ELECTRONIC CORPORA: AS POWERFUL TOOLS IN COMPUTATIONAL LINGUISTIC ANALYSES.

Mohamed Grazib

Djillali Liabes University: Computer science department Sidi Bel Abbes

E-mail: mfgrazib@hotmail.com

Abstract: Technology has emerged almost all the domains in our daily life. In computational linguistics, the uses of electronic corpora are very important. Nowadays it is possible to study linguistic phenomena by using statistical analyses: Concordances, collocations and frequencies have great influence in making linguistic researches more available, more adequate and more accurate.

Keywords: corpus, computational linguistics, corpus linguistics, Concordances, collocations and frequencies

1 Introduction:

Electronic Corpora are indispensable for computational linguistics; in addition to the availability and the accuracy the tasks can be done in few minutes. Nowadays both the qualitative and quantitative analyses of language are possible by the uses of Electronic corpora and computers. This article is an attempt to show the benefits of corpora in the English applied linguistic studies.

2. What is an electronic 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 or spoken form that can be available on computers as software or via internet. ***G. Cook*** (2003:73) suggests that the word corpus refers to a databank of language which has actually occurred-whether written, spoken or a mixture of the two. The written texts are originally from magazines, books, diaries, newspapers, letters, popular fictions……; however the spoken texts can be any recorded formal or informal conversations: Telephone conversations, dialogues, radio-shows, political meetings…….

3. What is computational linguistics?

The Association for Computational Linguistics defines computational linguistics as the scientific study of language from a computational perspective. Computational linguistics is a discipline between linguistics and computer science .It is a part of the cognitive sciences and it has a strong relation with artificial intelligence.
Computational linguistics originated from the 1950s, where the United States used computers to translate automatically texts from foreign languages into English, particularly Russian scientific journals. Traditionally, computational linguistics was usually performed by computer scientists who had specialized in the application of computers to the processing of a natural language.

4. What is Corpus linguistics?

Corpus linguistics is the study and analysis of data obtained from a corpus. The main task of the corpus linguist is not only to find the data but to analyse it. Computers are useful, and sometimes indispensable, tools used in this process. Corpus linguistics is based on two main software objects: a corpus, which is the body of data to be investigated, and a concordancer, a tool for searching that corpus. Corpus Linguistics is now seen as the study of linguistic phenomena through large collections of machine-readable texts: corpora. **Biber et al** (1998:23) said that: “Corpus linguistics makes it possible to identify the meanings of words by looking at their occurrences in natural contexts, rather than relying on intuitions about how a word is used or on incomplete citation collections”.

5. Size of corpora:

Corpora come in many shapes and sizes, because they are built to serve different purposes. Nowadays 1 million words is fairly small in terms of corpora. We can make a distinction between reference[1] and monitor corpora [2]: The following list shows a very limited sample of corpora’s sizes.

- **Bank of English**: about 400 million words.
- **COBUILD/Birmingham Corpus**: More than 200 million words.
- **Longman Lancaster corpus:** 30 million words.

[1] Reference corpora have a fixed size (e.g., the **British National Corpus**).

[2] Monitor corpora are expandable (e.g., the **Bank of English).**

- **British National Corpus (BNC**):100 million words.

- **American National Corpus (ANC):** 11.5 million words.

- **Brown corpus**: 1million words.

- **Lancaster-Oslo/Bergen (LOB**) corpus: 1 million words.

- **Northern Ireland Transcribed Corpus**: 400,000 words.

- **Corpus of Spoken American English (CSAE**):200,000 words.

What is evident is that the size of any corpus depends mainly on the purposes it was Created for, and that this size can vary from some hundred words to some million words.

6. Concordance, Collocation and Frequency.

In reality, a corpus by itself can do nothing at all; it is nothing other than a store of used language. A corpus does not contain new information about language but the software offers us new perspectives. Most readily available software packages process data from a corpus in three ways: showing, frequency, phraseology and collocations. **G. Cook** (2003:111).

6.1. Concordance:

A concordance is a screen display or printout of a chosen word or phrase in its different contexts, with that word or phrase arranged down the centre of the display along with the text that comes before and after it.
John Sinclair (1991:32) defines a concordance as a collection of the occurrences of a word-form each in its own textual environment. In the same context **S. Hunston** (2002:39) says that it is a programme that searches a corpus for a selected word or phrase and presents every instance of that word or phrase in the centre of the computer screen with the words that come before and after it to the left and right. The selected word appearing in the centre of the screen is known as the “node word”.
The following example illustrates the 10 concordances of the word **computer** from Web Concordancer **LOB.txt.**

1 *etition between the analogue **computer** and the digital computer. To a*

2 *g made on a Ferranti Mercury **Computer** at Meteorological Office, Duns*

3 *unnecessary devices that the **computer** can be made an economic propos*

4 *ouch with manufacturers about **computer** developments of special signi*

5 *he {0PIW} are compiled by the **computer** from data sheets (dictionary*

6 *seen that the problem of the **computer** is in no way related to the p*

7 *racy is required the digital **computer** is the only one to use and ele*

8 he {0ACE} digital electronic ***computer*** *of our laboratory and, further*
9 with the help of the digital ***computer*** *of the University of Toronto.*
10 al purpose electronic digital ***computer*** *to do the job. It is therefore*

6.2. Collocation:

Firth (1957) stated that “you shall know the word by the company it keeps”. The meaning of Firth’s citation here is to classify words not only on the basis of their meanings, but also on the basis of their co- occurrence with other words.
S. Hunston (2002:12) defines collocation as the statistical tendency of words to co-occur.
Collocation investigations can be a preliminary step for other research questions: investigating the distribution of word senses and uses, and comparing the use of seemingly synonymous words, because languages have many words that are similar, and dictionary definitions often characterise such words as identical or synonymous in meaning, however investigating the use and distribution of synonyms in a corpus allows us to determine their contextual preferences associated with other collocates or associated with register differences. **Biber et al** (1998:24).
The following table shows the five most **time**’s collocations with nouns, verbs, and adjectives.

	Nouns		Verbs		Adjectives	
	WORD	# TIMES NEARBY	WORD	# TIMES NEARBY	WORD	# TIMES NEARBY
1	YEARS	1933	WAS	17846	LONG	4850
2	YEAR	1703	IS	12614	GOOD	1587
3	PERIOD	1360	HAD	8128	SHORT	1522
4	PEOPLE	1334	BE	8023	OTHER	1202
5	DAY	1139	WERE	4298	RIGHT	1111

Table 1: collocations of the word **time** with (nouns, verbs, and adjectives) from **view.byu.edu** :

6.3. Frequency:

Frequency list tells us what words and phrases are used most often. **Biber et al** (1998:23) argue that frequency investigations tell us how often different words are used; allowing us to identify particularly common and uncommon words. Based on the evidence of the billion-word Oxford English Corpus, the 100 commonest English words found in writing around the world are as follows:

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*

Table2*: The first 100 frequent English words.*

As seen in the table above many of the most frequently used words are grammatical words (articles, auxiliaries, prepositions….); however the first noun position (time) is the 55th. We can also explore frequencies according to the main word classes: The frequencies of the main word classes in 1 million-word computer corpora of written English are given in the table bellow:

	%		*%*		*%*	
	Total 1 million words in each corpus		***Informative prose sections***		***Imaginative prose (fiction) sections***	
	US	*UK*	*US*	*UK*	*US*	*UK*
NOUNS	*26.80*	*25.2*	*28.50*	*26.9*	*21.77*	*20.0*
VERBS	*18.20*	*17.8*	*17.02*	*16.4*	*21.69*	*21.9*
DETERMINERS	*14.16*	*14.2*	*14.84*	*15.2*	*12.11*	*11.4*
PREPOSITIONS	*12.04*	*12.2*	*12.77*	*13.1*	*9.87*	*9.6*
ADJECTIVES	*7.07*	*7.3*	*7.65*	*7.8*	*5.35*	*5.7*
PRONOUNS	*6.56*	*7.1*	*4.75*	*5.0*	*11.94*	*13.1*
CONJUCTIONS	*5.92*	*5.5*	*5.94*	*5.5*	*5.86*	*5.4*
ADVERBS	*5.23*	*5.5*	*4.73*	*5.0*	*6.72*	*7.2*
OTHERS[1]	*4.02*	*5.2*	*3.80*	*5.2*	*8.49*	*5.8*
TOTAL	*100*	*100*	*100*	*100*	*100*	*100*

Table 3: The frequencies of the main word classes

By making a brief analysis[2], based on the information taken from the table above, we can notice that:

- The nouns are the most frequent used words
- Verbs are more frequent in conversation and in fiction.
- Pronouns are more frequent in spoken English and in fiction that in informative writing.
- Conjunctions have almost the same frequency in both corpora

7. Using electronic corpora in computational linguistics:

Many words have meanings that are similar, and yet the words are not able to be substituted one for the other. Dictionaries, which deal with words separately rather than comparatively, can be of little help, but observing typical usages of near synonyms can clarify differences in meaning. **S. Hunston** (2002:45).

[1] Includes wh- words, foreign words, numerals………

[2] The analyses are from the Brown Corpus of American English **(Francis&Kucera**.1982:547) and the Lancaster-Oslo-Bergen (LOB) Corpus of written British English (**Johansson&Hofland**.1989:15).

The study is about the following synonyms (**Sheer, pure, complete, utter and absolute**). The first analysis is upon a traditional research (using dictionaries)

In this context **Partington** (1998:33-46) gives examples of intensifying adjectives: "**sheer, pure, complete, utter, and absolute**". He points out that dictionaries tend to define those words in similar ways, and even give them as synonymous of each other:

- **The Collins COBUILD English Dictionary** (CCED), suggests that "complete" and "pure" are synonyms of "sheer".
- **The Longman Dictionary of Contemporary English** (LDOCE) gives "pure" as a synonym of "sheer".
- The earlier **Collins COBUILD English Language Dictionary** (CCELD) gives "absolute" as a super ordinate of "sheer".

In spite of this apparent similarity in meanings, the typical collocates of each adjective differ to quite a considerable degree. For example "**sheer**" is used with nouns of degree or magnitude (sheer weight, sheer number) often in the pattern (the sheer noun + of noun); e.g. (the sheer weight of noise).

The other adjectives do not collocate with these nouns. In addition, "**sheer**" alone is often used in expressions indicating causality (though sheer insistence; by sheer hard work; because of sheer hard work; his sheer integrity got him though; his enthusiasm and sheer hard work meant that things moved quickly). **Partington** (1998:36). He ends this analysis by making some statements: "**Complete**", is used with nouns indicating:

- Absence:(complete ban)
- Change:(complete revamping)
- Destruction: (complete collapse)
- Absolute is used with what **Partington** calls "hyperbolic" nouns, such as (chaos, disgrace, genius……). Ibid (1998:43)

7.1. The corpora analyses.

7.1.1. By frequency analyses

The table bellow shows the frequencies of the adjectives (complete, absolute, sheer, and utter):

DISTRIB	WORD/PHRASE	TOKENS REG1	PER MIL IN REG1 [100,000,000 WORDS]
1	COMPLETE	12594	125.94
1	ABSOLUTE	3432	34.32
1	PURE	3305	33,05
1	SHEER	2028	20.28
1	UTTER	652	6.52

Table 4: Frequencies from **(view.byu.edu)**

By using corpora we can see immediately that **complete** is the most used word (12594 times), followed by the adjective **absolute** (3432 times) , in the 3rd position we can find that **pure** is used (3305 times), **sheer** is used (2028 times) ; however **utter** is the last position with only a frequency of (652).

7.1.2. By register analyses

The following five tables show the frequencies of the adjectives concerned by registers:

REGISTER	**SPOKEN**	**FICTION**	**NEWS**	**ACADEMIC**	**NONFIC MISC**	**OTHER MISC**
TOKENS	23	31	30	144	139	217
SIZE (MW)	10.33	16.19	10.64	15.43	16.63	28.39
PER MIL	**2.2**	**1.9**	**2.8**	**9.3**	**8.4**	**7.6**

Table 5: The frequency of “**Complete**” by registers from (**view.byu.edu**)

REGISTER	**SPOKEN**	**FICTION**	**NEWS**	**ACADEMIC**	**NONFIC MISC**	**OTHER MISC**
TOKENS	56	422	218	177	259	585
SIZE (MW)	10.33	16.19	10.64	15.43	16.63	28.39
PER MIL	**5.4**	**26.1**	**20.5**	**11.5**	**15.6**	**20.6**

Table6: The frequency of “**Sheer**” by registers from (**view.byu.edu**)

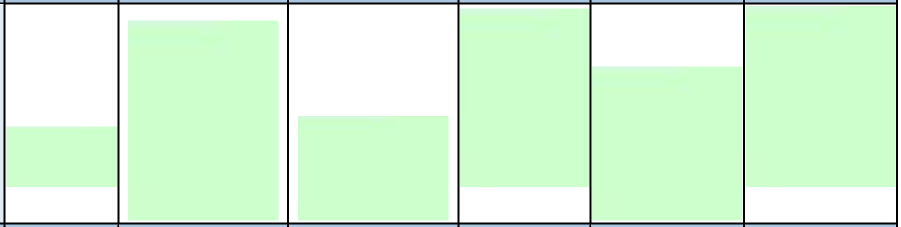
REGISTER	SPOKEN	FICTION	NEWS	ACADEMIC	NONFIC MISC	OTHER MISC
						
TOKENS	87	444	152	380	355	706
SIZE (MW)	10.33	16.19	10.64	15.43	16.63	28.39
PER MIL	**8.4**	**27.4**	**14.3**	**24.6**	**21.3**	**24.9**

Table7: The frequency of “***pure***” by registers from (**view.byu.edu**)

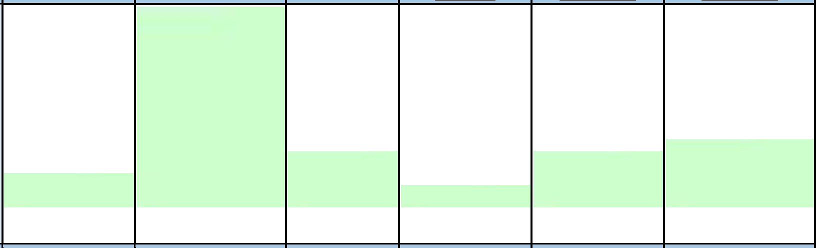
REGISTER	SPOKEN	FICTION	NEWS	ACADEMIC	NONFIC MISC	OTHER MISC
						
TOKENS	22	196	37	22	57	120
SIZE (MW)	10.33	16.19	10.64	15.43	16.63	28.39
PER MIL	**2.1**	**12.1**	**3.5**	**1.4**	**3.4**	**4.2**

Table8: The frequency of “**Utter**” by registers from (**view.byu.edu**) .

REGISTER	SPOKEN	FICTION	NEWS	ACADEMIC	NONFIC MISC	OTHER MISC
	🟩	🟩	🟩	🟩	🟩	🟩
TOKENS	318	370	223	925	676	920
SIZE (MW)	10.33	16.19	10.64	15.43	16.63	28.39
PER MIL	**30.8**	**22.8**	**21.0**	**59.9**	**40.6**	**32.4**

Table9: The frequency of “**absolute**” by registers from (**view.byu.edu**) .

The adjective **absolute** is used mainly in the academic register by a frequency of 59.9 per million words; however the adjective **pure** is also used in the fiction register by a frequency of 27.4 per million words, and in the other registers by a frequency of 24.9 per million words. The adjective **sheer** is used mainly in fiction 422 times which means 26.1 per million words, it is also used in news register by a frequency of 20.5 per million words; but in what concerns the adjective **utter,** it is mainly used in fiction register(only 12.1 per million words); however its use in the other registers is less important. The adjective complete, is less used if compared with the other adjectives, we can distinguish that it reaches only a frequency of 9.3 per million words

7.1.3. By collocation analyses.

The following table shows us the adjectives with their top 20th most frequent collocations.

DISTRIB	WORD/PHRASE	TOKENS REG1
1	PURE WHITE	104
2	COMPLETE NEW	29
3	SHEER HARD	25
4	PURE NEW	18
5	COMPLETE UNIFIED	17
6	SHEER PHYSICAL	17
7	PURE PUBLIC	12
8	ABSOLUTE BEST	11
9	COMPLETE SUPRACONAL	11
10	COMPLETE PHYSICAL	10
11	COMPLETE POLITICAL	9
12	COMPLETE SHORT	9
13	PURE ORAL	9
14	COMPLETE FINANCIAL	8
15	COMPLETE HUMAN	8
16	COMPLETE MENTAL	8
17	PURE ECONOMIC	8
18	ABSOLUTE MINIMUM	7
19	ABSOLUTE MORAL	7
20	COMPLETE MONETARY	7

Table 10: adjectives with the top 20th most frequent collocations from (view.byu.edu) .

In analysing collocation's table we can notice immediately that the words: (**white** 104, **new** 18 and **public** 12) are the most frequent words that collocate with the adjective **pure**; the word (**new** 29) collocates most frequently with **complete.** The first frequent word that collocates with **absolute** is (**best** 11); however the adjective utter does not exist in the top 20 words that collocate with the adjectives listed before.

8. CONCLUSION:

Corpus Linguistics has developed considerably in the last decades due to the great possibilities offered by the natural language processing with computers. The availability of computers and machine-readable texts has made it possible to get data quickly and easily. Linguistic domains are investigated by the use of computers; the results are very amazing if compared with the traditional research methods.

References:

1**. Biber, D., Conrad, S. & Reppen, R.** Corpus Linguistics**. Investigating language structure and use**. 1998. Cambridge: CUP. (1998).
2**. Cook. G. Applied linguistics**: Oxford. OUP. (2003)
3. **Firth, J.R.** 1957. **Papers in Linguistics** 1934-1951. Oxford: OUP(1957).
4.**Francis, W.N & Kucera, H**: **Frequency analysis of English usage** . Boston , MA: Houghton Mifflin. (1982)
5. **Hunston, S. Corpora in applied linguistics. Cambridge**: Cambridge University Press. (2002).
6. **Johannsson S. and Hofland K.**. **Frequency Analysis of English vocabulary and grammar**: based on the LOB corpus. Clarendon Press, Oxford.(1989).
7**. Partington, Alan**. Patterns and meanings: **Using corpora for English language research and teaching**. Amsterdam: John Benjamins PublishingCompany(1998).
8. **Sinclair, John McH.**. **Corpus, concordance, collocation**. Oxford: Oxford University Press. (1991).
9.**view.byu.edu** Mark Davies, Professor of Corpus Linguistics at Brigham Young University.
10.Web Concordancer **LOB.txt.**