

Electronic Corpora and Textbook Design

Mr: Mohamed Grazib
mfgrazib@hotmail.com
Tahar Moulay University. Saïda

1. Introduction:

Electronic Corpora and Textbook Design is an attempt to shed light on the importance of corpora in the design of the recent textbooks and the advantages presented if compared to the traditional ways to make linguistic investigations. 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 the text book design. In this paper we try to shed light on the importance of corpora and necessarily the British National Corpus which contains 100 million words in applied linguistics in general and in the textbook design in particular.

2. What is a corpus?

The word Corpus plural (corpora) or (corpuses) is derived from the Latin word 'corpus' which means: 'Body' in French 'corps'. A corpus is a large set of texts (electronically stored and processed); it may be used to refer to any text in written and/or spoken form. The size of a corpus varies according the total words it contains:

- Bank of English: 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.
- International Corpus of Learner English (ICLE): 1 million words.

In this paper we shall take only the BNC¹ as example.

¹ : The BNC (British National Corpus) is a reference for the English language .

3. Composition of the BNC (British National Corpus):
The BNC is composed mainly by 90 per cent (90 million words) of written texts and 10 per cent (10 million words) of spoken texts¹. It is formed by:

- 4054 texts
- 97,619,934 sentences
- 100 million words
-

2. 1. The written books:

The British National Corpus contains 90 million written words, and one printed page contains 500 words on average; so the BNC contains the equivalent of 180,000 pages of written texts. The written texts present the lion's part in the BNC 90%. The total number of the written books and periodicals is 2688 books, and they are mainly from:

- Books (58.58 %),
- Periodicals (31.08%)
- Others (10.3%).

In what concerns the selection of books, the first half was selected according to the following criteria:

- To be published by a British publisher
- To contain sufficient pages of text to make its incorporation worthwhile
- To cost less than a set price.

The second half of books was selected systematically to make balance between the different categories. The selection proceeded as follows: Bestsellers, Literary prizes, Library loans, Periodicals and magazines.

3.2. The domains:

The domains explored in the BNC written books are mainly:

- Commerce and finance and world affairs: 779 texts
- Social science: 527 texts
- Imaginative: 477 texts
- Leisure: 438 texts
- Arts: 261 texts

¹ The spoken texts are recorded files converted into written texts.

- Belief and thought: 164 texts
- Natural science: 141 texts

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 “deal” illustrates the importance of corpus annotation:

Words	Parts of speech	Frequencies
deal	singular common noun	115
deal	proper noun	1
deals	plural common noun	5
deal	base form of the verb	66
dealing	present participle	51
deals	third person singular form of the verb	20
dealt	past tense of the verb	14
dealt	past participle	17

Table 1: The relative frequencies of “deal” from the LOB Corpus

As seen in the table above, the corpus annotation is very important, so without this process (tagging and parsing) a corpus, this later will be unable to distinguish between “deals” as plural common noun and “deals” as third person singular form of the verb.

So the corpus presents results to the linguists in 3 ways:

- Frequencies
- Concordances
- Collocations

5. The benefits of the frequencies in the textbook design:

A frequency list is simply a list of all the part of speech types in a corpus with the number of occurrences of each type. The list can be displayed in frequency order or in alphabetical order.

Beautiful + Word	Frequency
woman	191
Girl	116
things	84
house	79
countryside	76
place	74
face	73
hair	61
eyes	61
scenery	45

Table 3: collocation list of “beautiful” From the BNC.

6. The benefits of collocations in the textbooks design:

The right and the left collocations of “*country*”

Left collocates of “country”		Right collocates of “country”	
the country	13	country has	6
your country	8	country needs	4
a country	7	country In	3
each country	3	country can	2
by country	1	country however	2
alternative country	1	country should	2
foreign country	1	country side	2

Table 4: The right and the left collocations of “*country*”

The results obtained from corpora are of paramount importance because linguists find very precise information about a target word:

- Left collocations
- Right collocations
- And their respective frequencies

If we want to replace the adjective **beautiful** with its near synonyms, one can select immediately the adjective **pretty** (or perhaps another adjective). It seems to us that there is no difference between them, but in reality we have so much information to detect from corpora.

The first information is that “**beautiful**” is the most used word among a list of synonyms¹:

Words	frequency
Beautiful	8394
Lovely	6037
Attractive	5051
Wonderful	4671
Striking	2576
Superb	2063
Magnificent	1972
Handsome	1578
Charming	1346
Delightful	1069

Table 2: synonyms of “beautiful” From the BNC.

The second information is that “**Beautiful**” is mainly used with: woman, girl, house, countryside, place, face, hair, eyes...

But “**Beautiful woman**” is more used than “**beautiful girl**”.

“**Beautiful woman**” is used 191 times

“**Beautiful girl**” is used only 116 times.

However if an advanced research will be made, one will find that “**pretty**” is differently used:

“**Pretty girl**” is used 127 times

“**Pretty woman**” is used only 68 times²

¹ Frequencies obtained from the British National Corpus

² Hunston, S. Corpora in applied linguistics. Cambridge University Press. (2002)

7- The benefits of the concordances in the textbooks design:

Concordances for *school*¹

- 1- To learn more after **school**
- 2- With roads, shops, and **school** buildings.
- 3- I was very excited and arrived **school** at 8.30 am to help out moving equipment.
- 4- And necessary machine not only at **school** and office, but also at home.
- 5- We join together at **school** at about 8:30 a.m.
- 6- But at **school**, we don't really have enough chance.
- 7- At **school**, we learned the process of doing subjects
- 8- We apply the theory we have learnt in **school**.
- 9- In school, we do the practical exercise everyday.
- 10- I understand what I have learnt in **school**.

Concordances are very important in textbook design because they present data as it is naturally used from native speakers. From these concordances and thanks to computer software one can pick up directly the main collocations used with the target word.

So in order to train students to write topics about schools for example, teachers must find such sentences in their textbooks and should present these words and encourage students to form the same sentence types.

8. Conclusion:

Nowadays, almost all the linguistic domains are investigated with the help of corpora or computational linguistics. Via the text books, corpora present abundant numbers of real examples from authentic language used by native speakers. So, corpora are important resources to the text book designers, who can easily collect and then select words, expressions, and sentences.

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

¹ Concordances for *(school)* from the BNC

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.

Bibliography:

- Biber Douglas. , Susan Conrad. & Randi Reppen. (1998) Corpus linguistics: "Investigating language structure and use" . Cambridge: University Press.
- Biber, D, et al (1999).Longman grammar of spoken and written English. Longman.
- Burnard, L. (2000). Reference Guide for the British National Corpus . Oxford: O.U.P.
- Firth, J.R. (1957) Papers in Linguistics: 1934-1951, London: Oxford University Press.
- Francis, W,N. & Kucéra,H. (1982) Frequency analysis of English usage . Boston,MA: Houghton Mifflin.
- Hunston, S. (2002) Corpora in applied linguistics. Cambridge University Press.
- Kennedy G. (1998).An introduction to corpus linguistics . London :Longman..
- Long M,H. & Richards J,C (2002) "Series editors preface" for corpora in applied linguistics. S,Hunston. Cambridge :University.Press.
- McEnery, T. & Wilson, A. (2001, 2nd ed.) Corpus linguistics. Edinburgh University Press.
- Sinclair J.M. (1991). Corpus Concordance Collocation. Oxford :OUP.
- Stubbs M.(1996). text and corpus analysis .Oxford: Blackwell.
- bnc.byu.edu (professor: Mark Davies: Brigham Young University)