

# Electronic Corpora as an ICTE tool

## The case of 1st year LMD (English)

*Mohamed Grazib*

*English teacher*

*Faculty of literature, languages, social and human sciences*

*Dr Moulay Tahar University*

*E-mail : mfgrazib@hotmail.com*

**Abstract:** the ICTE (Information and Communication Technology for Education) is one of the important modules integrated in almost all the specialities of the LMD system. It is why the integration of electronic corpora as an ICTE tool for the 1<sup>st</sup> year LMD (English) mainly at Moulay Tahar University as a first experience in Algeria is of a great importance. The uses of electronic corpora as the British National Corpus (BNC) which contains 100 million words for students learning English as a foreign language will enhance their proficiency and their competencies to develop their skills in computational linguistics by using statistical analyses through concordances, collocations and frequencies which are available via the net.

**Keywords:** ICTE (Information and Communication Technology for Education) **corpus**,

**BNC** (British National Corpus) **computational linguistics**,  
**corpus linguistics**,  
**frequency, collocations and concordances.**

### **1 Introduction:**

Nowadays Electronic Corpora are available via internet or on Compact Disks in the form of software. Electronic corpora are indispensable in computational linguistics. Corpora and computers permit qualitative and quantitative analysis of language. This article is an attempt to show the benefits of corpora in linguistic studies and how they can be useful for the students learning English (LMD system) in the ICTE module.

### **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 or spoken form or a mixture of the two that can be available on computers as software or via internet.. 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. 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.

### 4. What is Corpus linguistics?

Corpus linguistics is the study and analysis of data obtained from a corpus. Computers are useful and 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 **concordance software**, a tool for searching that corpus.

### 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. 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.
- **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.
- **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 millions of words to some hundred words.

## 6. Concordance, collocation and frequency.

An electronic corpus by itself such as the BNC can do nothing at all; it is just a store of used language. A corpus does not contain new information about language but the software offers the possibility to investigate and analyse data obtained from the corpus in three ways: showing, **concordances**, **frequencies**, and **collocations**.

### 6.1. Concordance:

A concordance is a screen display of a target word or phrase arranged in 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 8 concordances of the word "**beautiful**" from Web BNC concordance software.

- 1- There are **beautiful** waving weeds in the water and everything is blue and green.
- 2- The birds, bright and **beautiful**, dipped their wings exultantly in the lucid flow of air.
- 3- It's the most **beautiful** thing that ever happened to me.
- 4- The scene is still **beautiful** but in a calmer and less vibrant way.
- 5- She had three years to spend in a **beautiful** city, reading; and that was what she was supposed to be doing.
- 6- Hope then built a **beautiful** shrine at the other end of the village.
- 7- She has a **beautiful** voice, beautiful.
- 8- Aren't these **beautiful**?

### 6.2. Collocation:

**Firth** (1957) stated: "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.

The following table shows the 10 most **time**'s collocations with nouns, verbs, and adjectives.

| NOUNS | # TIMES<br>NEARBY | VERBS | # TIMES<br>NEARBY | ADJECTIVES | # TIMES<br>NEARBY |
|-------|-------------------|-------|-------------------|------------|-------------------|
|-------|-------------------|-------|-------------------|------------|-------------------|

|               |      |             |       |                  |      |
|---------------|------|-------------|-------|------------------|------|
| <u>YEARS</u>  | 1933 | <u>WAS</u>  | 17846 | <u>LONG</u>      | 4850 |
| <u>YEAR</u>   | 1703 | <u>IS</u>   | 12614 | <u>GOOD</u>      | 1587 |
| <u>PERIOD</u> | 1360 | <u>HAD</u>  | 8128  | <u>SHORT</u>     | 1522 |
| <u>PEOPLE</u> | 1334 | <u>BE</u>   | 8023  | <u>OTHER</u>     | 1202 |
| <u>DAY</u>    | 1139 | <u>WERE</u> | 4298  | <u>RIGHT</u>     | 1111 |
| <u>PLACE</u>  | 984  | <u>BEEN</u> | 4203  | <u>NEW</u>       | 973  |
| <u>WORK</u>   | 936  | <u>ARE</u>  | 3915  | <u>FULL</u>      | 897  |
| <u>LOT</u>    | 895  | <u>HAS</u>  | 3908  | <u>GREAT</u>     | 733  |
| <u>LIFE</u>   | 828  | <u>WILL</u> | 3592  | <u>DIFFERENT</u> | 635  |
| <u>MONEY</u>  | 819  | <u>HAVE</u> | 3584  | <u>AVAILABLE</u> | 583  |

**Table 1:** The word “time” with the 10 most frequent (nouns / verbs /adjectives) from theBNC.BYU.EDU

#### 6.2.1. Interpretations:

According to the statistics shown by the table (1), the results taken from the BNC web affirm that the word “time” collocates mainly with:

##### \*NOUNS:

- **Year(s):** with a frequency of 1933 and 1703 times
- **Period:** with a frequency of 1360 times
- **People:** with a frequency of 1334 times
- **Day:** with a frequency of 1139 times

##### \*VERBS:

The word “time” collocates mainly with the auxiliary to “be” in past tenses rather than the present tenses.

- **Was:** with a frequency of 17846 times
- **Is :**with a frequency of 12614 times
- **Be:** with a frequency of 8023 times
- **Were:** with a frequency of 4298 times
- **Been:** with a frequency of 4203 times
- **Are:** with a frequency of 3915 times

##### \*ADJECTIVES:

The word “time” collocates mainly with short adjectives rather than long adjectives:

- **Long:** with a frequency of 4850 times
- **Good:** with a frequency of 1587 times
- **Short:** with a frequency of 1522 times
- **Different:** with a frequency of 635 times

### 6.3. Frequency:

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. The following list shows the 100 commonest English words found in writing around the world (taken from the Oxford English Corpus):

|    |      |    |       |    |        |     |         |
|----|------|----|-------|----|--------|-----|---------|
| 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.

Frequencies can be shown in a general form as shown in the table 2, or according to the main word classes as shown in table3.

|       | % |    | %           |  | %           |
|-------|---|----|-------------|--|-------------|
| Total |   | 1  | Informative |  | Imaginative |
|       |   | 31 |             |  |             |

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

**Table 3:** The frequencies of the main word classes

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).

#### **7- The corpus methodology:**

The most suitable methodology concerning corpora and the ELT practice is that of the **task based approach** which can be divided into three steps:

The **first step**, teachers have to teach students about corpora:

- What is a corpus?
- How it is formed?
- What is its content?
- And some things about corpus linguistics: **frequencies, collocations, concordances....**

The **second step** is to teach and to train students:

- How to use electronic corpora,
- How to use the concordance software.
- How to extract: concordances, collocations and frequencies.

The **third step** is to exploit all what is found with the help of the corpus, this consists in the interpretations of the results found and to use them in the teaching process.

#### **8. CONCLUSION:**

Integrating Electronic Corpora in the ICTE module for the LMD system (English) is without any doubt very fruitful. In one hand,

it gives the opportunity to use the technological tools: Computers, internet, corpora software such as the BNC via the web, and making linguistic analyses for the foreign language they are learning (English). The experience has seen its first steps with the students of the first year LMD English in the ICTE module, but some obstacles still exist such as the lack of computers and internet connections.

#### References:

1. **Biber, D., Conrad, S. & Reppen, R.** . *Corpus Linguistics. Investigating language structure and use.* 1998. Cambridge: CUP. (1998).
2. **Firth, J.R.** 1957. *Papers in Linguistics 1934-1951.* Oxford: OUP(1957).
3. **Francis, W.N & Kucera, H:** Frequency analysis of English usage . Boston , MA: Houghton Mifflin. (1982)
4. **Cook, G.** Applied linguistics: Oxford . OUP. (2003)
5. **Halliday M.A.K.** System and function in language (edited by G. Kress). Oxford: OUP. (1976).
6. **Hunston, S.** *Corpora in applied linguistics.* Cambridge: Cambridge University Press. (2002).
7. **Johannsson S. and Hofland K..** Frequency Analysis of English vocabulary and grammar: based on the *LOB corpus*. Clarendon Press, Oxford.( 1989).
8. **Partington, Alan.** *Patterns and meanings: Using corpora for English language research and teaching.* Amsterdam: John Benjamins Publishing Company(1998).
9. **Sinclair, John McH..** Corpus, concordance, collocation. Oxford University Press. (1991).
10. [view.byu.edu](http://view.byu.edu)
11. Web **BNC** concordancer.