

Dictionary Approaches by Corpora

GRAZIB Mohamed
Université de Saïda

Abstract

This research is an attempt to shed light on the importance of corpora in the design of the recent dictionaries and official textbooks. The present paper gives examples taken from dictionaries based on corpora such as the British National Corpus BNC and the Oxford Corpus where all the analyses are made automatically with the help of computers and specific software that present the results in the form of frequencies, collocations and concordances.

1. Introduction

Electronic corpora are large set of written and/or spoken texts electronically stored and processed and presented in software forms in computers or via the net. They 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. By adding information to corpora, it will be easy to retrieve data from them in three ways:

- Frequencies
- Concordances
- Collocations

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

- 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
- Brown corpus: 1million words

Some corpora have a fixed size and they are not expandable (e.g., the BNC). They are known as reference corpora; 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 and occupies 1.5 G Bytes. In total, it comprises just over 100 million orthographic words.

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. It is formed by:

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

2. 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:

(light) as adjective:

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

(light) as verb:

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

(light) as noun:

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

Table 1: Parts of Speech of (light)

As seen in the table 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 the corpus presents the results to linguists in 3 ways:

¹. From “just the word”

- Frequencies
- Concordances
- Collocations

3. Corpora and Dictionaries

In the Oxford Advanced Learners Dictionary 7th edition we find what follows:

- 183,500 British and American words , phrases and meanings
- 85,000 example sentences
- 2,000 new words
- 7,000 synonyms and opposites
- 5,000 study words from the sciences, literature, computing and business
- 700 World English words
- 2,600 cultural words
- 96 study pages(essay writing, emails, grammar reference)

The dictionaries based on corpora can show various information of a word starting from its frequency and its collocations. Hunston.S. (2002:96) states: “*Corpora have so revolutionised the writing of dictionaries and grammar books for language learners*”²

Here are some comparisons between dictionaries written without corpora and others designed with the help of corpora. *The (Longman Dictionary of Contemporary English* 2nd edition 1987 :) which was written without a corpus and three

². HUNSTON.S.(2002:96) Corpora in Applied Linguistics. Cambridge University Press.

dictionaries written with the aid of a corpus (Hunston.S.2002:97)³. (*Collins COBUILD English Dictionary* : 1995), (*Longman Dictionary of Contemporary English*:2001) and (*Oxford Advanced learners Dictionary*:2006).

Words	Longman 1987	COBUILD 1995	Longman 2001	Oxford 2006
Know	20 senses	+30 senses	49 senses	+60 senses
Matter	10 senses	+20 senses	36 senses	+40 senses
May	7 senses	+15 senses	22 senses	19 senses
place	20 senses	+30 senses	39 senses	39 senses

Table 2: Comparisons between dictionaries written without corpora and others designed with the help of corpora.

Thanks to corpora which can permit an advanced research by:

- **Category:** Headwords, phrasal verbs, idioms, definitions....
- **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 use words (historical), polite words ,slang...

³. HUNSTON.S.(2002:97) Corpora in Applied Linguistics. Cambridge University Press.

- **Region:** British English, American English, Australian English, Canadian English, South African English....
- **Topic:** Biology, chemistry, communication technology, computer science technology, education, finance and business, medicine, politics, religion sports....

This technology permits also to make an advanced research by using syntax for searching words:

- **Starting with:** (e.g. starting with *bac**):
The results obtained (among a long list) are: *Baccalaureate, bachelor, back, backache, bacon, bacteria*....
- **Ending with:** (e.g. ending with **ive*): *Adjective, attractive, collective, expensive, formative, inductive*...
- **Containing specific characters:** (e.g. containing **ord**): *Accordance, border, coordination, Nordic, oxford, sordid, subordinate*...

4. Concordances, collocations and frequencies:

Working with corpora means working with statistics which exist in concordances, collocations and frequencies.

4.1. Concordances:

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

Concordances for *school*⁴

⁴. 10 first concordances of (school) from the BNC

- 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**.*

4.2. Collocations

Sometimes, we are not only interested by the target word alone or by its immediate meaning, but we are more interested by its surrounding (words that are used before and after). Collocations are necessary to master the real use of the target word as Firth.J.R (1957:196) stated⁵: “*you shall know the word by the company it keeps*”. The following list shows the left and right collocations of **school**⁶

Left collocates		Right collocates	
in	6	we	6
the	6	results	3
at	5	buildings	2
and	2	and	2
primary	2	at	2

⁵. Firth, J.R. (1957) *Papers in Linguistics: 1934-1951*, London: Oxford University Press.

⁶. Left and right collocations of “school” from the BNC

from	2	canteen	1
students	1	Comment	1
high	1	facilities	1
teenagers	1	From	1
Secondary	1	has	1
London	1	pupils	1

Table3: Left and right collocations of “school”

4.3. Frequencies

Nobody can deny the roles of dictionaries in the linguistic domains; they are among the main resources for both teachers and learners. Corpora and Computers make analyses easiest and reliable; they can present a detailed list of a target word. The following example studies the two synonyms: (*big*) and (*large*), if we refer to a traditional dictionary we will find that (*big*) and (*large*) are full synonyms but in reality and according to the statistics obtained from a corpus⁷.

- (*Large*) is used at 57.1% in the English language
- (*Big*) is used at 42.9% in the English language

WORD1: <i>LARGE</i> (57.1%)		Frequency	WORD2: <i>BIG</i> (42.9%)		Frequency
1	Tentacle	140	1	Bang	394
2	Quantity	117	2	Guy	62
3	Bowel	103	3	Bloody	60
4	Pores	77	4	Aye	55
5	Oral	72	5	Daddy	54

Table 4: Collocation of (*large*) and (*big*).

⁷. Frequencies from the BNC (BYU)

Through a simple analysis one can distinguish that (*large*) is used more than 57%, however (*big*) is used less than 43%. In contrast, we can find that when these words are used in contexts, statistics will be quite different. (*Big bang*) is the most frequent collocation (394), however the most frequent collocation with large is (*large tentacle*) (140).

The results obtained in the above table show that more we have recent dictionaries more we get various definitions of the target word with rich information about its lexicology

5. Synonyms from dictionaries based on corpora

According to traditional dictionaries (*strong*) and (*powerful*) are complete synonyms; but if we refer to the dictionaries based on corpora we will find very important information about the 2 synonyms⁸.

Word	Frequencies
<i>Strong</i>	15768
<i>Powerful</i>	7070
Total	22838

Table 5: frequency distribution of (*strong*) and (*powerful*)

The word (*strong*) is used with a frequency of 15768 times in the British National Corpus; however the word (*powerful*) has a frequency of 7070.

By using dictionaries designed from electronic corpora, we can obtain valuable information for the two synonyms.

⁸ From the BNC (byu.edu)

5.1. Collocations of (*strong*)⁹ (Distribution according to the first 100 collocations)

Distribution*	WORD/PHRASE	Frequencies
1	Strong sense	209
2	Strong support	152
4	Strong position	124
6	Strong feeling	113
9	Strong opposition	83
18	Strong influence	57
20	Strong argument	55
86	Strong impact	19
94	Strong women	18
100	Strong body	16

Table 6: Collocations of (*strong*) from the BNC

5.2 Collocations of (*powerful*)¹⁰ (Distribution according to the first 100 collocations)

Distribution *	WORD/PHRASE	Frequencies
1	Powerful influence	72
2	Powerful man	64
5	Powerful body	42
10	Powerful position	31
15	Powerful argument	24
21	Powerful impact	17
33	Powerful sense	15
50	Powerful support	11
51	Powerful feelings	10

⁹. Distribution according to the first 100 collocations of “strong” from the BNC

¹⁰. Distribution according to the first 100 collocations of “powerful” from the BNC

57

Powerful opposition 10

Table 7: Collocations of (powerful) from the BNC

The analyses taken thanks to the corpus show that “*strong sense*” and (*powerful sense*) have not the same balance in lexicography because for the first word (*strong*) it is the most important and the first collocation with 209 times but for the second word (*powerful*) it is the 33rd most important collocation with only 15 times. The same remark could be said to the collocation (*influence*), it is the first most important collocation for the word (*powerful*) with a frequency of 72 times however it is considered as the 18th most important collocation for the word “*strong*” with a frequency of only 57. We can take another example such as the verb (*to learn*) which has 2 different forms in the past simple and past participle forms: Regular and irregular forms (learnt and learned). At The first sight these two words are typically the same but in reality each word is quite different from the other it depends on the use and the context where the word is used. (*Learned VS Learnt*)¹¹

DISTRIBUTION	WORD/PHRASE	TOKENS
<u>1</u>	LEARNED	5230
<u>2</u>	LEARNT	2153

Table 8: frequency distribution of)Learned) and (learnt)

According to the British National Corpus (BNC) statistics learned (the regular form) is preferable than the irregular form (learnt).

¹¹. Statistics from the BNC(byu.edu)

WORD/PHRASE		WORD/PHRASE	
Learned friend	413	Learnt English	4
Learned gentleman	127	Learnt French	4
Learned member	95	Learnt lessons	3
Learned profession	21	Learnt German	3

Table 9: Collocations of learned and learnt (Distribution according to the 4 first collocations)

6. Dictionaries and Register Studies

Dictionaries designed by electronic corpora permit also to make register studies of the target words. Thanks to data collected from native speakers' source and software, it is possible to make a partition of the registers where the word is strongly used and where not. The last example is about two words (few) and (a few); the results obtained from the corpus are invaluable. The first word (few) is more used in fiction (571, 9) followed by news (416, 8) than in spoken (345, 5) and in academic (319, 1); however for the word (a few) is more used in fiction (426) followed by spoken (255, 1) than in news (233, 3) and academic (143, 9)¹².

REGISTERS	Spoken	Fiction	News	Academic
few	345.5	571.9	416.8	319.1
A few	255.1	426.0	233.3	143.9

Table 10: Frequencies of (few) and (a few) from sub-registers of the (BNC. byu.edu)

¹² Statistics from the BNC (byu.edu)

7. Conclusion

Nowadays, almost all the linguistic domains are investigated with the help of corpora or computational linguistics. With dictionaries based on corpora and mainly the electronic dictionaries, it is possible to make very advanced research such as:

- Search by subject, registers
- Search a word according to the parts of speech: Abbreviations, adjectives, adverbs, verbs, auxiliaries, linking verbs, modal verbs, phrasal verbs, combining-forms, conjunctions, definite/indefinite articles, infinitive markers, nouns, numbers, ordinal numbers prefixes, suffixes, prepositions, pronouns...
- Search according to registers: Dialect, formal, informal, humorous, ironic literary, slang, taboo...
- Search according to the subject area: architecture, art, biology, business, computing, economics, linguistics, mathematics, philosophy, politics, psychology, religion, sport, statistics...

Dictionaries designed by electronic corpora present a powerful resource of data for both learners and teachers. Nowadays almost all the official textbooks are designed either by electronic corpora or by dictionaries based on electronic corpora.

References

- BIBER, D. SUSAN, C & RANDI, R (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.
- COOK, G. (2003). Applied Linguistics. Oxford University Press

- FIRTH, J.R. (1957) *Papers in Linguistics: 1934-1951*, London: Oxford University Press.
- HUNSTON, S. (2002) *Corpora in Applied Linguistics*. Cambridge University Press.
- JOHANSON, S. & HOFLAND, K. (1989). *Frequency Analysis of English Vocabulary and Grammar*. Oxford: Clarendon Press.
- SINCLAIR, J.M. (1991). *Corpus Concordance Collocation*. Oxford : OUP.
- SINCLAIR, J.M. (1992). "The Automatic Analysis of Corpora". Starvik (ed.)

Dictionaries

- Cambridge Learner's Dictionary (2001)
- Collins COBUILD English Dictionary (1995).
- The Earlier Collins COBUILD English Language Dictionary (CCELD)
- Longman Dictionary of Contemporary English 2nd Edition, (1987)
- Longman: Active Study Dictionary. 1991 Oxford university press.
- Longman Dictionary of Contemporary English 3rd edition (1995)
- Oxford Advanced Learner's Dictionary (2006)

Web-sites

- <http://www.oxford.com/dictionaries>
- <http://www.cambridge.com/dictionaries>
- www.dictionary.cambridge.org
- [http:// www.dictionary.oed.com](http://www.dictionary.oed.com) (Oxford)
- [http://www. bnc. byu.edu](http://www.bnc.byu.edu)
- [http://www. Jtw.edu](http://www.jtw.edu)