

Automatic Text Dissection by Corpora

GRAZIB Mohamed & HADJI Mohamed
Université de Saïda

Abstract

The paper focuses on the methods used to explore ordinary texts that will be dissected into atomistic parts by using powerful technological tools. The study will be based mainly on electronic corpora which are of paramount importance in computational and corpus linguistics. “Automatic Text Dissection by Corpora” attempts to explore deeply texts that could be analysed scientifically via charts, graphs and statistics. It shed light on how this invaluable data could be used by specialists and linguists in order to make the right interpretations and formulate the right critics about the real image of texts processed by advanced software.

1. Introduction

The study aims to analyse a text by using electronic corpora such as the British National Corpus (BNC), and the Corpus of the Contemporary American English (COCA) and other advanced tools that will be mentioned below.

The chosen text is entitled “*Fast-Food*”, it is a first year middle school text taken from “*Spotlight on English*”¹ the official textbook of the ministry of the national education. The technological tools used in this investigation are mainly electronic corpora mentioned above and other advanced software tools. The investigation will be made mainly on general and range frequencies, parts of speech partitions, genre

¹ F. Bouhadiba et al (2004). *Spotlight on English*

investigations, diachronic explorations and comparative studies.

As a sample, we have taken a short text from the first year middle school textbook, it contains approximately 54 words². The selected text is mainly about food, but it remains as a very important example in what concerns food's target vocabulary³: *food, fast-food, restaurant, hamburgers, chips, pizza, hot-dogs, sandwiches and tacos*.

2. The target text⁴

Young people like fast- food. What is fast -food? It's a small restaurant where you can have hamburgers, chips, pizzas, hot-dogs, sandwiches, tacos, and drinks. You can eat your food here or take it away. Fast- food doesn't cost much and doesn't take much time; to prepare and to eat but is it healthy?

3. Range Study

Once we input the text in the space provided for this purpose, we'll immediately obtain the following results.

² "Fast-food" is considered as one word

³ Vocabulary investigation made by the online Collins Concise English Dictionary HarperCollins Publishers

⁴ A short text about fast-food (2004:112)

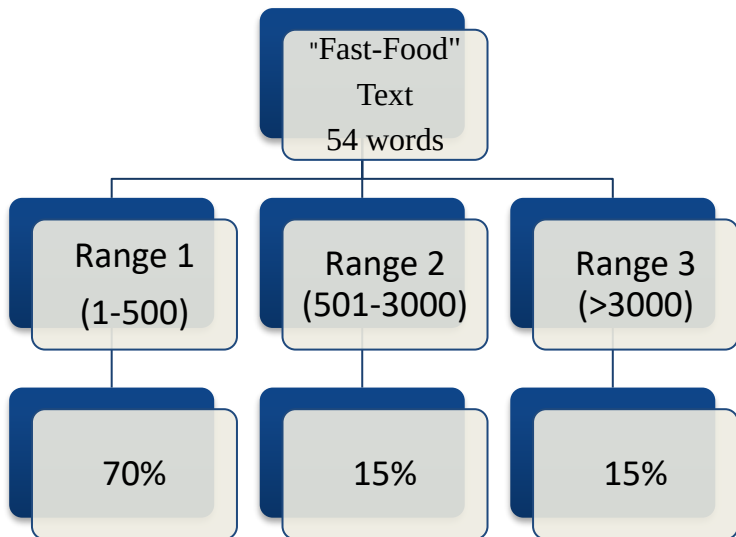


Figure 1: COCA frequencies of “Fast-Food” text

As seen in figure 1, the dissection of “Fast-Food” text leads to three main ranges:

- Range 1: from 1 to 500 first frequent words used in the English language.
- Range 2: from 501 to 3000 first frequent words used in the English language.
- Range 3: > 3000 first frequent words used in the English language.

Among the options offered by the corpus software, the user can explore each range alone; by a simple click, we can obtain a large list of detailed frequencies about each category:

Range 1:

Frequencies	Words
03	And, it
02	Can, do, is, much, , take, to, , you,
01	A, away, but, cost, food, have, here, like, or, people, small, time, where, young, your

Table 1: Range 1 list

Range 2:

Frequencies	Words
02	eat
01	Hot-dogs, drinks, healthy, prepare, restaurant

Table 2: Range 2 list

Range3:

Frequencies	Words
03	Fast-food
01	Chips, Hamburger, Pizza, Sandwich, taco

Table 3: Range 3 list

One can notice that the text's partition is logically acceptable, simply because more than the 2/3 of the total number of words used in this text are among the first 500 most frequent ones in the English language (70%); however the two other ranges present the 1/3 of the text with 15% for each range. Linguists and specialists

recommend that only 3000 words⁵ are necessary to use the English language adequately.

4. General Distinctive Parts of the Text

After obtaining the most important characteristics of the text, we can have a glance about the distinctive parts of the target text in what concerns the parts of speech categories and their respective frequencies.

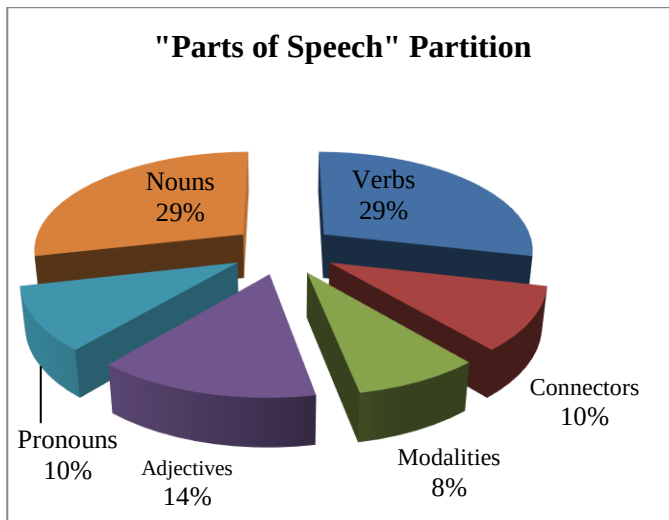


Figure 2: "Parts of Speech" Partition

This option permits to make deeper dissection where the main parts of speech and their respective frequencies are made available in order to analyse the target text. The previous results obtained in the figure above lead to make a partition that makes us able to have an immediate idea about the text's composition mainly in what concerns the parts of speech. In this case, our text responds logically to the

⁵ 3000 necessary words : Oxford 3000 TM

universal rules of the BNC parts of speech proportions where, the nouns are dominants, followed closely by verbs and where adjectives are in the third position.

5. Detailed Distinctive Parts of Text

The same option permits to explore deeply each part alone. In order to analyse correctly a text, we need very detailed and precise statistics. The following example offers large opportunities to specialists and non specialists to make very advanced investigation in this domain. The following investigation is just a replica to the previous one (parts of speech investigation); with more implicit characteristics and details.

Distinctive Parts of the Text	Detailed frequencies	General Frequencies
Verbs	- Static: 57% (8) - Active: 43% (6)	14
Nouns	- Countable and uncountable	14
Connectors	- Opposition: 20% (1) - Addition: 60% (3) - Disjunction: 20% (1)	05
Modalities	- Negation: 50% (2) - Place: 50% (2)	04
Adjectives	- Objective: 57% (4) - Subjective: 43% (3)	07
Pronouns	- You : 29% (2) - It: 43% (3) - What: 14% (1) - Your: 14% (1)	07

Table 4: Detailed Distinctive Frequencies

In this short passage, the words related to “food” present the most important part in the analysed text. This consists of 11 words: Fast-food (3times), restaurant, hamburgers, chips, pizza, hot-dog, sandwiches, tacos, drinks and food (1 time for each word). As a matter of fact, and in the light of the obtained results by the used software, the word “fast-food” appears as the most frequent word in this short text with 3 apparitions. It is why; we have selected this item as having the status of the most frequent word in this text. By the way the selected word was used to make many parallel investigations with other words: Fast-food Vs restaurant, their genre study, as well as their usage in different periods of time.

6. Genre Study

6.1 “Fast-food” in all Genres

This investigation permits to explore deeply any ordinary word and obtain invaluable statistics that will be analysed and interpreted. In this case and as mentioned above, we have studied the word “fast-food” in the main genres: (spoken, fiction, magazine, newspaper and academic).

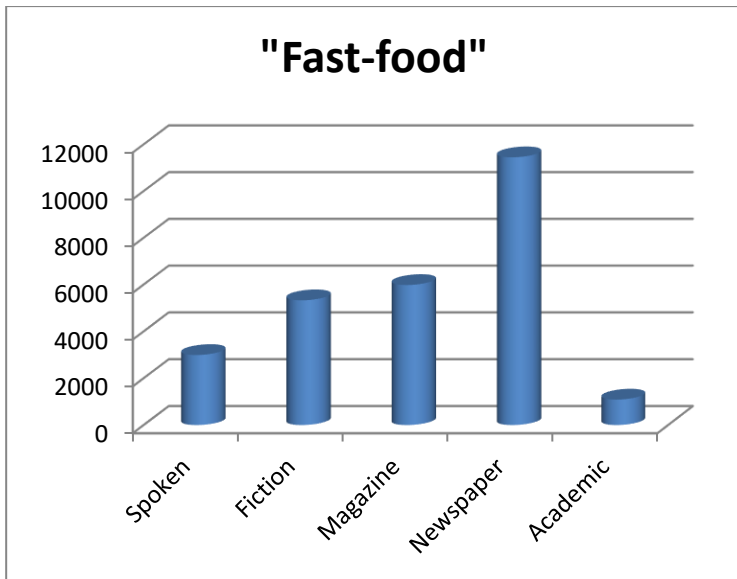


Figure 3: “Fast-food” in the Main Genres

As seen in the above figure, the word “fast-food” remains a newspaper and magazine’s genre word rather than being a spoken word. It’s probably due to the advertisement influence and the written papers published about this category of restaurants that invaded all societies in the world.

6.2 “Restaurant” in all Genres

As far as “fast-food” investigation is concerned, we’ll make the same study with an almost equivalent word which is “restaurant” that is strongly challenged by “fast-food” in these modern times.



Figure 4: “Restaurant” in the Main Genres

The scrutiny leads that not only the word “restaurant” is mainly used in newspaper and magazine genres, but it is also used with excellence in fiction. As a consequence the word “restaurant” remains strongly a romantic and emotional place in all the fiction works rather than an ordinary “fast-food”.

The figure below shows clearly that the word “restaurant” is strongly used in all genres if compared with its equivalent word “fast-food”. Not only the “restaurant” is a newspaper, fiction and magazine genre word by excellence but it is also considered as an academic word.

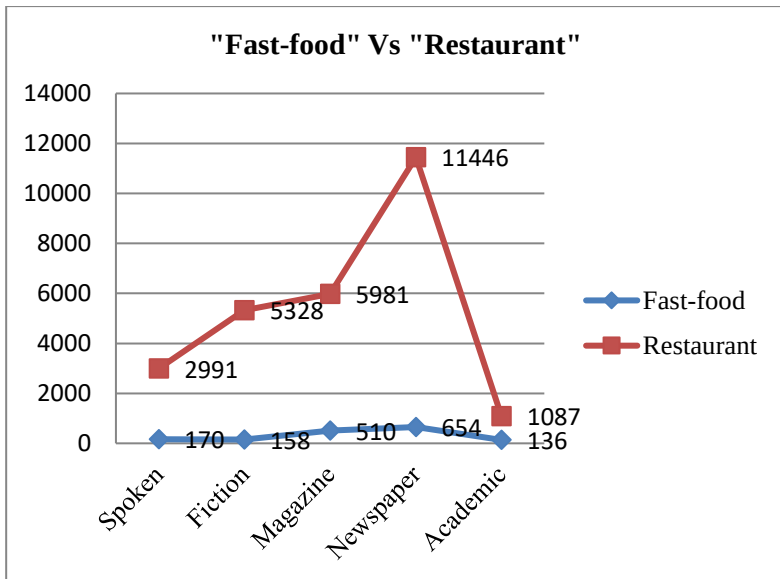


Figure 5: Comparative Study

7. Diachronic Study

7.1 “Fast-food” Diachronic Study

In this sub-section, we’ll study the strong and weak moments of “fast-food” during different decades in the newspaper genre.

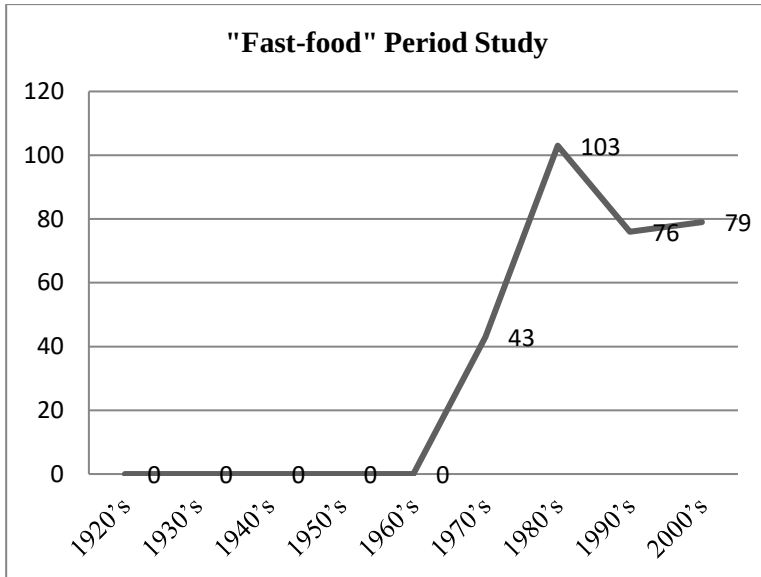


Figure 6: "Fast-food" Period Study

As seen from the figure above, "fast-food" is typically a new word. It appeared in the 1970's and knew its strong moments in the 1980's, but it decreases during the two last decades.

7.2 "Restaurant" Diachronic Study

In this sub-section, we'll study the strong and weak moments of "restaurant" during different decades in the newspaper genre.

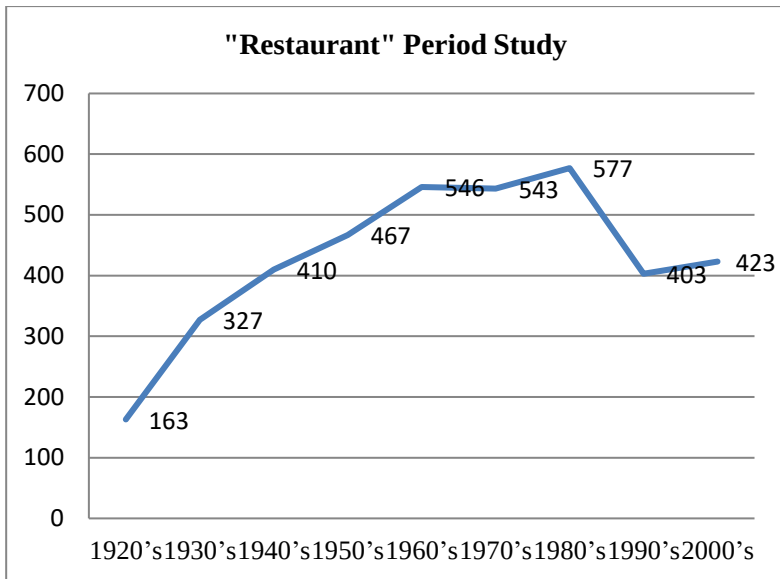


Figure 7: "Restaurant" Period Study

The period investigation led to important results that could interest linguists, sociologists, etymologists... Even when the word "restaurant" is a very old word that was used during centuries; but it hasn't lost its social, linguistic and romantic rank in many languages. We have restricted the study to the last century and the beginning of the present one. The results show that the word "restaurant" remains almost the same in what concerns the usage frequency in the newspaper genre. It was not affected at all by the strange and enormous events known in the modern era. So, even the crash of Wall-Street, the Second World War and the Cold War had not affected at all the linguistic, social and the romantic sides of the word "restaurant".

Conclusion

Using corpora to dissect texts for linguistic purposes is an experience that could be very helpful to very interested persons (linguists, textbook designers, teachers, students...) it will offer a lot of issues to dissect texts, paragraphs, sentences, and even words to detect the least information about them. So, corpora are and remain as very important resources to linguists and researchers, who can easily collect and then select words, expressions, and sentences from authentic language.

Statistics such as general and detailed lists of the distinctive parts of the analysed text(s), and frequencies about each related word category, are really very precious and invaluable information about the raw text(s) that was in the form of abstract paragraphs, sentences and words which are transformed and presented in statistics, charts, tables and graphs containing a lot of interpretations according to the objectives and the purpose of the study.

The uses of advanced technology such as electronic corpora and computer software will let their finger prints in computational and corpus linguistics.

What we have got as results are only drops from an ocean, but what remains will be without any doubt of paramount importance.

References

- BURNARD, L. (2000). *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Service.
- HUSTON, S. (2002) *Corpora in Applied Linguistics*. Cambridge. University Press.
- JOHANSSON, S. & HOFLAND, K. (1989) *Frequency Analysis of English Vocabulary and Grammar*. Clarendon Press.
- SINCLAIR, JM. (1992). “*The Automatic Analysis of Corpora*”. Starvik (ed.)
- BOUHADIBA, F et al. (2004). *Spotlight on English*. 1st Year Middle School Textbook.
- Text: “*Fast-food*”. From: *Spotlight on English* (2004:112); 1st Year Middle School Textbook.
- British National Corpus (BNC Online Version)
- Corpus of Contemporary American English (COCA Online Version)
- Mark Davies. BNC BYU Brigham Young University.
- Lexical Tutor. V.6.2
- Oxford 3000 TM. Topes V.8.1
- Wordnetweb.Princeton.edu
- The Online Collins Concise English Dictionary. Harper Collins Publishers.