Electronic Corpora as Very Rich Resource in Text Book Design

Mohamed GRAZIB
Université de Tahar Moulay -Saida

Abstract:
"Electronic corpora as very rich resource in text book design" is an attempt to shed light on the importance of the uses of electronic corpora such as the British National Corpus (BNC) or/and other English corpora in the design of the official English text books in English for non native speakers.
Thanks to advanced technology, electronic corpora have become now an important resource in applied linguistics and in text book design they present many advantages if compared to the manual and traditional ways to make linguistic investigations.
These powerful tools such as the British National corpus (BNC), the American National Corpus (ANC) and other important corpora are freely available on the web, they present results in the form of concordances, collocations, frequencies and word lists that are of great importance to linguists and text book designers to reach very best results.
The study will also focus mainly on the methodology used in order to adopt such powerful tools in the text book design and the necessary parts needed to be exploited in this process which consists on the design of pedagogical supports for pupils and students in order to learn English as a foreign language.
Key words: Corpora, BNC (British National Corpus), applied linguistics, text book design, word lists, frequencies, concordances and collocations.

1. Introduction:

Electronic corpora are of great importance in linguistic studies. Nowadays, it is impossible to investigate any linguistic field without the uses of computational linguistics or corpus linguistics. The developed countries give more importance to text books and to the way they are designed, mainly in what concerns the content which is the first purpose for which these text books are created.

"Electronic corpora as very rich resource in textbook design" will shed light on the importance of electronic corpora in textbook design such as the British National Corpus which contains 100 million words available in software form. The speed, the amount of data processed in few seconds, the storage capacity and the accuracy are very important elements that help linguists to perform their tasks in very suitable conditions and let positive impact on the learning process in general.

The focus will be on the necessary types of data used by the text book designers when exploring corpora and how they can use this very rich resource in text book design?

2. What is a corpus?

A corpus is a large set of written and/or spoken texts electronically stored and processed; it is presented in software forms in computers or via the net; as Mc Enery and Wilson (2001:197) say: "A corpus is a finite collection of machine-readable texts, sampled to be maximally representative of a language or variety".

Corpora are often subjected to a process known as annotation (tagging and parsing). Annotating corpora is somehow the process which makes the inert corpora in live.

The size of corpora varies from some thousands of words to some millions of words:

- Bank of English Corpus: About 400 million words
- British National Corpus (BNC):100 million words
- Longman Lancaster Corpus: 30 million words
- American National Corpus (ANC): 11.5 million words
- Brown corpus: 1million words

Reference corpora are corpora with a fixed size and they are not expandable (e.g., the BNC); in contrast other corpora are expandable and texts are continuously being added (e.g., the Bank of English). The BNC World Edition contains 4,054 texts the equivalent of 100 million words.

3. Composition of the British National Corpus (BNC):

The BNC is composed mainly by 90 per cent (90 million words) of written texts and 10 per cent (10 million words) of spoken texts, Burnard, L. (2000). It is formed by:

- 4,054 texts
- 97,619,934 sentences
- 100 million words

3.1. Domains

Domains	Texts	%	Domains	Texts	%
Imaginative	477	21.91	Applied science	370	8.21
Arts	261	8.08	Social science	527	14.80
Belief and thought	146	3.40	World affairs	484	18.39
Commerce/Finance	295	7.93	Unclassified	51	1.93
Leisure	438	11.13	Natural/pure science	146	4.18

Table 1: Domains (L.Burnard: 2000)

As seen in the above tables, the written texts present the lion part in the BNC (3,144 texts), however the spoken texts to be converted into written texts present only 10 %(910 texts). The written texts cover all the domains such as arts, literature, commerce, science, leisure... Their origins are mainly from books (58.58%), periodicals (31.08%) and others (10.3%).

4. The useful corpus

In order to make corpora useful and ready for doing linguistic research, they are often subjected to a process known as annotation (tagging and parsing). Annotating a corpus means that information about each word's part of speech (verb, noun, adjective, etc.) is added to the corpus in the form of tags. By adding information to a corpus it will be easy to retrieve data easily and with great precision.

The following example of "***light***" [1] illustrates the importance of corpus annotation:

Examples	Frequencies
Light touch	59
Light wind	49
Light breeze	41
Light rain	28

Table 2: Parts of speech of (light) as adjective

[1] : from "Just the word"

Examples	Frequencies
Light a cigarette	681
Light fire	227
Light a candle	133

Table 3: Parts of speech of *(light)* as verb

Examples	Frequencies
red light	321
Traffic light	275
Shed light	194
Green light	191

Table 4: Parts of speech of *(light)* as noun

As seen in the tables above, the corpus annotation is very important, so without this process (tagging and parsing), a corpus will be unable to distinguish between *(light)* as noun, *(light)* as adjective and *(light)* as verb.

So thanks to annotation process, corpora can present results to linguists in 3 ways:

-Frequencies
-Concordances
-Collocations

5. Frequencies, Concordances and Collocations

Working with corpora means working with statistics which exist in the form of concordances, collocations and frequencies. These three forms will give linguists, lexicographers and text book designers, precious data about any word. (Sinclair, J.M.1991).

6. The benefits of frequencies in text book design:

Hunston.S. (2002:67) says: "Frequencies is a list of all types in a corpus together with the number of occurrences of each type ". Nobody can deny the roles of corpora in linguistic domains; they are among the main resources for both teachers and learners.

The following table shows the list of the 100 commonest English words found in writing around the world:

1	the	26	they	51	when	76	come
2	be	27	we	52	make	77	its
3	to	28	say	53	can	78	over
4	of	29	her	54	like	79	think
5	and	30	she	55	time	80	also
6	a	31	or	56	no	81	back
7	in	32	an	57	just	82	after
8	that	33	will	58	him	83	use
9	have	34	my	59	know	84	two
10	I	35	one	60	take	85	how
11	it	36	all	61	people	86	our
12	for	37	would	62	into	87	work
13	not	38	there	63	year	88	first
14	on	39	their	64	your	89	well
15	with	40	what	65	good	90	way
16	he	41	so	66	some	91	even
17	as	42	up	67	could	92	new
18	you	43	out	68	them	93	want
19	do	44	if	69	see	94	because
20	at	45	about	70	other	95	any
21	this	46	who	71	than	96	these
22	but	47	get	72	then	97	give
23	his	48	which	73	now	98	day
24	by	49	go	74	look	99	most
25	from	50	me	75	only	100	us

Table 5: The first 100 commonest English words (from the BNC)

We can also explore frequencies according to the main word classes:

Nouns		Verbs		Adjectives	
1	time	1	be	1	good
2	person	2	have	2	new
3	year	3	do	3	first
4	way	4	say	4	last
5	day	5	get	5	long
6	thing	6	make	6	great
7	man	7	go	7	little
8	world	8	know	8	own
9	life	9	take	9	other
10	hand	10	see	10	old
11	part	11	come	11	right
12	child	12	think	12	big
13	eye	13	look	13	high
14	woman	14	want	14	different
15	place	15	give	15	small
16	work	16	use	16	large
17	week	17	find	17	next
18	case	18	tell	18	early
19	point	19	ask	19	young
20	government	20	work	20	important
21	company	21	seem	21	few
22	number	22	feel	22	public
23	group	23	try	23	bad
24	problem	24	leave	24	same
25	fact	25	call	25	able

Table 6: Frequencies according to the main word classes (from the BNC)

7. The benefits of concordances in text book design:

Concordances are the most frequent sentences used by native speakers in real situations; they can serve as citations to explain difficult words in recent dictionaries rather than using famous persons' citations such as

Shakespeare. John Sinclair (1991:32) says: "A concordance is a collection of the occurrences of a word-form each in its own textual environment ".
The following list shows **"Thank you"**[1] concordances taken form the BNC. (from Web concordancer)

*1- **Thank you** very much.*
*2- **Thank you** for your company.*
*3- Very well, **thank you**.*
*4- **Thank you** very much for your letter.*
*5- **Thank you** for helping me.*
*6- I wanted to **thank you**...*
*7- I can't **thank you** enough for this evening.*
*8- **Thank you** for your letter giving further information about this application.*
*9- `That was very kind of you, **thank you**'; she said gratefully.*
*10- Yes, **thank you** very much.*
*11- **Thank you** again for writing.*
*12- **Thank you** very much indeed.*

By using these results taken from the British National Corpus software the text book designers will find any difficulty to select:

- Very easy sentences
- Very frequent sentences
- Very understandable sentences

---

[1] : A random selection of **12** solutions from the **9598** found in the BNC software.

- Expression taken from the source (the language that is really used by native speakers)

8. The benefits of collocations in text book design:

J. Firth (1957) stated that "you shall know the word by the company it keeps". Collocations are the statistical tendencies of words that co-occur with other words.

This table shows the "**time's**" collocations with **"verbs"**, and "**adjectives"**:

	Verbs	**Frequencies**
time +	was	17846
	Is	12614
	had	8128
	Be	8023
	were	4298

Table7: collocations of time (+) verbs (from the BNC software)

Adjectives		**Frequencies**
long	**+ time**	4850
good		1587
short		1522
other		1202
right		1111

Table 8: collocations of time (+) adjectives (from the BNC software)

Sometimes the concordances can't give precise information to the text book designers, who are more

interested by the immediate environment of the target word(s), that is/are at the left and/or the right of the target word to form very useful expression needed to develop the language capacities of learners.
For example:
The word "**time**" is mainly used with the auxiliary (to be) and with short adjectives:
- "**Long time**": is used **4850** times in the British National Corpus.
- "**Good time**": is used **1587** times in the British National Corpus.
- "**Short time**": is used **1522** times in the British National Corpus.
So thanks to the statistics taken from the corpus, text book designers will find any obstacle to select what they see necessary and important for the learners and mainly for the beginners. All what the text book designers are asked to do is to make logical analyses and serving priorities for which the books are designed for.

9. The benefits of corpora in text book design:
At the end, one can notice the great importance of corpora and the rich information they present to linguists in general and to text book designers in particular.
Today corpora have made a revolution in linguistics, they have changed the way analyses and investigations are performed qualitatively and quantitatively. They are used in many linguistic domains and fields among them: lexicography, dictionaries, sociolinguistics, register studies, ESP, translation and other fields.... So thanks to electronic corpora; now, and even for beginners, it is possible to make advanced research by:
- Frequency: Essential /Advanced

- Parts of speech: Nouns, verbs, adjectives, adverbs, pronouns.....
- Grammar: Countable nouns, uncountable nouns, singular nouns, plural nouns, transitive verbs, intransitive verbs....
- Usage: Child's words, teenagers' words, formal, informal, Internet, legal, literary, old fashioned words, old used words (historical), polite words ,slang...
- Region: British English. American English, Australian English, Canadian English, South African English....
- Topic: Biology, chemistry, communication technology, computer science technology, education, finance, business, medicine, politics, religion, sports....
- Registers: Academic, fiction, ESP

This leads immediately to help text book designers to select:

- Very easy words
- The most frequent words in spoken and written English
- Very easy sentences
- Very frequent collocations
- The necessary list of words that a learner needs first in his text book in order to understand and use gradually the target foreign language.

10. Conclusion:

Nowadays, almost all the linguistic domains are investigated with the help of corpora or computational linguistics. Via text books, corpora bring to classrooms abundant examples of authentic language used by native speakers.

So, corpora are important resources to text book designers, who can easily collect and then select words, expressions, and sentences from authentic language.

What we have got as results are only drops from an ocean, but what remains will be without any doubt of paramount importance. These results and analyses can never be done without the helps of computers and corpora software. Corpora can't replace definitively the existing text book design methods, but they are used mainly to assist, enrich and enhance them.

To sum up, we can notice that corpora are just tools to save very large texts in electronic forms; software that accompanies these corpora are the main elements that make the inert corpora useful and alive what remains, is that linguists, lexicographers and text book designers perform the adequate analyses from the results obtained in the forms of concordances, collocations and frequencies.

Bibliography:

- Biber Douglas. , Susan Conrad. & Randi Reppen. (1998) Corpus linguistics: "Investigating language structure and use" . Cambridge: University Press.
- Burnard, L. (2000). Reference Guide for the British National Corpus (World Edition). Oxford: Oxford University Computing Services.
- Firth, J.R. (1957) Papers in Linguistics: 1934-1951, London: Oxford University Press.
- Francis, W,N. & Kucéra,H. Frequency analysis of English usage . Boston,MA: Houghton Mifflin. (1982)
-Hunston, S. Corpora in applied linguistics. Cambridge University Press. (2002)
-Kennedy G. An introduction to corpus linguistics . London :Longman. (1998).
- Leech, G. (1997). Introducing Corpus Annotation. Garside et al (eds.), pp 1-18 - McEnery, T. & Wilson, A. Corpus linguistics. Edinburgh University Press. (2001, 2nd ed.)
-Sinclair J.M. Corpus Concordance Collocation. Oxford :OUP.(1991).
- Stubbs M.(1996). text and corpus analysis .Oxford: Blackwell. (1996).

Web-sites:

- http://www.bnc.byu.edu
- http://www. Jtw.edu

CV

Mohamed GRAZIB
Teacher at the English Department
Interests: Didactics and computational linguistics
E-mail: mfgrazib@hotmail.com