Center for Advanced Language Proficiency Education and Research
The Pennsylvania State University

Using a Corpus in Language Teaching

by

Michael McCarthy
CALPER & University of Nottingham

October 2004
CALPER Professional Development Document (CPDD) 0410

What is a corpus?

Nowadays, the types of information we need to create good language teaching materials and resources, and information which can help us understand better how learners develop, can be greatly enhanced by the use of a computerized **corpus** (plural: **corpora**). A corpus is simply a collection of texts, written or spoken, usually stored in a computer database. A corpus may be quite small, for example, containing only 50,000 words of text, or very large, containing many millions of words. The largest corpora now hold in excess of half a billion words of text, and new corpora are being constructed all the time.

Written texts in corpora are usually drawn from books, newspapers, or magazines that have been scanned or downloaded electronically. Such corpora help us to see how language is used in contemporary society, how our use of language is changing, and how language is used in different text types.

Spoken corpora are made up of transcripts of spoken language (and, increasingly, transcripts accompanied by sound and/or video files). These transcripts may be of ordinary conversations recorded in people's homes and workplaces, out in the shopping mall, or transcripts of phone calls, business meetings, radio broadcasts, or TV shows. Like written corpora, spoken corpora show us how language is actually used in real life and in different contexts.

People build corpora of different sizes for specific reasons. For example, corpora used in dictionary production usually contain tens of millions of words, because the corpus has to include as many examples as possible of all the words and expressions in the language. A medium-sized corpus of a few million words, on the other hand, consisting of transcripts of lectures and seminars, could be used to design materials for learners who need academic language for their studies. Other corpora are even more specialized and much smaller.

One of the most important kinds of corpora are learner corpora, where the written or spoken production of learners is stored in the computer and can be tracked over time, following the development of an individual, of a group, or of a particular cohort (e.g. females, sophomores, Spanish learners in high school, etc.). Learner corpora can give us unique insights into how learners acquire language over time. Developing learner corpora and helping teachers to build their own learner corpora is one of CALPER's goals.

Once a corpus is stored in a computer, we can use readily available software to analyze it and "search" for information in the same way we use search engines to find keywords on the Internet, but with more sophisticated tools. By searching a corpus we hope to get answers to questions such as:

- What are the most frequent words and phrases in the language in question?
- What are the differences between spoken and written usage?
- Which grammatical structures do people use most frequently, or least frequently?
- How many words must a learner know in order to participate in everyday conversation?
- How many different words do native speakers generally use in conversation?
- How well are my learners progressing? Is their vocabulary range growing?
- Are some learners doing better than others? In what ways is this manifested?

Corpora now exist for many languages and are collected for different purposes. Learner dictionaries, grammar reference materials, vocabulary learning materials, and, more recently, course books have all benefited from the information in corpora. And learner corpora are making a difference in our understanding of questions such as *What does it mean to be an advanced user of a language?*

A corpus in itself is not a theory of language learning or a teaching methodology, but it does influence our way of thinking about language and the kinds of texts and examples we use in language teaching.

Analyzing corpora

The most basic tool for analyzing the texts in a corpus is the **frequency list**. A frequency list tells us what words and phrases are used most often. Here is a frequency list for the top 50 words in spoken North American English, based on a sample of spoken data from the American National Corpus. The most frequent word – *I* – is at the top of the list.

As we can see, all the top 50 words occur thousands of times in this corpus, so there is a huge amount of information we can learn about each word. The top 20 words occur more than 40,000 times each.

	word	frequency
1	I	180,977
2	and	149,925
3	the	145,918
4	you	123,771
5	uh	112,031
6	to	105,596
7	a	101,731
8	that	93,381
9	it	82,708
10	of	76,347
11	yeah	67,740
12	know	65,808
13	in	57,835
14	like	48,098
15	they	45,205
16	have	43,455
17	so	42,941
18	was	41,453
19	but	40,892
20	is	40,068
21	it's	38,768
22	we	37,362
23	huh	36,495
24	just	32,650
25	oh	31,263

	word	frequency
26	do	30,330
27	don't	29,231
28	that's	29,188
29	well	29,059
30	for	28,687
31	what	27,038
32	on	26,360
33	think	25,020
34	right	24,383
35	not	23,123
36	um	22,998
37	or	22,779
38	my	22,539
39	be	22,325
40	really	20,838
41	with	20,797
42	he	20,732
43	one	20,552
44	are	20,347
45	this	20,239
46	there	20,008
47	I'm	19,802
48	all	19,713
49	if	19,263
50	no	18,908

Figure 1: The top 50 words in a north American spoken English corpus

Here are some observations a corpus analyst would typically make from this list:

- In a spoken corpus, *I* and *you* are among the most frequent words of all. This is because conversation is very interactive; it's not surprising that *you* and *I* feature prominently. In a written corpus, however, *I* and *you* appear less frequently, because written texts are usually about "the world out there," so third-person subjects (politicians, celebrities, etc.) predominate.
- Most of the top 50 words are grammar words (pronouns, prepositions, articles, demonstratives, conjunctions, auxiliary verbs, etc.), but not all of them.
- *Know, right, really,* and *think* are not grammar words; they occur frequently because of the expressions *you know* and *I think*, because we use *right* to agree with someone, and because of the way we use *really* to react to things people say.

Even from this short list of 50 items we can learn a lot about how people communicate, and this information can be used to design appropriate materials and activities for the conversation class.

Frequency lists also help us to set different levels for language learning. For example, in English, the top 1,800 or so words in the spoken frequency list are much more frequent than all the other words in the list. There is a sharp fall-off in frequency after the first 1,800, which shows us that these 1,800 words work much harder than all the others. In fact, these 1,800 words make up more than 80 percent of all the words in all the texts in the corpus. We can therefore say that learners of English who want to be able to participate in everyday conversation must know at least these 1,800 words, or they will simply not be able to put together even a basic string of sentences. They will of course need a lot of other words to talk about themselves and the world around them, but the basic 1,800 words are the cement that holds the whole language together. This will be more or less true for any language being learnt, not just English.

Even a very small corpus can provide us with a lot of grammatical information. In a corpus of only 6,000 words of Spanish newspaper texts, we get hundreds of examples of the core grammar words such as *de* (*of* in English), the definite articles *el* and *la*, and the preposition *en* (*in*), all of which can be analyzed in greater detail using concordances (see below).

Figure 2: The top 30 words in a Spanish newspaper corpus

	word	frequency		word	frequency
1	DE	409	16	PARA	39
2	LA	271	17	SU	32
3	EL	167	18	NO	31
4	EN	160	19	AL	28
5	QUE	132	20	ES	26
6	Y	125	21	HA	26
7	A	112	22	HAN	20
8	LOS	109	23	ENTRE	17
9	DEL	102	24	LO	17
10	LAS	82	25	MÁS	17
11	POR	69	26	SUS	17
12	SE	67	27	COMO	15
13	UN	58	28	CUANDO	15
14	CON	55	29	FUE	15
15	UNA	51	30	CONTRA	14

When words occur thousands of times in a corpus, it is difficult to make sense of all the uses of them, so software designers have come up with simple tools to reduce the workload. One of these is the **concordance**. A concordance is a screen display or printout of a chosen word or phrase in its different contexts, along with the text that comes before and after it. Figure 3 is a sample (the first screen-shot) of a concordance for the word *de* in the Spanish news corpus. The researcher can look at screen after screen, and see *all* the different occasions in which the writers in the corpus have used *de*. This screen shows a random sample of contexts. The software arranges ***de*** vertically in the middle of the computer screen and, in this case, arranges what follows it in alphabetical order. Here we can see that what follows ***de*** is often a time expression, referring to years, months, parts of the day, etc (*de 1997, de la mañana, 27 de agosto, 7 de marzo*). This is useful information if you want to teach expressions involving the months or days (e.g. in English *June 28th , 5 o'clock in the morning*, etc.). We also see that ***de*** is used with people's ages (*de 23 años, de 72 años*).

1	ido u otro la situación creada a partir	de	1997, en que la inesperada victoria d
2	o surgieron, según se ha dicho, a fines	de	1999, como resultado de una escisión
3	ran contienda electoral de la primavera	de	2002, que enfrentará a la derecha y l
4	mo de los detenidos fue Jon Etxeberria,	de	23 años, durante la madrugada del pas
5	omingo, cuando Saphir Bghioua, un joven	de	25 años y con antecedentes por robo d
6	en vigor inmediatamente para un periodo	de	30 días, pero el Congreso Legislativo
7	illo para permitir el acceso al colegio	de	45 niñas, todas menores de 11 años, m
8	tomática a Jean Faret, un ex legionario	de	72 años, jefe de gabinete del alcalde
9	s que dos de los detenidos el pasado 27	de	agosto tras la caída del "comando Bar
10	do José Ramón Acedo Espina el pasado 29	de	agosto. Al día siguiente, la Guardia
11	llada por la Guardia Civil el pasado 24	de	agosto. Tras la desarticulación, agen
12	en el tabaco se ha vendido a los medios	de	comunicación nada más y nada menos qu
13	tituciones benéficas, empresas y medios	de	comunicación para donar alimentos y m
14	ha volado de nuestras manos. Los medios	de	comunicación, espoleados por el márke
15	mó nota" de la citada Carta en el texto	de	conclusiones de la cumbre de Niza. Pa
16	ación Portillo no contaba con la sequía	de	este año, que se ha juntado con el ba
17	breza que se han agravado por la sequía	de	este año. La decisión de decretar el
18	ntre los franceses. Los últimos sondeos	de	julio le dan, incluso, claramente ven
19	tras la dimisión de Trimble el pasado 1	de	julio. Las dificultades están centrad
20	gratuita de los suplementos literarios	de	junio y, después, se encuentran, de l
21	aza al hombre" se prolongó hasta las 11	de	la mañana del domingo, cuando Saphir
22	durante el consejo de ministros regular	de	la mañana, dijo Park Jun Yong, aunque
23	idense, George W. Bush. La aprobación	de	la moción de censura rompió la coalic
24	ios de junio y, después, se encuentran,	de	la noche a la mañana, con unos periód
25	homicida del agresor, quien a lo largo	de	la noche telefoneó varias veces a com
26	un confuso horizonte en la articulación	de	la nueva Europa, Francia se prepara p
27	ad de la doble gran contienda electoral	de	la primavera de 2002, que enfrentará
28	amanecer a la decadencia"- la historia	de	los últimos 500 años de cultura occid
29	los combates políticos más trepidantes	de	los últimos tiempos. El año 2002 no
30	a en una gasolinera tras ser movilizado	de	madrugada. El faro giratorio instalad
31	ellos. Se legitiman a sí mismos. El 7	de	marzo pasado, cuando la caravana del
32	status penal del presidente. A primeros	de	octubre, en efecto, el Tribunal de Ca
33	ntrée" singularmente caliente. A partir	de	septiembre, la vida pública francesa

Figure 3: Sample from a concordance of *de* in a Spanish newspaper corpus

Corpora are also useful for studying **collocations**, which are one of the aspects of acquiring a foreign language vocabulary which many learners have difficulty with. Collocation is the likelihood that two words will occur together. So, for example, the word *fair* is likely to be used with *hair*, but not with *car* or *jacket*. We say that *fair* **collocates** with *hair*, but *beige* does not.

With a large corpus of millions of words, we can use our software to generate collocation statistics. The software presents us with statistics in the form of simple tables that show us which words occur together most frequently. Here is a sample of the top 10 adjectives that follow the adverb *pretty* (e.g. *it was pretty good*), and how often they occur in a seven million-word spoken sample of the American National Corpus.

	pretty +	frequency
1	good	1032
2	nice	102
3	cool	92
4	bad	87
5	big	87
6	close	84
7	neat	70
8	soon	70
9	sure	64
10	easy	53

Figure 4: The 10 top adjectives that follow the adverb *pretty*, from the American National Corpus (spoken sample)

Imagine if you had to count those 1032 examples of *pretty good* by hand! The software does it in a split second. With this knowledge, we can present the most frequent and useful contexts for *pretty*, or any other words or phrases in the corpus. Learning words along with their most frequent collocates is a good learning habit that can start right from the lowest levels, based on reliable information.

Applications in language teaching

An example from the teaching of English as a second language shows how corpora have been used in the design of teaching materials. A major new English course called *Touchstone* (published by Cambridge University Press) is based on a large North American English corpus. The *Touchstone* authors spent several years researching the corpus, finding the most useful grammar and vocabulary for learners from a basic to an intermediate level, and finding out how people communicate in everyday situations, especially in conversation.

One way the *Touchstone* authors used the Corpus was to look for the most frequent and typical uses of everyday words. For example, how do people most typically use the verb *can*? As well as having the meaning of "ability" (e.g., *I can swim under water*), which most English teachers are familiar with, conversations in the spoken corpus showed that an extremely common use of *can* occurs when people talk about what it is possible to do in

different places and situations (e.g., *In New York, you can go to the top of the Empire State Building*). So the coursebook includes this meaning and gives it priority.

Of great interest to all language teachers is the information we can get about how learners deal with language and how their development can be tracked. By entering student essays and other written work, learner journals, emails, internet chat or web-log ('blog') data, transcripts of oral examinations, transcripts of learners doing tasks, etc., we can track development over time. We can measure whether the vocabulary they are using becomes larger, more varied, whether a grammatical structure is being used correctly, and how long it is before a structure really becomes embedded and 'learnt'. This is all becoming easier now that students are doing a lot of their work on computers and online. The CALPER research team at Penn State is particularly interested in gathering and analyzing these kinds of data.

The future

In the future we can expect bigger corpora, consisting of billions of words, and software that will be able to transcribe conversations automatically. We can expect more sophisticated tools to do the kinds of searching that at present cannot be done automatically.

We can also expect more and better corpus-informed teaching materials, perhaps in electronic format, on DVD, or accessible on the Internet. These materials will include hyperlinks to actual corpora or to corpus samples so that teachers and students can explore and investigate language more easily for themselves. And we can expect more user-friendly home- and classroom-based facilities with which teachers and learners can build and explore their own corpora. Corpora will become more sophisticated and ever more finely tuned to our needs. The future is exciting, and corpora are here to stay.

Further reading

First and foremost, visit our corpus pages at CALPER: http://calper.la.psu.edu/corpus.php, then click on the CORP links for a wide range of information about corpora.

The following books and articles are recommended if you want to learn more about corpora.

Biber, Douglas, Susan Conrad, and Randi Reppen. (1998). *Corpus Linguistics. Investigating language structure and use.* Cambridge: Cambridge University Press.

Hunston, Susan. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

O'Keeffe, Anne, & Fiona Farr. (2003). "Using language corpora in initial teacher education: pedagogic issues and practical applications." *TESOL Quarterly* 37 (3): 389–418. This article contains a useful list of Web sites connected to the study of corpora.

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

McCarthy, Michael (2004). "Using a corpus in language teaching" (CALPER Professional Development Document 0410). University Park, PA: The Pennsylvania State University, Center for Advanced Language Proficiency Education and Research.

This CALPER PDD can also be downloaded at http://calper.la.psu.edu/publications.php

Contact Information:
James P. Lantolf and Karen E. Johnson, CALPER Co-directors
Center for Advanced Language Proficiency Education and Research
The Pennsylvania State University
5 Sparks Building
University Park, PA 16802-5203
Tel: (814) 863-1212
Fax: (814) 865-1316
Email: calper@psu.edu

This publication was supported by a grant from the U.S. Department of Education (CFDA 84.229, P229A0200). However, the contents do not necessarily represent the policy of the Department of Education, and one should not assume endorsement by the Federal Government.