the process, the students are making not only observations of language per se, but of a procedure that they can employ in their language study and in their practical use of English. Click here[viii] to see the word sketch of *device* and here[ix] to see the first one hundred concordances of *adjective + device*. Note: if you click on any of the buttons in these examples, you will be asked for a password. Click the Cancel button and you will be able to register for the Sampler version of the program.

Correcting written work[x] is another sphere of activity in which teachers use corpus data. Whether free writing or translation, students' deployment of words can be compared with attested native speaker language. Since the process of improving one's foreign language skills manifests in using the language more and more idiomatically, the statistical probability of words being used in each other's environments needs to be considered. And a corpus can provide this. Some examples follow.

A student recently submitted a paper which included *In my point of view*. By simply typing *point of view* into the phrase field, *from my point of view* is immediately apparent. In the same paper, *to have to their disposal* appeared. By typing in *disposal*, *at* is the most frequent preposition – 597 times, the next being *of*, 196 times, and that reveals a different meaning of the word. He also wrote *copiously repeated mistakes.* The most frequent adverbs preceding *repeated* which indicate *a number of times* are *often* (17 time), *frequently* (11), *endlessly* (10), *constantly* (8), *regularly* (5), *oft* (4), *consistently* (4), *widely* (3), *usually* (3), *persistently* (3), *continually* (3), *perpetually* (1), *interminably* (1). The less frequent of these have the negative connotation that was probably intended by *copiously.*

We shall now turn to students' use of corpora. Tim Johns[xi], the father of Data Driven Learning (DDL), evolved his approach around the time when John Sinclair et. al. were developing the first COBUILD dictionary. The BU in the acronym stands for Birmingham

University where they were both working. DDL has its pedagogical foundation in such thinking as Tarone and Yule (1989:11) who recommend:

> a task-based, problem-solving, interactive learning approach for fostering sociolinguistic competence with the learner as ethnographer, making observations from data they find. (Tarone & Yule, 1989:11)

While their statement was not made with any reference to Johns' work, DDL answers their call admirably, for this is largely how a *kibbitzer* works. A kibbitzer is in some ways like action research on an isolated linguistic item in that it presents the question or quandary, the process and the data, and the results. Click here[xii] to see some examples of kibbitzers on display at MICASE[xiii], the Michigan Corpus of Academic Spoken English.

One of the basic tenets of Dalton Education[xiv] is *if the teacher does all the work, the students don't learn anything.* Applied to DDL, the process of researching language to answer one's own queries is maximally involving. For example, students can use corpora to check forms of words, infer meaning, find collocations and colligations, observe register, genre, mode, etc, and observe the contexts and co-texts in which words are used. This usually works as guided discovery activities.

Such an involved and multi-faceted process also enriches students' linguistic awareness. Whether or not students need this linguistic sophistication is a moot point[xv]. It is my view that the more information someone has, the better equipped one is to make choices while speaking and writing.

However, the practicality of engaging students in DDL tasks is not without problems. The reality of learning styles and classrooms and teachers and textbooks and examinations cannot be denied. A basic issue here is that students can be overwhelmed with language that is incomprehensible due to its richness in cultural references, figurative language, undecodable syntactic structures, and the like, in short, the very elements that make such language desirable input. This richness is a far greater contribution to *learner input* than many an artificial sentence, which typically lacks any sense of anchoring in time or place, are devoid of cultural or attitudinal stance, and seem committed to a matchstick scaffolding for the word or phrase. Such a poverty of input cannot lead to a healthy and vigorous *learner output.* One solution offered to the problem of incomprehensible data has been the creation of a corpus of readers, i.e., of simplified language. However, research undertaken by Ramesh Krishnamurthy[xvi] demonstrated that this compromised language did not constitute a rich linguistic diet.

The sheer volume[xvii] of the data presented can also overwhelm, so it is fortunate that the newer concordancers are able to present user-friendly summaries of large amounts of data.

Some complain that the time taken to solve a quandary is disproportionate to the information gleaned, while others believe that in working with the language so closely, one is incidentally gaining additional language experience in terms of quantity and quality. This is in addition to learning a skill with the potential life-long benefits of learner independence.

A more principled solution then is to adapt the task, not the language. A few examples of task type follow.

1. Lexical Support

Words are sometimes used in the environments of other words which have a similar meaning, force or function. This idea of *lexical support* can be observed simply by observing the frequent collocates and by examining concordances. For example, the top 20 collocates of the word *disgusting,* are *disgusting, revolting, ugh, disgraceful, vile, gust, sill, urgh, Camille, shocking, obscene, filthy, horrible, Lydia, absolutely, fucking, unpleasant, bloody, ugly, dirty*. Here are four sentences from the BNC that exemplify this.

- It was absolutely filthy , horrible and scuzzy , with disgusting stains on the floor.
- They said 'It stinks , it 's disgusting , it 's horrible stuff!'
- It is difficult to imagine any of the jargon-junkies who preside over American psychology writing , for example , that ` nothing filthy , disgusting , foul , loathsome , nauseous , offensive , revolting , vile , squalid , feculent , or obscene ' seems to have escaped the attention of modern ` artists ' .
- It is disgusting and immoral and a disgrace.

2. Polysemy

Here are three sentences containing *abandon*. And, following them, three of the meanings from the *Oxford Advanced Learner's Dictionary* online[xviii]. The students are required to decide which of these meanings is employed in each sentence. They are also required to explain how they arrived at their conclusion. And finally, they should locate some more illustrative sentences for each case.

1. Some teachers, in starting from "what was there", even abandoned the attempt to expose students to "the best that has been thought and said".
2. The Communist Party had not yet abandoned its attempts to gain control of the ILP, despite the assurances made in the previous year.
3. This is not to imply that expressions of sophisticated learned eloquence should be abandoned in favour of popular writing.

    a) to stop supporting or helping sb; to stop believing in sth
    b) ~ sb (to sth) to leave sb, especially sb you are responsible for, with no intention of returning
    c) to stop doing sth, especially before it is finished

3. Colligation

Which prepositions follow these verbs? (a) believe, (b) depend, (c) rely, (d) hope

Which prepositions follow these adjectives? (a) keen, (b) enthusiastic (c) good (d) interested

What difference does the choice of prepositions make with (a) dream (b) struggle (c) laugh (d) die.

Which of these verbs is followed immediately by a *to* infinitive? (a) let (b) make (c) manage (d) allow

4. Combined skills

In this activity, the students have to choose the only possible word from among the underlined words.

Two to three hundred Czech doctors are deserting/leaving/going for western Europe every month, according to digits/numbers/figures from the Czech Doctors Association given/released/published in Monday's Mlada fronta Dnes. The Association bases its digits/numbers/figures on applications it gets/receives/takes for a certificate needed to work abroad. Britain is one of the most popular/desirable/trendy destinations for Czech doctors, with some of them commuting home to the Czech Republic at weekends, the paper writes/says/reports. *[Cesky rozhlas, June 2005]*

Conclusion

This article has been concerned with some theoretical issues and practical applications of using a concordancing program. We have done so using a monolingual snapshot corpus of general English, namely the BNC – it is a representative sample of English. Another type of corpus is the *monitor corpus* which is continually added to, and there are *bi-lingual* and *parallel* corpora which have texts in two or more languages. As mentioned above, there are many specific corpora representing a domain, a genre, an author, etc. Of particular interest in pedagogical spheres are *learner corpora*, which contain language written by non-native speakers. This is used in error analysis, language acquisition and interlanguage studies. We can also make our own corpora of song lyrics, fairy stories, news items, texts about fishing or swimming, and of our students' writing.

As a weapon in the armoury of language study and teaching, it is still early days for the use of corpora and concordancers. Given that many teachers and students have ready access to computers and the internet, that DDL came with a sound pedagogical pedigree, and the steady growth in e-learning, it seems likely that sooner than later, consulting corpora will become a standard instrument in revealing the unknown knowns in language deployment.

References

Biber, D., Johannson, S., Leech, G., Conrad, S. & E.Finegan (1999). *Longman Grammar of Spoken and Written English*. Longman

Hanks, P. (1998). *The New Oxford Dictionary of English*. Oxford: Oxford University Press

Hatch, E. & C.Brown (1995). *Vocabulary, Semantics, and Language Education*. Cambridge: Cambridge University Press.

Lewis, M. (1993). *The Lexical Approach*. Hove: Language Teaching Publications.

Krishnamurthy, R. (2001) "Learning and Teaching through Context - A Data-driven Approach." http://www.developingteachers.com/articles_tchtraining/corpora1_ramesh.htm

Norris, R. (2001). *Ready for First Certificate*. Oxford: Macmillan.

Rundell, M. (Ed.) (2002). *Macmillan English Dictionary for Advanced Learners.* Oxford: Macmillan.

Sinclair, J.M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J.M. (ed.) (1995). *Cobuild English Dictionary*. 2nd edition. Collins

Tarone, E. and G. Yule (1989). *Focus on the Language Learner*. Oxford: Oxford University Press.

Thornbury, S. (2002). *How to Teach Vocabulary*. Harlow: Pearson Education.

Notes

[i] http://nlp.fi.muni.cz/lexicom2005/
[ii] http://www.lexmasterclass.com/
[iii] http://www.rotten.com/library/bio/usa/donald-rumsfeld/

[iv] http://www.collins.co.uk/Corpus/CorpusSearch.aspx. Some years ago I created a web-site called “A Ten-step Introduction to Concordancing through the Collins Cobuild Corpus Concordance Sampler” which can be found at http://www.fi.muni.cz/~thomas/CCS/.

[v] The Word Sketch Engine evolved from the program *Bonito*. It is a web-based concordancing program. The sampler version which can be found at http://www.sketchengine.co.uk/ uses the British National Corpus. To register for the sampler, go to http://www.sketchengine.co.uk/reg/reg.cgi/registration_form. There is also another website linked to that explaining its functions and how to create searches: The Sketch Engine Uswer Guide at http://www.sketchengine.co.uk/Sketch-Engine-User-Guide.htm
[vi] http://www.natcorp.ox.ac.uk/
[vii] http://www.fi.muni.cz/~thomas/EAP/take_WSE
[viii] http://www.fi.muni.cz/~thomas/EAP/device_WSE
[ix] http://www.fi.muni.cz/~thomas/EAP/adj+device_WSE
[x] See also http://www.iatefl.org.pl/call/j_soft18.htm
[xi] http://www.ecml.at/projects/voll/our_resources/graz_2002/ddrivenlrning/whatisddl/resources/tim_ddl_learning_page.htm
[xii] http://www.lsa.umich.edu/eli/micase/kibbitzer.htm
[xiii] http://www.lsa.umich.edu/eli/micase/index.htm
[xiv] http://www.edith.nl/telmie2/reforped/princ/princ.html
[xv] "moot" occurs 67 times as an adjective in the BNC, 43 times in the phrase "moot point".
[xvi] in personal correspondence, Sept 2004.
[xvii] This collocation occurs 51 times in the BNC. This is the fifth most frequent adjective preceding *volume* after *large, total, free, high.*
[xviii] http://www.oup.com/elt/catalogue/teachersites/oald7/?cc=global