Democratic and Popular Republic of Algeria
Ministry of Higher Education and Scientific Research

Djillali Bounaama University, Khemis Miliana
Faculty of Letters and Foreign Languages
Department of Foreign Languages

THE WORD "ALGERIA"AS USED IN ENGLISH

CORPUS ANALYSIS

Dissertation Submitted to the Department of Foreign Languages in Candidacy for the LMD Master in English Language and Communication

Candidate:

Rim Benaichouba

Ouissam Mesbaiah

Manel Boucherit

Supervisor:

Mr. Cherifi Hamza

2021/2022

Declaration

I hereby declare that the substance of this dissertation is the result of my investigation due reference of acknowledgment is made when necessary to the whole of other researchers.

Rim Benaichouba

Ouissam Mesbaiah

Manel Boucherit

DEDICATION I

This thesis dedicated to:

*The soul of my father whose presence colored my life with joy, and whose absence left nothing but longing. I love you Daddy and I will owe everything to you for as long as I live.

*My mother who has always been the source of my inspiration and encouragement.

*My second father who raised me as one of his daughters, and my aunt.

*My husband Imad who has been a source of strength, support and motivation for me all the time

*My adorable sisters Imane, Ikram, Fella and Moufida.

*My brother Mohammed and his wife Youssera.

* My only and lovely niece Anais.

*My amazing friend Aya.

Rim Benaichouba

DEDICATION II

I cannot express enough thanks to our supervisor Mr. Hamza Cherifi for his support and guidance throughout the making of this work. I offer my sincere appreciation for the learning opportunity provided by him.

My completion of this project could not have been accomplished without the support of the pillars of my life, my loving mom and dad, my two brothers and my best friend and

The love of my life Anas.

Finally to my cousin Mouni and my friend Sanaa, I shared with you all my sentiments of frustration and you were always there to provide me with comfort and reassurance.

My heartfelt thanks.

Ouissam Mesbaiah

DEDICATION III

I dedicate this thesis to all my beloved family,

My wonderful friends,

My dear teachers,

My colleagues,

And to everyone who helped me finish this work

Manel Boucherit

ACKNOWLEDGEMENT

We would like to offer our gratitude for our esteemed supervisor Mr. Hamza Cherifi for his guidance and priceless advice. Also our sincere appreciations for teachers of the Department of Foreign Languages of Djilali Bounaama University.

Abstract

Corpuslinguisticsuseslargecollectionsofbothspokenandwrittennaturalwordsandtextsthat are stored on computers. By using a variety of computer-based tools by web data, corpus linguists can explore different questions about language use. Learners provide an extremely powerful tool for the analysis of natural language and can provide tremendous insights as to how language use varies in different situations, such as spoken and written issues represented byelectronicdatabases:theBritishNationalCorpus(BNC),theMichiganCorpusofAcademic Spoken English (MICASE) and the Web Corpus. This study investigates the word Algeria as it is presented in the corpus .The findings and discussion are the primary focus and The obtained results will be evaluated and discussed in order to make an appropriate assessment According to data-analyses, the word Algeria is rarely mentioned in English particularly in academic contexts for a variety of reasons; the lack of the academic and scientific research Additionally, the French colonialism which played a significant role in Algeria's growth. Finally, the geographical location of Algeria regionally, continentally and globally that facilitate the economic commercial andmovements.

Keywords: Algeria, BNC, corpus, linguistics, MICASE, Web corpus.

Résumé

La linguistique de corpus utilise une grande collection de mots et de textes naturels orale et écrits qui sont stockés sur des ordinateurs. En utilisant une variété d'outils informatiques par desdonnéesWeb,les linguistes de corpus peuven texplorer différentes questions sur l'utilisation de la langue. Les apprenants fournissent un outil extrêmement puissant pour l'analyse du langage naturel et peuvent fournir des informations considérables sur la façon dont l'utilisation de la langue varie dans différentes situations, telles que les problèmes orale etécritsreprésentés par des bases de données électroniques : BNC, MICASE et le Web Corpus. Cette étude porte surlemotAlgérietelqu'ilestprésentédanslecorpusLesrésultatsetladiscussionsontl'objectif principal, ils obtenus seront évalués afin de faire une évaluation appropriée. Selon les analyses desdonnées,lemotAlgérieestrarementmentionnéenAnglais, enparticulierdanslescontextes académiques pour diverses raisons ; le manque de recherche académique et scientifique De plus,lecolonialismefrançaisquiajouéunrôleimportantdanslacroissancedel'Algérie.Enfin, lasituationgéographiquedel'Algérieauniveaurégional,continentaletmondialquifaciliteles échanges économiques et commerciale.

Mots Clé : BNC, Corpus, L'Algérie, Linguistique, MICASE, Web Corpus.

ملخص

تستخدم لغويات المدونة مجموعات كبيرة من الكلمات والنصوص الطبيعية المنطوقة والمكتوبة المخزنة على أجهزةالكمبيوتر. باستخداممجموعة متنوعة من الأدوات المستندة إليها حسب بيانات الويب ، يمكن لخبراء اللغة استكشاف أسئلةمختلفة حول الاستخدام اللغوي.حيث توفر أداة قوية للغاية لتحليل اللغة الطبيعية ويمكنهم تقديم رؤى هائلة حول كيفيةاختلاف استخدام اللغة المنطوقة والمكتوبة التي تمثلها قواعد البيانات الإلكترونية BNC, MICASE :ومدونة الويب .دراسة كلمة الجزائر كما وردت في المدونة ادت الي تقييم النتائج التي تم الحصول عليها ومناقشتها وفقًا لتحليلات البيانات ،نادرًا ما تم ذكر كلمة الجزائر في اللغة الإنجليزية بشكل عملي في السياقات الأكاديمية لعدة أسباب منها نقص البحثالأكاديمي والعلمي بالإضافة إلى الاستعمار الفرنسي للجزائر. وأخيراً الموقع الجغرافي للجزائرإلاقليمي والقاري والعالميمما يسهل الحركة الاقتصادية والتجارية.

الكلمات المفتاحية: مدونة, الويب, الجزائر, مدونة الويب, MICASE, BNC.

List of Abbreviations

CL: Corpus Linguistic

BNC: British NationalCorpus

MICASE: Michigan Corpus of Academic Spoken English

OUP: Oxford UniversityPress

BE: British English

FL: ForeignLanguages

ELI: English Language Institute

UM: University of Michigan

ICE: International Corpus of English

ICAME: International computer Archive of Modern English

SEU: The Survey of English Usage

UCL: University College London

COBUILD: Collins Birmingham University

ANC: American National Corpus

ICLE: International Corpus of Learners of English

LOB: Lancaster-Oslo/Bergen

CSAE: Corpus of Spoken American English

List of Tables

List of Figures

.

Table of Contents

Chapter one, Introduction

1. General Introduction

Corpus linguistics is the study of a language as it is expressed in its text corpus, or body of real world text. Corpus linguistics proposes that a reliable analysis of a language is more feasible with corpora collected in the field—the natural context of that language—with minimal experimental interference. The text-corpus method derives the set of abstract rules that govern any natural language from a corpus of texts written in that language. These findings can be used to investigate the connections between the subject language and other languages that have undergone a similar analysis. The first such corpora were derived manually from source texts, but that work is now automated. Corpora have been used not only for linguistics research, but also to compile dictionaries and grammar guides, such as Comprehensive Grammar of the English Language, which was published in 1985. Experts in the field have differing perspectives on corpus annotation. These perspectives range from John McHardy Sinclair, who advocates for minimal annotation so that texts can speak for themselves, to the Survey of English Usage team, who believes that annotation allows for greater linguistic understanding through rigorous recording. Some of the earliest attempts at grammatical description relied, at least in part, on corpora of religious or cultural significance. Prtikhya literature, for example, described Sanskrit sound patterns as found in the Vedas, and grammar of classical Sanskrit was based, at least in part, on analysis of that same corpus. Similarly, early Arabic grammarians paid special attention to the Quranic language Concordances were created in the Western European tradition to allow for detailed study of the language of the Bible and other canonical texts. Corpora in English the publication of Computational Analysis of Present-Day American English in 1967 was a watershed moment in modern corpus linguistics. The work, written by Henry Kuera and W. Nelson Francis, was

based on an analysis of the Brown Corpus, a contemporary compilation of approximately a million American English words carefully selected from a wide range of sources. Kučera and Francis ran various computational analyses on the Brown Corpus before combining elements of linguistics, language teaching, psychology, statistics, and sociology to create a rich and varied opus. Randolph Quirk Towards a description of English Usage published in 1960, was another important publication that introduced the Survey of English Usage. Soon after, Boston publisher Houghton-Mifflin approached Kuera about providing a million-word, three-line citation base for its new American Heritage Dictionary, the first dictionary to be compiled using corpus linguistics. The AHD took the novel step of combining prescriptive and descriptive elements. Other publishers quickly followed suit. The Bank of English was used to compile the British publisher COBUILD monolingual learner dictionary, which was designed for users learning English as a foreign language. The Survey of English Usage Corpus was used in the development of one of the most important Corpus-based Grammars, A Comprehensive Grammar of the English Language, written by Quirk and published in 1985.

The Brown Corpus spawned a number of similarly structured corpora, including the LOB Corpus, Kolhapur, Wellington, Australian Corpus of English, Frown Corpus, and FLOB Corpus. Other corpora include the International Corpus of English and the British National Corpus, a 100 million word collection of various spoken and written texts created in the 1990s by a consortium of publishers, universities, and the British Library. Although work on the American National Corpus for modern American English has paused, the 400+ million word Corpus of Contemporary American English is now available through a web interface. The Montreal French Project created the first computerized corpus of transcribed spoken language in 1971, containing one million words, which paved the way to create a far bigger corpus of spoken French in the Ottawa-Hull area. Corpora Multilingual Many of the early breakthroughs in statistical methods in natural-language programming happened in the field

of machine translation in the 1990s, thanks in large part to work at IBM Research.

There are corpora for languages other than English. The National Institute of Japanese Language and Linguistics in Japan, for example, has developed a variety of corpora of spoken and written Japanese. Corpus of ancient languages Aside from corpora of current languages, computerized corpora of collections of texts in ancient languages have also been created. The Andersen-Forbes database of the Hebrew Bible, which has been in development since the 1970s, parses each phrase using graphs reflecting up to seven levels of syntax and tags each segment with seven fields of data. The Quranic Arabic Corpus is an annotated corpus for the Quran Classical Arabic. This is a new project with various levels of annotation, including as morphological segmentation, part-of-speech labeling, and dependency grammar syntactic analysis. Corpora from a variety of fields Apart from pure linguistic inquiry, corpus linguistics had begun to be applied to other academic and professional disciplines, such as the burgeoning sub-discipline of Law and Corpus Linguistics, which aims to analyze legal texts using corpus data and methods. Methods Corpus linguistics has spawned a slew of research techniques aimed at tracing a line from data to theory. The perspective (annotation, abstraction, and analysis) was first introduced by Wallis and Nelson.

The application of a scheme to texts is known as annotation. Structured markup, part-of-speech tagging, parsing, and a variety of other forms are examples of annotations. The conversion of terms in the scheme to terms in a theoretically motivated model or dataset is known as abstraction .During this research, we will try to shed light on the development of corpus linguistics , present its significance in the field of linguistics and try to implement it as a tool to track the word Algeria , using a computer-based empirical analysis of language by employing large , electronically available collections of naturally occurring spoken and written texts, so-called corpora.

The present dissertation is composed of three chapters: The first chapter is basically a

theoretical one ; it consists mainly of general information about corpus linguistics and an introduction into the main event leading to the emergence of this field , it also reveals some crucial approaches to the domain of corpus linguistics and their significant impact .The second chapter practically investigates linguistically three types of corpora which are the British National Corpus, the Michigan corpus of academic spoken discourse and the web corpus. It focuses on representing these electronic databases and describing how language is used in different context furthermore these websites offer unparalleled insight into variation in English. The third chapter is primarily concerned with the investigation of the word Algeria as it appears in the corpus. The emphasis will be primarily on the findings, and discussions. The obtained results will be analyzed and discussed so that an appropriate evaluation can be made.

2. Objective of theStudy

This study is based on a semantic analysis and a corpus study of the word "Algeria." There are two research goals:Definition and History of CorpusLinguisticsinvestigation of the word Algeria as it appears incorpus to address the first research goal we focus on the development of corpus linguistics in retrospect and the main events that led to the development of this field; additionally, wereveal some critical approaches to the domain of corpus linguistics and their significant impact. To address the second research objective, we traced the word Algeria using corpus evidence from the analysis. The outcomes, and comments will be the mainfocus.

3. Dissertation Structure

This dissertation is divided into sixsections: an introduction, four chapters and a general conclusion. The introduction presents the main research questions about corpus linguisticsand the aim of the study. The first chapter or the literature review contains collected information from other studies about the main points of the research. The second chapter represents the methodologyusedtocollectandanalyzethedata (thetools we used to gather the data). The next chapter shows the results and the findings of the meaning and the use of the word Algeria as

it appear in English. Then the discussion chapter where the findings are discussed with giving recommendations. Then the conclusion chapter where the general summary of the research findings and results.

Chapter two, The Literature Review

1. Introduction

The first chapter deals mainly with the theoretical parts of corpus linguistics in general. The first section contains various expert definitions of corpus linguistics, and the focus then shifts to revealing the distinctive qualities of a corpus. The second part will be devoted to surveying the very earliest work (before approximately 1955) that prognosticates characteristics of modern corpus linguistics. Then it will be dissected into three sections: early antecedents, Chomsky's opposition, and the emergence of English corpus linguistics in the 1960s and 1970s. After that, this chapter will illustrate the neo-Firthian Approach to corpus linguistics. Then the chapter at hand will reveal the impact of corpora: grammatical description while focusing upon "Randolph Quirk," the founder of the Survey of English Usage (SEU). After that, another aspect will be highlighted in this chapter, which will be Corpora beyond Corpus Linguistics. In the same context, this chapter will demonstrate the size of a number of popular corpora. After that, it will reveal limitations that are linked to corpus as an investigation method. In the end, it will be summarized in a general summary.

2. An Introduction into Corpus Linguistics

Students rarely investigate the nature of language beyond the classical definitions and the general knowledge of placement of words, While it comes from the fact that a corpus consists on thousands, if not millions, of written or spoken text samples taken directly from: maximally representative= instances of language in use. A corpus views these "naturally occurring language" as a credible source for a more intensive investigation on the nature of words in terms of implications and interpretations in order to derive an analysis through the

obtained data from a corpus.

The word corpus comes from the Latin word corpus, which means 'body' in French. The plural versions of "corpus" are referred to using the word 'corpora' or 'corpuses'. The term corpus refers to a collection of things.

It also refers to a vast collection of texts that have been stored and processed electronically; the contents may be in written or spoken form.

The oxford companion to the English Language (1992) defines corpus as "a body of knowledge"; "a collection of texts, utterances, or other specimens that are thought to be more or less representative of a language and are frequently kept in an electronic database". Neselhauff (2011) defines corpus linguistics as “a methodology of linguistic analysis that views naturally occurring language as a credible source for the investigation and classification of linguistic structures” and according to Hanks (2012) Corpus linguistics is primarily concerned with interpreting language patterns in word meaning or syntactic composition. Also John Sinclair, 1991says that corpus is defined as “a collection of naturally occurring language text, chosen to characterize a state or variety of a language”.

The British National Corpus (BNC) has about 100 million words in it, as transcripts, written texts account for 90% of the total and spoken texts account for 10% .in the same context Cook suggests that:

"The word corpus refers to a databank of languages 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 ".... G. Cook (2003:73). Applied Linguistics

To sum up, in light of what has been discussed so far, definitions imply that a corpus is the collection of texts that makes up a corpus. Linguists, professors, and students can benefit greatly from a corpus in this regard. It enable researchers to collect and analyze very large amounts of data with computational technologies. The written texts come from a variety of

sources, including magazines, books, diaries, newspapers and popular culture. In addition, any recorded formal or informal spoken text can be used such as dialogues, radio shows, classrooms discussions, and political debates. This spoken texts are captured and written down.

Corpus, in general, is any electronic collection of documents or information that have been sampled to be as representative of a language as possible.

3. Qualities of a Corpus

A corpus differs from a text analysis in various ways. A corpus has the following properties, according to McEnery and Wilson (1996):

Representativeness: in linguistics, we are more interested in complete languages than individual writer's texts. We use a sample of language that is more or less typical of the variety because analyzing the entire speech of a given variety is impractical.

One of the most important qualities of a corpus as a text is its representativeness. When creating a corpus for a specific language variety, we want to make it as representative as possible so that we may get the most accurate picture of the data's tendencies and frequencies. It would not be representative of mid-nineteenth-century English to limit our examination to Charlotte Bronte's books, nor would it be representative to confine to the category of literature.

Limited size: The corpus is not a random set of words with collocation stops (large). For example has 100.000.000 words. The monitor body, which is an open structer that continuously adding text, is an exception. The corpus is intended for lexicographical work involving new words or word inflections.

Readers of Machines: For a long time, the term corpus was only used to refer to printed material. The phrase now denotes that the text's body is machine-readable. A computerized

corpus has the advantage of requiring an end-to-end search to analyze, for example, the frequency of the word "book". Such a work can be completed in seconds thanks to indexing software.

Standard reference: Most corpora provide a standard reference that identifies the variety they represent. The Brown corpus of written American English, the LOB corpus of written British English, and the London-lund corpus of Spoken British English are only a few examples. Cherifi (2019).

4. Background of Corpus Linguistics

Before adopting corpus analysis as a primary method for data collection and examination as well as the developments that led to the creation of such vast corpora, the process relied on native speakers' intuition and assessment (Devitt 2020).

As a dependent field of study, corpus linguistics is significantly distinct from other fields and sub-disciplines of linguistics .unlike such fields as phonology, the study of social variation in language, or the critical discourse, corpus linguistics is not limited to the analysis of a specific aspect of language . Rather, corpus is a set of methods and tools that is applicable to many variant fields of study within linguistics.

The field of corpus linguistics, which is acknowledged as a sub-discipline of linguistics has exploded in popularity since the late 1950s. This is due in part to its near complete reliance on computerized text, which had to advance to the extent where it could manage and modify vast amount of machine readable text before anything resembling modern corpus linguistics could exist. That said, corpus linguistics did not appear out of nowhere in the late 1950s and 1960s which was the case for any other intellectual movement. Forerunners can be identified whose work in retrospect appears to have been leading toward the creation of corpus linguistics in its modern form. In particular, approaches to language assessment in reliance on

natural usage data long predate modern corpus linguistics, wherever such approaches made use of large datasets, or was dependent on issues of quantitative analysis the incentive of what we might consider to be corpus linguistics.

5. Early Antecedent

One of the most primal and basic techniques of corpus analysis is the frequency list, despite the fact that compiling frequency lists from hundreds of thousands or millions of words of texts presented an extremely challenging proposition prior to the advent of the computer. For example Kading (1897) investigated letter sequence frequencies in a corpus of 11 million words of German. A more relatable focus was the frequencies of words in lists used in foreign language teaching (Thorndike 1921, Palmer 1933,Fries and Traver 1940, Bongers 1947, and West 1953). However, corpus data was used in other forms as well, for instance in Fries (1952) corpus based descriptive grammar of English, or in Eaton (1940) study of the frequency of word meaning. Alongside this embryonic corpus based work, some fields of investigation such as field linguistics (Boas, 1940) or the study of language acquisition (1924) were fundamentally based almost entirely on the analysis of collections of observant language data, an approach severely identical in principle to that underlying corpus linguistics, thought the datasets in question were mostlylimited in terms of resources by the standards of modern electronic corpora and were manually examined.

Survey based on large text corpora begun to consider computer as the essential tool as soon as computers existed, indeed the first move was not to electronic computers but to the precursor technology of the punched card machine.in (1951) Roberto Busa, working on a corpus of the poetry of Thomas Aquinas, utilized this technology to produce the first machine generated concordances (Winter1999). Bursa's pioneering work was the depart point for the field of humanities computing, or, as it is currently known as digital humanities. Following

this instance, equally prominent research by Alphonse Juilland set the needed foundation for some of the other vital principles of work with corpora of electronic text, such as the significance of meticulously balanced corpus sampling, or the importance of taking into consideration statistics of dispersion without neglecting raw frequencies (Juilland and Chang Rodrigues 1964).

6. Chomsky's Opposition of Corpus Data

Noam Chomsky, a resurrected American specialist in syntaxis both a prominent and a peculiarity central figure in the history of corpus linguistics. He is peculiar because his initial intentions were intended to oppose corpus linguistics.

It is worth delving into Chomsky's philosophical justification for dismissing corpus based linguistic analysis. Chomsky's contribution was to suggest that competence, not performance, or as he subsequently expressed it, internal language rather than exterior language is the essential phenomenon that a language theory should address. There is an abstract knowledge in the mind of the native individual speaker that does not correspond to any particular finite collection of observed utterances; it is this knowledge, for example that allows a speaker to intuitively know that a certain sentence is grammatically incorrect in their language, regardless of whether or not they previously encountered the particular sentence they are judging. Chomsky felt linguists should strive to account for native speakers' intuitive capacity to produce an endless number of legitimate sentence forms. Chomsky did not merely outsource performance research to a lower priority. He explicitly advocated that performance data may present little or to nothing about competency because the number of possible sentences in a language is unlimited, and any collection of performance data can reflect more than a finite, biased and distorted of the language. Moreover, Chomsky deemed naturally occurring language data to be of "degenerate quality" (Chomsky 1965:8), containing

performance defects and non-grammatical forms that do not accurately reflect the ability of the speakers who created them. Chomsky's stance on observed linguistic evidence has stayed consistent over time, even as other components of his ideas have developed. When Andor (2004:97) interviewed him he had this to say:

"Corpus linguistics does not mean anything. It is like saying suppose physics and chemistry decide that instead of relying on experiments, what they are going to do is take videotapes of things happening in the world and they will collect huge videotapes of everything that is happening and from that maybe they will come up with some generalization or insights. Well you know, scientists do not do this. But maybe they are wrong. Maybe the scientists should just collect lots and lots of data and try to develop the results from them"

And added that data collection for the sake of implying it toinvestigate a case study is not an approved research methods by scholars in other fields and proceeded to further more interrogate the effectiveness of the research method.

Chomsky's arguments contain a plethora of points, for example argues that a corpus that is able to completely represent a language is unattainable; Chomsky's proposition that the set of possible sentences is infinite. The aftermath of these arguments contributed to cause a shift within the structure of linguistics away from empirical usage data that can be found in a corpora, and towards rationalism and reliance on intuition and introspection as primary data. This shift was not perfect and definitive. Fields such as phonetics, child language acquisition, the profession of language variation, and other aspects of sociolinguistics all proceeded to depend on observed natural data to a large extent. On the same context, the trend away from empirical data was not as pronounced in the east European tradition. A broadly empirical approach to linguistics retained its importance in the former eastern bloc.

But despite these exceptions, Chomsky's arguments may not have been sufficient to justify the change toward rationality that did occur. For instance, a meaningful corpus analysis does not uniquely require that the corpus contain the entirety of the language been analyzed, as the Chomskyan opposition proposes. It simply requires the corpus to present a representative

sample of thelanguage. Moreover, Chomsky's perspective on how natural sciences work is not inclusive and evasive of other science fields and their work strategy. There exist entire fields of science that meets the exact description of Chomsky and that are dependent on the methodology of analyzing collected data-For instance, Astronomy and Geology. Perhaps as a result of Chomsky's anti-corpus argument inadequacies, the pendulum eventually moved towards empiricism, resulting in asignificant expansion of corpus linguistics. Small groups of researches persisted in exploring the computer-based analysis of large text corpora the period when introspection was the dominant method of mainstream linguistics theory, in many cases using these corpora as sources of frequency data, which was largely inaccessible to introspection. Later, in the 1980s and 1990s, corpus data was greatly rehabilitated in linguistics, to the point where, in the twenty-first century, employing corpus datais no longer regarded as unconventional and inadmissible. Corpus data are becoming increasingly important in the research of linguists. Corpus linguistics started to be evaluated on the basis of the corpus's results rather than on an abstract philosophical argument. Corpora have been demonstrated to be extremely valuable in a variety of fields of linguistics, including contrastive linguistics (Johanson2007), Theoretical linguistics (Wong 2006, Xiao and McEnery 2004), and as a source of data for language description (Hanks 2009). It is becoming increasingly impossible to identify a linguistics field where a corpus technique has not been used.

As a result Chomsky's objection has almost completely dissolved as a major impediment to the development of corpus linguistics as a field. therefore it impacted the emergence of the characteristics of the research groups that dove corpus research forward in the 1960s and 1970s, hence English corpus linguistics study at this timeframe was crucial to the development of modern corpus linguistics.

7. English Corpus Linguistics

Despite the importance of non-English work, English corpus (ECL) served as the crucible for the development of modern corpus linguistics. Within ECL, corpus linguistics provided a significant, systematic contribution to the enhanced description of language's lexis and grammar. Key concepts like collocation and corpus annotation were also created and enhanced within ECL

Early work in ECL was usually concentrated in certain researches centers, with major researches at those universities often leading a formal or informal research group that contributed to corpus-based language studies. Many of these organizations communicated with one another on a regular basis through an organization called ICAME (international computer Archive of Modern English), which was created in the 1970s. According to Leech and Johnson's (2009) account of ICAME's founding and early years, the organization provided a critical framework for a network of scholars to collaborate in the development of ECL as a field. In addition, ICAME gathered and disseminated English corpora. Another aspect contributing to the rapid development of ECL, particularly research focused on British English those were not only more often produced corpora, but they were also among the first to be archived, providing researchers other than their founder's relatively simple access to the resources.

The brown corpus was one of the most important corpora distributed by ICAME. The brown corpus was created first and only later added to the ICAME archive. It was originally given the considerably more lengthy official title of "A standard corpus of present-day Edited American English for use with digital computers" (Frances and Kucera1964) when it was built at Brown University in the United States of America. Notably, the brown corpus was designed from the start to be electronic corpus; unlike the survey of English Usage (SEU), which had begun work a few years prior, it did not begin on paper. Furthermore, Francis and

Kucera focused on two issues that continue to bother corpus builders today, namely balance and representativeness, which were also significant among the concerns raised in the English Usage study. A careful sampling rams was used to create the brown corpus. The writings to be sampled were written in American English in 1961 and published in books or journals. Except for poetry and drama, each text sample had to be at least 2.000 words long (or as close as possible to 2.000 words without breaking up a sentence). And the corpus was planned to have 500 of these samples, totaling one million words. This was a sizable corpus at the time, and it is still sizable by the standers of today. When corpora hundreds of times larger can be managed with ease. The 500 texts were distributed among fifteen categories, writings were picked at random from categories such as (press coverage), (press Editorial), (science fiction), (popular lore), (generic fiction) and (humour).

The brown corpus had a significant impact. For a long time, it was the sole publicly available corpus of American English and as a result, it was utilized in a large number of corpus linguistic research and corpus-based computational linguistics. Second, the model it gave, which combined a precise sampling technique with a well-constructed Genre breakdown, established the parameters for one of the major trends in following corpora, the sample corpus. A sample corpus aims to reflect as accurately as possible a specific variant of a given language at a specific point in time, hence the term "snapshot corpus" for this sort of corpora. Later corpora modeled themselves after the Brown corpus, applying the same genre breakdown and other design requirements.

8. The Neo-Firthian Approach to Corpus Linguistics

The history of corpus linguistics that has been detailed has been done so with the goal of presenting one particular approach to corpus linguistics, or school of corpus linguistics. This is the viewpoint that corpus linguistics. Is basically a method that may be applied to a variety of disciplines of linguistics, testing and refining previous discoveries while also providing no

solutions to fundamental language difficulties? However, this is not the only viewpoint from which the history of corpus linguistics can be examined. Another different approach to corpora that John Sinclair pioneered. From the late 1960s to 1970s, this approach was principally developed at Sinclair's university, the University of Birmingham. Parallel to the tradition that was forming at UCL, Lancaster, and other universities at the time. Most notably, in Sinclair's view, corpus linguistics has a theoretical significance that raises it above its role as a methodology. Sinclair's work is referred to as ‘neo-firthian’ because it is based on the ideas of the linguist firth first, which the neo-firthians extended by applying them to computer-aided corpus analysis.

The word and its discourse context are typically the central focus of investigations in neo-firthian corpus linguistics; thus, word and discourse meaning, phraseology, and collocation, i.e. the pervasive regular lettering of co-occurrence among words in corpus data, have been major, though not exclusive concerns. For examples, in the 1980S, Sinclair COBUILD research units pushed corpus lexicography forwards, resulting in the publication of a series of highly influential dictionaries; corpus data is now the central resource for lexicographers; and, slightly later, a neo -firthian approach to English grammar was developed, in which words and grammar are inextricably bound, in what has been termed 'lexico-grammar' (Halliday 1985 b, Sinclair 1991). In neo-firthian thinking on language, the phenomena of collocation in particular has grown to have not only descriptive but also theoretical relevance. An explanation on some of this conceptual development in the work of Sinclair and other scholars linked with this school will be done.

Given the school's lexicographic foundations, some neo-firthian methods go so far as to make the word, its phraseology, and its collocational properties the keystone of linguistic description and theory. Many elements of language that were previously explained in terms of grammar are now explained in terms of the lexicon, which is the concept of lexical grammar

that we have already mentioned. While most neo-firthians believe that in in linguistic description, Lexis must be prioritized over grammar, the core concept has been expressed in a variety of ways. The centrality of collocation is linked above all the the prominence of meaning in Sinclair's writings.The idea of assigning a central role in linguistics to 'meaning in context, like the concept of collocation itself,comes from Firth but also further developed by Sinclair (2004: 18-20) contends that meaning in running text are not confined to specific words, but have broader scopes and that the beginning and end points of the presentation of a given meaning may not be obvious based on the occurrence of collocation in corpus data. The idea that words are not chosen independently by speakers arises from the predominance of collocation. Rather, meaning units are chosen, and each meaning unit brings with it a string of multiple words. Sinclair (1991: 110) defined this as the 'idiom principle', which states that a speaker «has available to him or her large number of semi-reconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments14. Units that extend beyond the word and hence include collocations of words are referred to as 'extended units of meaning' or a 'lexical item', according to Sinclair (Stubbes 2001: 60). The majority of what appear to be synthetically formed phrases are actually the result of retrieving idioms and other collocation units from the lexicon , rather than the result of 'on-the-fly' operation of a conceptual set of rules (as proposed by many formalist models of grammar in particular).

Only a small percentage of language is created or understood according to rules. According to the open-choice principle which Sinclair coined.

Michael Halliday, whose approach to data affected Sinclair's position (Sinclair 2004 : vii) , and Michael Stubbs , who established an approach to the analysis of discourse in which collocation plays a prominent role, are two researchers affilliated with the neo-firthian school (see Stubbs 1996) . Susan Hunston, Michael Hoey, Bill Louw, Wolfgang Teubert, and Elena

Tognini-Bonelli were among the linguists who worked with Sinclair at Birmingham. In terms of corpus-driven linguistics, Tognini-Bonelli (2001) is responsible for one of the most significant characterizations of the theoretical basis of the neo-firthian method. Teubert's (2004, 2005) attempts to ground the neo-firthian method in order philological traditions and to investigate its conceptual roots are notworthy. Louw (1993) is credited with popularizing the concept of semantic prosody (connotation), which is crucial in neo-firthian methods to word meaning and discourse analysis. While Sinclair's lexico-grammatical approach to language still evokes the lexicon and the grammar as two different (implicitly independent) systems, Hunston's and Hoey's methods go even further, not only highlighting the function of lexicon but also unifying the description of lexis and grammar.WhenevercompletelyPattern grammar, Hunston's contribution to this research, is a paradigm in which language is built up as a succession of connected sequences of fuzzy structures, with collocation providing both coherence and meaning (Hunston and Francis 1999). Meanwhile, in 2005, Hoey developed the lexical priming framework, the first comprehensive attempt to develop a theory of language based on neo-firthian ideas, with Hoey making specific claims about the relationship between frequency and collocation in text corporal, as well as the nature of language in the mind. Words are primed to co-occur with other words, according to this theory, so that a person observing generating a given word X is psychologically primed, or prepared in some manner, to anticipate make one or more of the words that X is associated to in the mind. However, several other neo-firthians, especially Teubert (2005: 2-3), disagree with Hoey's hypothesis of psychological role for collocation. Collocation, and corpus linguistics in general, for Teubert, is a tool for examining meaning in language rather than a window into mind-internal event. One of the most striking features of the recent history of the neo-firthian school of corpus linguistics has been the emergence of such divergent viewpoints from a common root in Sinclair's ideas on collocation and lexicogrammar.

Hoey was one of the researchers who moved from Birmingham to the University of Liverpool in the 1990s. This resulted in the establishment of a new and widely known center for the study of corpus linguistics, similar to Birmingham model. Moves like this contributed to the spread and popularization of Sinclair's ideas outside of his own institution. So, while the conceptual underpinnings of neo-firthian corpus linguistics differ from those of methodologist corpus linguistics, the overall story of how these two schools arose is largely the same: a pioneering center in the postwar UK exploited emergent computing technologies to take a data intensive approach to linguistics, using spontaneously occurring speech and writing as the basis for their investigation of language. Similarly, both the neo-firthian and methodologist schools have developed new insights into language and have successfully propagated their pioneering approaches, in part through the dissemination of ideas and techniques and in part, through the training of new generations of linguists who have gone on to further pursue that approach to language study.

9. The Impact of Corpora: Grammatical Description

To give an idea of the impact of corpus linguistics, we will concentrate on Randolph Quirk, because he is largely responsible for the development of corpus linguistics. In 1959, Quirk founded the Survey of English Usage (SEU) at University College London (UCL). The word SEU, or simply the survey, refers to both the research unit founded by Quirk and the corpus whose development was its first project. This was the first attempt to compile a comprehensive compilation of current English. The SEU was a forerunner of later corpora such as the British National Corpus and the American Corpus (IDE and reppen2004), as it attempted to balance its approach to the American National Corpus by recording both written and spoken English and sampling them across a wide range of genres and contexts. The SEU was a true pioneer in corpus linguistics, with its early phases predating even the Brown Corpus. The SEU was not initially kept on a computer. It was first kept on file cards and only

later converted to computerized format, with the spoken portion known as the London-lund corpus (Svartik 1990). The corpus had one million words of grammatically examined modern British English when it was digitized. The team at UCL, directed by Sydney Greenbaum, would eventually take the lead in building the international corpus of Englishness, which includes Australian, British, Hong Kong, Indian, and Irish English and is still the largest corpus for the comparative study of variants of English.

One of the characteristics that distinguished UCL's contribution to corpus linguistics from the beginning was its involvement with corpus parsing (deep grammatical annotation). Given the importance of the grammatical analysis in the UCL method, it's no surprise that one of the UCL team's most significant contributions to establishing the utility of corpus linguistics was in the domain of grammar production. The reference grammars of English by Quirk. The 1985 grammar was also the first widely distributed modern corpus informed grammar, making its publication a watershed moment in corpus linguistics history. Some of the early grammars, such as the Fries 1940, used corpus evidence. Quirk et al. 1985, on the other hand,established the tone for grammars that followed.

Despite the fact that the survey of English put a lot of work towards producing spoken corpus information, the early UCL grammars all have the written language as their primary focus. The grammars of the time were heavily influenced by a view of speech as a degraded form of communication, as well as hesitancy, slips of the tongue and interruptions. In this respect, it's worth noting that there was no big corpus of spoken English accessible until recently on which to base a study of grammarian spoken English. There had been some pioneering early work, such as Fries (1940), that used small collections of transcribed speech to explore grammar in spoken English, but there were no substantial corpora of spontaneous speech until the creation of the London-land corpus, the spoken section of the British National Corpus and the Santa Barbara Corpus of spoken American English. When such corpora

generated, a wide range of opinions about speech grammar emerged quickly. The orthodox position, according to which grammar in speech exists only in a mangled form, subject to interference from useless performance factors, is one extreme. The development of spoken corpora enabled the development of the polar opposite. Speech is grammatical, but it also has its own grammar that is unique from writing.

10. Corpora Beyond Corpus Linguistics

One striking element of corpus-based research's recent history is the rise in the number of academics who do not identify as corpus linguists but have begun to use corpus-based approaches in their study. Corpus approaches have thus prompted new advances, new findings, and, in some cases, critical reflection on previous work in other domain of linguistics. In the fields of functionalist linguistic theory, cognitive linguistics, and psycholinguistics, corpus-based research has found particularly fertile ground. Consider the adoption of corpus methods within functionalist linguistics as an example of how corpus linguistics has been used within the theoretical Frameworks of various areas of language study.

In terms of theoretical position, much corpus-based research can be defined as functionalist in the broadest sense; this is obvious in more ways than just corpus linguistics and functionalist theory's shared rejection of (Chomsky) formation. The basic goal of functionalism is to comprehend linguistic structure in terms of meanings it conveys and the purposes it fulfills, and then to put it to use. This is a goal that many corpus-based studies share. Biber's (1988) multi-dimensional approach to text-type (register) seeks functional explanations for formal (grammatical) differences, which is exactly what functionalism is concerned with. Similarly, the role of grammatical in much diachronic corpus linguistics research into the history of the English language makes it a functionalist activity ; nevertheless, there is a growing body of work that use corpus-based analysis inside research

addressing essential components of functionalism.

A distinctive mode in functionalist linguistics, the goal is to look at one or more specific grammatical structures and try to find functional motives for their use, such as semantic, pragmatic, or processing variables that can explain grammatical form. Many studies use an approach that involves looking at a series of examples of the structures in issue, assessing each one for the presence or absence of a variety of functional elements, and then looking for a structural link between these aspects and the syntactic form. The use of corpus data as the source of the set of examples to be analyses is a common application of corpus data in functionalist linguistics. The practice of using less rigorously sampled sets of instances gave way to the use of corpora throughout time. As an example, Birner's (1994) research on non-interrogative subject -verb inversion in English (sentences with the order XVS instead of SV) is based on 1,778 examples of this structure found in various written and spoken sources by Birner and others. Notably, before corpus linguistics became a well-known discipline, any collection of linguistics examples even one gathered on arbitrary basis was commonly referred language as a 'corpus'. However, we observe the beginning of a movement towards the incorporation of corpus data (in the ordinary sense) into functionalist-theoretical analysis in some studies using relatively small and arbitrary 'corpora', particularly in the 1980s and 1990s. Cardin's (1972) study of backwards anaphora (anaphora) is based on a 'corpus' of twelve texts (six children's books); while this is remarkably small to be recognized precised and representative , Cardin's study clearly demonstrates an awareness of the significance of representativeness.

More functionalist research have used large and/or standardized corpora of the type familiar to corpus linguistics from the mid to late 1990s. These can be classified based on how thoroughly the corpus evidence is handled. Some syntax studies treat the corpus purely as a repository of instances, rather than taking a systematic method to dealing with the corpus

evidence as a whole. Decklerk and reed (2000), for example, utilize the cobuild, Brown, LOB, and ice corpora to show examples of English clauses indicated by conjunction. Other studies employ corpus data in a more methodical manner. Valera (1998), for example, investigates the phenomenon of subject- directed adverbs. Valera's database is derived from the LOB corpus in a methodical and comprehensive manner. Similar approaches were used by Mckoon and Macfarland (2000) to investigate the transitivity and subject/object types of internally versus externally caused change of state verbs, using a combination of large English Corpora , by Tempeley (2003) in an investigation into the factors underlying the use of zero relative pronouns in English, using the Penn Treebank, and by Holman (2005) .

The important point here is not only that the corpus provides fresh views on language analysis; it is also important to emphasize that many (intra-) disciplinary barriers to the widespread use of corpus data have now been overcome. Corpora can be utilized to solve a wide range of linguistic problems, not only by self-identified corpus linguists, but also by researches that do not identify themselves as corpus linguists but are willing to apply corpus-based methodologies.

11. Size of Corpora

The size of a corpus is an important factor to consider. In today's world, millions of words are saved and analyzed in a matter of seconds (L. Burnard). What it is clear is that the size of every corpus can range from thousands to millions of words to a few tens of millions of words. As a result, there is a balance between huge corporations with specific interests and small businesses. And also aiding linguists and lexicographers in the creation of dictionaries

Smaller corpora (a few thousand words), are on the other hand tailored for specialized research. The following section displays a small selection of corpus sizes:

Bank of English: 400 million words.

COLLINS Birmingham University - international language Database (COBUILD Corpus):

more than 200 million characters.

British National Corpus (BNC): 100 million words

Longman Lancaster Corpus: 30 million words

American National Corpus (ANC): 11.5 million words

ARCHER Corpus: 2 million words

Helsinki corpus: 1.5 million words

International Corpus of Learners of English (ICLE): 1 million words

Brown Corpus: 1 million words

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

Freiburg Corpus: 1 million words

Survey of English Usage Corpus (SEU): 1 million words

International Corpus of English (ICE): A collection of corpora, each corpus contains 1 million words-in-progress collection.

Helsinki Corpus of Older Scots: 830.000 words

Corpus of Spoken American English (CSAE): 200.000 words

Melboume-survey Corpus: 100.000 words

In today's world, there are millions of corpora in various languages, sizes and objectives.

They are primarily used for linguistic, sociolinguistic, and educational purposes around the world.

12. Capability and Limitations of Corpus Data

Researchers can use corpus data to interrogate and deduce the meaning of words based on recurrent grammatical parents as well as the collection of words in question. It is feasible to make well-informed judgment about how words behave in specific contexts and registers with the help of large banks of text.

It is important to mention that , regardless of how large the source material is , a corpus

and its findings are merely a reflection of what is chosen from the (ultimately finite) data it includes . As a result, according to Basarally (2011), absolute linguistic accountability is impossible to achieve. As a result, Leech (in Kennedy 1998: 23) advises caution ' some sentences won't occur because they are obvious, others because they are false, still others because they are impolite.

Furthermore, corpus data can only assist academics in determining the range of any underlying meaning a word may have and determining whether collocations are plausible rather than linguistically acceptable. As a result, observations should not be treated as concrete representations of language usage, but rather as reasonable evidence based on the information provided.

Conclusion

This chapter has primarily focused on the theoretical aspects of corpus linguistics, the retrospective development of methods and means of investigation, some of the key figures who have had a significant impact on this field of linguistics and some of the most popular corpora in our day and age. In the next chapter, the focus will be narrowed to focus on three particular tools of investigation to set the foundation for result analysis.

Chapter Three, Methodology

1. Introduction

This chapter focusses on how language is used in different context according to the National British Corpus, the Michigan Corpus of Academic Discourse and the web corpus. This chapter is divided into three parts. The first section of the chapter describes an entry to the corpora in general meaning and it focuses upon major English corpora, which are grouped in terms of their primary uses so that readers will find it easier to choose corpus resources suitable for their particular research questions. The second section discusses the historical background and structure of the each corpus moreover, how it used by learners and their purpose.

2. Methodology Choice

This chapter presents object of the study, approach of the study, and method of the study. Firstly, the object of the study explains how "Algeria" is used in the real-world text. Secondly, the approach of the study discusses the tool to examine the problems of this research. Finally, the method of the study describes the processes in which the data were collected and analyzed.

2.1. Aim of the Study

The object of this research was clauses that were taken from British National Corpus at https://www.english-corpora.org/bnc/. British National Corpus is a collection of written and spoken world text, which contains more than a hundred million clauses that are published in this corpus. Because it contained lot of clauses, these objects were collected from many sources, which those sources collected from newspapers, fictions, politics, laws, and educations. Therefore, the use of "Algeria" in real world application is wider and gets various data.

2.3. Approach of the Study

The approach of this research is semantics. The research utilized lexical analysis to identify the context. The goal of using Linguistic context is to understand how the speakers or users employ the language in proper situation to avoid ambiguity. Thus, it is appropriate to reveal the Linguistic context of the dominant meanings of "Algeria".

2.4. Method of the Study

In process of conducting the research, two steps were executed. First was collecting the data and second was analyzing the data. Those steps were presented below.

2.5. Data Collection

This research applied qualitative method on several samples. The clauses from British National Corpus were filtered and selected to analyze. The steps of data collection were as followed.

First step to do was collecting the first 100 clauses containing the word "Algeria" that were taken from British National Corpus Using a computer, the researcher wrote the word "Algeria" as a keyword at https://www.english-corpora.org/bnc/. From the first 100 clauses, only the clauses that used the word "Algeria" as a country and word were selected. The clauses containing "Algeria" in phrasal verb and prepositional verb were eliminated and were not used as the data. After the filtering, 60 clauses remained and were used as the data of this research. Additionally, this research used 60 clauses which contained of spoken and written texts.

3. The National Corpora

National corpora are normally general reference corpora which are supposed to represent the national language of a country. They are balanced with regard to genres and domains that typically represent the language under consideration. While an ideal national corpus should cover proportionally both written and spoken language, most existing national corpora and

those under construction consist only of written data, as spoken data is much more difficult and expensive to capture than written data. This section introduces a number of major national corpora.

4. The British National Corpus

4.1. History of the British National Corpus

According to Burnard (1998) the BNC involved a collaboration of three publishers, including Oxford University Press, Longman, and W. & R. Chambers, as well as two universities, the University of Oxford and Lancaster University, and the British Library. The BNC was established in 1991 under the direction of the BNC consortium, and the project was completed in 1994. After 1994, no new samples were added, but the BNC was slightly revised before the introduction of the second edition BNC World 2001 and the third edition BNC XML Edition 2007.The BNC was the vision of computational linguists who seeks to develop a corpus of modern, naturally occurring language in the form of speech, text, and writing that could be studied by computers. As a result, it was constructed as a broad corpus to lay the foundations for computerized corpus linguistics search and processing. One of the ways the BNC set itself apart from other corpora at the time was to make the data available for commercial and educational purposes as well as academic study.

The corpus was limited to British English only and did not include the world's versions of English. This was partly due to the fact that the British government was funding a large chunk of the research, and the British government was logically interested in promoting documentation of its own language variation. The BNC needs funds from commercial and academic institutions as well, due to its possible unprecedented size. BNC data were available for commercial and academic research as a result.

4.2. The British National Corpus Structure

The British National Corpus (BNC) is considered as a monolingual corpus. It is a collaboration between commercial and academic partners. It is a 100 million words collection of written and spoken language from variety of sources, intended to represent a board cross-section of British English from the late twentieth century also, it aims to be representative of a whole range of English language currently used in speech and writing. The BNC is divided into two parts: the written part and the spoken part. First, the written part is a 90% of the whole BNC in which it includes newspapers, academic books, letters, and essays. Second, the spoken part which is only 10% that contains informal conversations, radio shows, etc.

The main tasks of corpus development can be listed as: corpus design, acquisition and preparation of data, corpus processing and making corpus material available to end-users. The focuses on four main areas of application, i.e. linguistic research, reference publishing, natural language processing by computer and language teaching.

BNC is extremely important for the study of spoken and signed language because, unlike written language, speech, sign, and gestures vanish once they are produced, requiring the use of multimodal corpus to investigate interactive face-to-face communication.

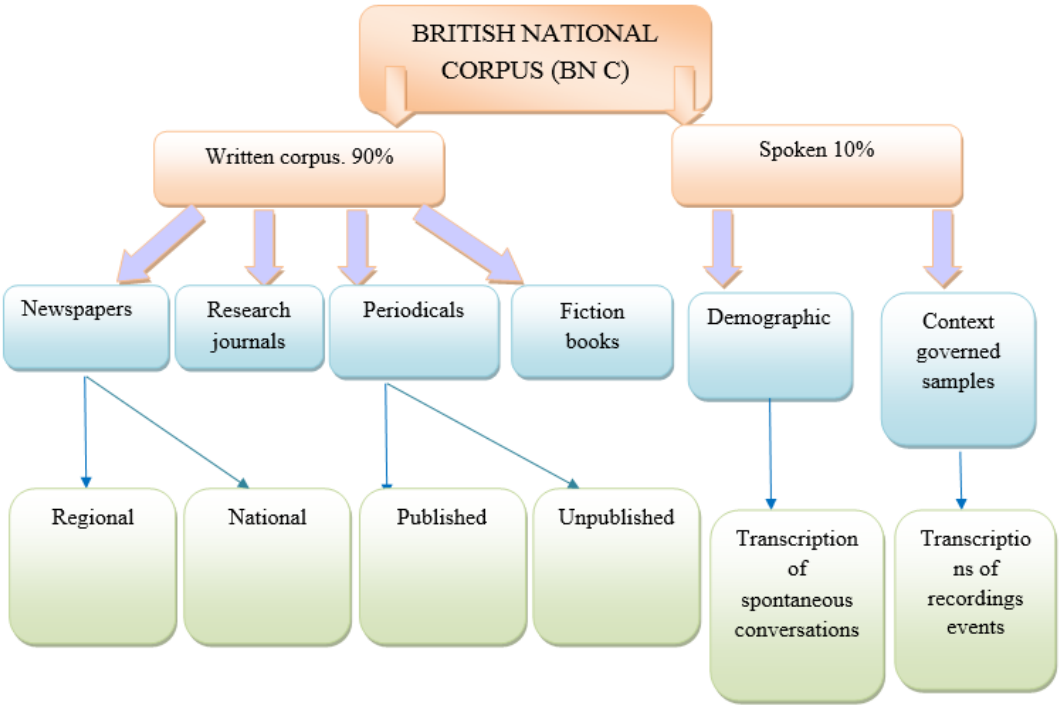


Figure01: The British National Corpus Structure

4.3. The Uses of the British Nation Corpus

For text analysis, learners use both the BNC and the Corpus of contemporary American English (COCA), the main principal applications of corpus use are:

- Using both helps ensure that the user gains better overall understanding of the global use of English, not only British English.
- Language is a living thing and many words traditionally are considered to belong to American English are used by British English speakers, and vice versa.
- Each has their own advantages over the other. For example, the BNC includes more informal, everyday conversation whereas the COCA is much larger in size and was created more recently. This means they complement each other well.
- Reference Book Publishing: Dictionaries, grammar books, teaching materials, usage guides, thesauri. Increasingly, publishers are referring to the use they make of corpus facilities: it's important to know how well their corpora are planned and constructed.

- Linguistic Research: Raw data for studying lexis, syntax, morphology, semantics, discourse analysis, stylistics, sociolinguistics...
- Artificial Intelligence: Extensive data test bed for program development.
- Natural language processing:

Taggers, parsers, natural language understanding programs, spell checking word lists...

- English Language Teaching: Syllabus and materials design, classroom reference, independent learner research.

4.4. The British Nation Corpus Construction

The British National Corpus is a construction of 100 million word corpus of modern British English for linguistic research. Oxford University Press (OUP), Longman Group UK Ltd., Chambers, Lancaster University's Unit for Computer Research in the English Language (UCREL), Oxford University Computing Services (OUCS), and the British Library collaborated on this pre-competitive endeavor. The UK Department of Trade and Industry and the Science and Engineering Research Council are sponsoring the initiative under their Joint Framework for Information Technology.

4.5. Sorts of the BNC

According to Burnard (2009) there are several sorts of, such as:

Monolingual:

It deals with modern British English, not other languages used in Britain. However non-British English and foreign language words do occur in the corpus.

Synchronic:

It covers BE of the late twentieth century, rather than the historical development which produced it.

General:

It includes many different styles and varieties, and is not limited to any particular subject field, genre or register. In particular, it contains examples of both spoken and written language.

Sample:

For written sources, samples of 45,000 words are taken from various parts of single-author texts. Shorter texts up to a maximum of 45,000 words, or multi-author texts such as magazines and newspapers, are included in full. Sampling allows for a wider coverage of texts within the 100 million limit, and avoids over-representing idiosyncratic texts.

4.6. Creating the BNC

Making the BNC was a joint effort of a large number of participants; organizations and individuals. It comprised two main stages: the planning (design stage) and the execution (creation stag).

4.6.1. The Design stage

The BNC project started with a careful planning stage where the design principles for the corpus were drawn up. These established a number of selection criteria which were then used for identifying suitable texts to be included in the corpus. In addition to the selection criteria for the written and spoken components, a large number of classification features were identified for the texts in the corpus.

A /Selection Criteria: Written texts

Texts were selected for inclusion in the corpus according to three independent selection criteria: domain, time, and medium. Target proportions were defined for each of these criteria, as listed below.

Domain: The domain of a text indicates the kind of writing it contains:

- 75% of the written texts were to be chosen from informative writings: of which roughly equal quantities should be chosen from the fields of applied sciences, arts, belief & thought, commerce & finance, leisure, natural & pure science, social science, world affairs.
- 25% of the written texts were to be imaginative, that is, literary and creative works.

Medium: The medium of a text indicates the kind of publication in which it occurs. The classification used is quite broad:

- 60% of written texts were to be books
- 25% were to be periodicals (newspapers etc.)
- Between 5 and 10% should come from other kinds of miscellaneous published material (brochures, advertising leaflets, etc.)
- Between 5 and 10% should come from unpublished written material such as personal letters and diaries, essays and memoranda, etc.
- Small amount (less than 5%) should come from material written to be spoken (for example, political speeches, play texts, broadcast scripts, etc.)

Time: The time criterion refers to the date of publication of a text. Being a synchronic corpus, the BNC should contain texts from roughly the same period. The intention was that no text should date back further than 1975. This condition was relaxed for imaginative works only, a few of which date back to 1964, because of their continued popularity and consequent effect on the language.

B/Classification features: Written texts

In addition to the selection criteria, a large number of classification features were identified for the texts in the corpus. No fixed proportions were specified for these features, although the intention was to make sure that there should be an appropriate level of variation within each criterion. The classification criteria include such things as:

- Sample size (number of words) and extent (start and end points).

- Topic or subject of the text.
- Author's name, age, gender, region of origin, and domicile.
- Target age group and gender.
- "Level" of writing (a subjective measure of reading difficulty): the more literary or technical a text, the "higher" its level.

Information was added when available which means that the amount of information added to each text varies.

4.6.2. The Spoken Component

There are two parts to the 10-million word spoken corpus: a demographic part, containing transcriptions of spontaneous natural conversations made by members of the public and a context-governed part, containing transcriptions of recordings made at specific types of meeting and event.

All the original recordings transcribed for inclusion in the BNC have been deposited at the National Sound Archives of the British Library.

A/The Demographic part of the Spoken Corpus

A total of 124 volunteers were recruited by the British Market Research Bureau. The volunteers came from four social groupings (AB, C1, C2, and DE). There were male and female volunteers from a wide range of ages, and they lived at 38 different locations across the UK. Recruits were chosen in such a way as to make sure there were equal numbers of men and women, approximately equal numbers from each age group, and equal numbers from each social grouping.

Recruits used a personal stereo to record all their conversations unobtrusively over two or three days, and logged details of each conversation in a special notebook. Those who took part in the recordings were asked after the conversation to give permission for their speech to be included in the corpus.

Information about the participants, such as age, sex, accent, occupation, was recorded when available.

B/The Context-Governed part of the Spoken Corpus

The intention was to collect roughly equal quantities of speech recorded in each of the following four broad categories of social context:

- Educational and informative events, such as lectures, news broadcasts, classroom discussion, tutorials.
- Business events such as sales demonstrations, trades union meetings, consultations, interviews.
- Institutional and public events, such as sermons, political speeches, council meetings, parliamentary proceedings.
- Leisure events, such as sports commentaries, after-dinner speeches, club meetings, radio.

- Phone-ins. Information about the participants, such as age and sex, was recorded when available.

5. The Michigan corpus

5.1. The Michigan corpora: MICASE + MICUSP

In this paper, we provide a detailed account of the steps that were central to designing and compiling the Michigan Corpus of Upper-level Student Papers (MICUSP). MICUSP is a new collection of 829 papers (around 2.6 million words) written by University of Michigan students in their final undergraduate year or in their first three years of graduate education. The papers come from sixteen disciplines, ranging from Humanities and Arts to Physical Sciences, and represent a range of different text types. In this paper, we offer an overview of

the design of MICUSP, the online submission process used to collect papers, and the text-type classification of the papers.

The two significant language corpora, the Michigan Corpus of Spoken Academic English (MICASE), and the Michigan Corpus of Upper-Level Student Papers (MICUSP) were developed at the English Language Institute. Teachers and researchers across the world use these, and materials historically important collections of language for linguistic analysis, classroom activities development.

MICUSP: The Michigan Corpus of Upper-Level Student Papers.

MICASE: The Michigan Corpus of Academic Spoken English.

5.1.1. History of the MICASE

In 1997, the English Language Institute (ELI) at the University of Michigan started the MICASE project. Dr. Rita Simpson was the original project manager. The project was driven by two questions:

1. What are the characteristics of contemporary academic speech—its grammar, its vocabulary, its functions and purposes, its fluencies and dysfluencies?
2. Are these characteristics different for different academic disciplines and for different classes of speakers?

Because MICASE aimed to record a wide range of academic speech, our sampling goals spanned fifteen different types of speech events and four major academic divisions within those types (Humanities and Arts, Social Sciences, Biological and Health Sciences, and Physical Sciences). We adopted stratified random sampling. Each recording is classified according to speech event type, a pre-assigned number indicating the academic discipline, two letters representing the majority of participants in the event (e.g. junior undergraduate, senior faculty, staff), and a final three digit sequence to track chronologically when the tape was recorded.

All recordings were made with a digital audio tape recorder with two external stereo microphones, and at selected events, a video recorder. Two researchers attended most speech events in order to identify speakers and facilitate transcription by taking field notes about nonverbal contextual information; however, in small groups (e.g. advising sessions, office hours, study groups) where an observer's presence would have been intrusive, the research assistants left the room after the equipment was set up. All speech was recorded with written consent from the major speakers and verbal consent from other participants. Demographic information (sex, age group, university position, and native language) was collected from each speaker on a form distributed at the end of each event. The speaker information is included in the header of each transcript and is also entered into a separate database. All DAT recordings were captured and stored as MP3 format sound files for use with our computer transcription program, Sound Scriber, and have also been re-digitized as WAV format files and transferred to data CD for archival purposes.

In June 2001, the first phase of the project was completed, with over 190 hours of academic speech recorded. In April 2002, the transcription and proofing of all transcripts was completed (approximately 1.8 million words).

Then, in May 2002, the original search interface was launched, with a redesigned version released in June 2007. It has grown in popularity each year since its release, approaching nearly 140,000 hits in 2006. In 2009, we are excited for the release of a number of new features and support tools, including new MICASE online demos and new resources for EAP/ESL teachers

5.2. The MICASE:

The Michigan Corpus of Academic Spoken English (MICASE) is a collection of nearly 1.8 million words of transcribed speech (almost 200 hours of recordings) from the University of Michigan in Ann Arbor, created by researchers and students at the U-M English Language

Institute. MICASE contains data from a wide range of speech events (including lectures, classroom discussions, lab sections, seminars, and advising sessions) and locations across the university.

Project leader: Dr. Ute Römer, University of Michigan

Time of compilation: 1997–2002

Size: 1.8 million words

Language: English

Number of texts/samples: 152 (over 190 hours)

Period: 1997–2001

Released: 2002; new interface 2007

Funding: English Language Institute, University of Michigan

Project home page: http://quod.lib.umich.edu/m/micase/

Figure 02: The MICASE Home page

5.2.1. The MICASE Structure:

The Michigan Corpus of Academic Spoken English contains approximately 1.8 million words in the form of 152 transcripts of nearly 200 hours of recordings of 1,571 speakers, focusing on contemporary university speech within the domain of the University of Michigan. Table 1 shows the structure of the corpus

Criterion	Distribution
Speaker gender	Male (46%) Female (54%)
Academic role	Faculty (49) Students (44%)
Language status	Native speakers (88%) Non-native speakers (12%)
Academic division	Humanities & Arts (26%) Social Sciences & Education (25%) Biological & Health Sciences (19%) Physical Sciences & Engineering (21%) Other (9%)
Primary discourse mode	Monologue (33%) Panel (8%) Interactive (42%) Mixed (17%)
Speech event type	Advising (3.5%) Colloquia (8.9%) Discussion sections (4.4%) Dissertation defenses (3.4%) Interviews (0.8%) Labs (4.4%) Large lectures (15.2%) Small lectures (18.9%) Meetings (4.1%) Office hours (7.1%) Seminars (8.9%) Study groups (7.7%) Student presentations (8.5%) Service encounters (1.5%) Tours (1.3%) Tutorials (1.6%)

Table01: The MICASE corpus Structure

Academic events vary widely in their tone, substance, and length. The MICASE corpus includes speech events that range in length from 19 to 178 minutes, with word counts ranging from 2,805 words to 30,328 words. In the MICASE corpus, academic speech is defined as that speech which occurs in academic settings. In other words, it is not pre-defined as something like "scholarly discussion." In academic settings, we might, for example, find such speech acts as jokes, confessions, and personal anecdotes, as well as definitions, explanations, and intellectual justifications. Therefore, the MICASE researchers have taken pains to record a wide variety of academic speech events. Most speech events are fully recorded, from beginning to end, because the beginnings and ends of academic speech events may be of particular interest to researchers.

5.3. The British Academic of Spoken English Corpus:

According to P. Thompson and H.Nesi (2001) the British Academic Spoken English (BASE) corpus, which is designed as a British counterpart to the MICASE, is under construction at the Universities of Reading and Warwick. The corpus currently comprises a collection of recordings and marked up transcripts of 160 lectures (63 from Reading and 97 from Warwick, totaling 127 recording hours) and 39 seminars (from Warwick, 32 hours). The lectures and seminars spread evenly across four subject areas, as shown in Table 02

Subject area	Lectures	Seminars
Arts and Humanities	42	10
Social Studies and Sciences	40	11
Physical Sciences	40	8
Life and Medical Sciences	38	10
Total	160	39

Table 02: The British Academic of Spoken English Corpus

6. The Web Corpus

6.1. Corpus Linguistics and the Web

According to McEnery and Wilson (1996) any collection of more than one text can be called corpus. But the term corpus when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition using the Web as a Corpus is a recent challenge in corpus linguistics.

The practical issues such as appropriate linguistic search tools for accessing the internet, the issue of the issue of register variation, and methods for culling data from the web. It includes a variety of case studies on morphology, syntax, lexis, and synchronic and diachronic variation in English, and shows the two approaches the web in corpus linguistics- web-as-corpus and web-for-corpus-building-are used.

6.2. The Web Corpus

The Web corpus contains 14 billion words about 14 times the size of COCA in 22 million web pages. It is related to many other corpora of English that we have created, and they offer unparalleled insight into variation in English.

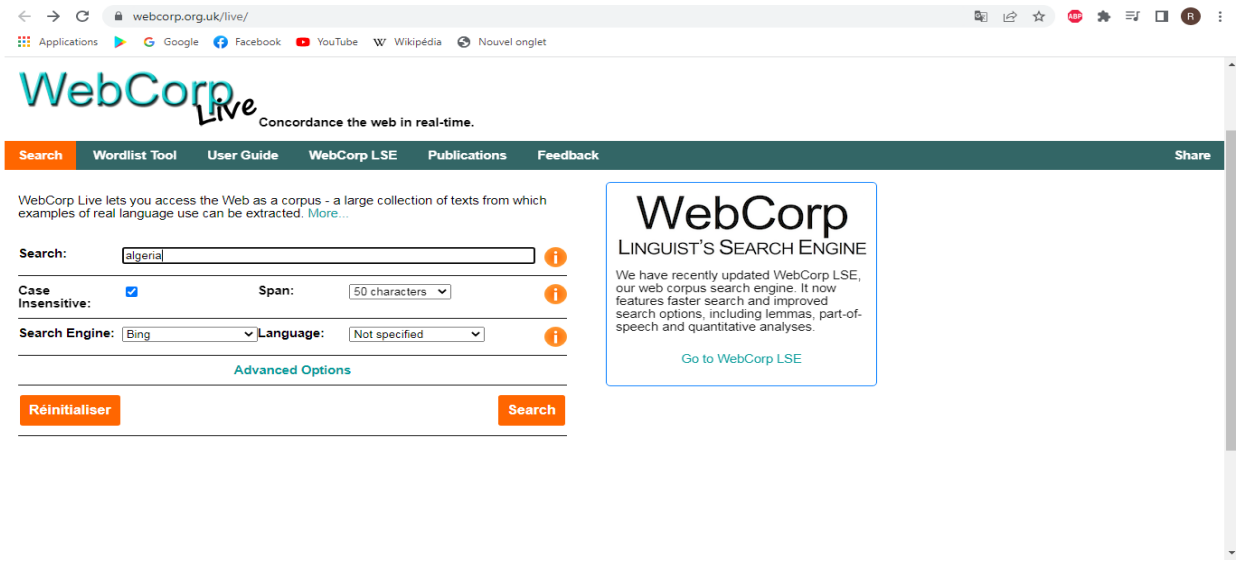


Figure03: the web corpus home page

Conclusion

This chapter provides an overview of approaches within corpus linguistics that address each of the BNC, the MICASE and the web corpus and how they investigate the use of the word Algeria in English. The first half of the chapter provides an introduction to corpora in general, focusing on significant English corpora that are organized by principal uses to make it easier for readers to identify corpus resources that are appropriate for their research objectives. The second half delves into the history and structure of each corpus, as well as how learners use it and what it's for.

Chapter Four, Results and Findings

1. Introduction

The discussion and the elaboration of the data analysisiswillbedone in two Parts. The first stepistracking the occurrence of the word "Algeria" depending on the sentences in which the wordissuedfrom the clauses in British National Corpus and number of frequency. In this part, all the possible meanings that are found from the Part of speech willbediscussed by providing the semanticfeatures to see the distinctions and similarities. The second part is examining the context of those possible meanings based on theirlinguist contexts.

2. The Word Algeria as itAppears in "British National Corpus"

The first step is to enter the site Britch National Corpus at: https://www.englishcorpora.org/bnc/ and log in.

2.1. How frequent the word Algeria is?

After logging in, we go to the search box. We type the word«Algeria" and click on findingmatching string and here we have the frequency table; it shows us how many times the word "Algeria" appeared in the corpus.

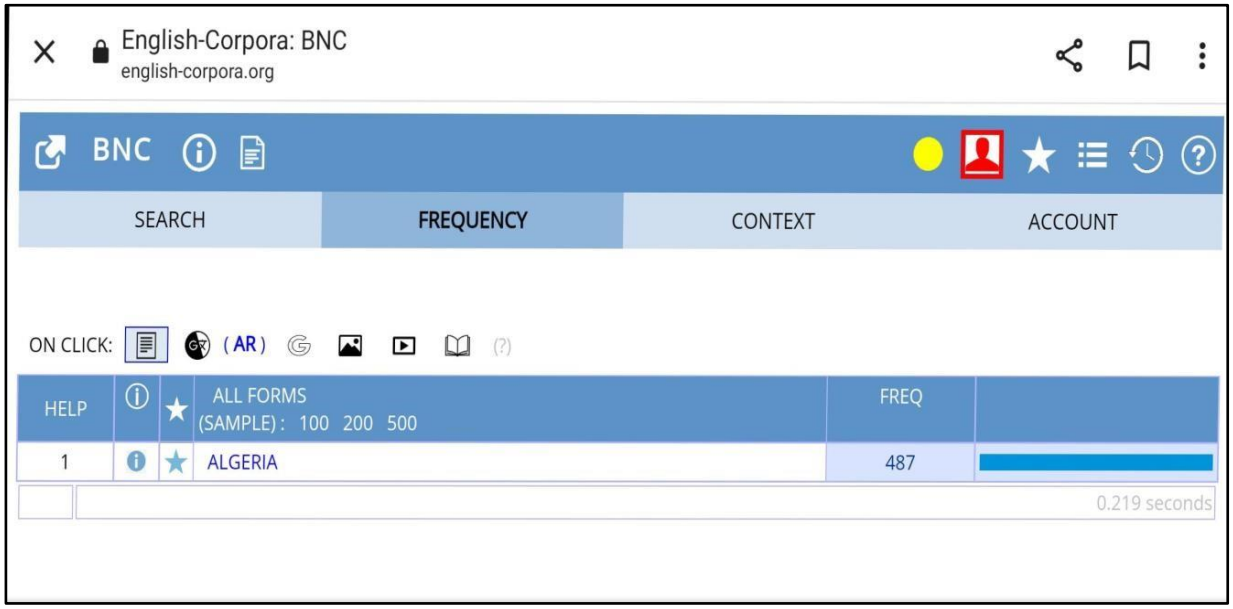
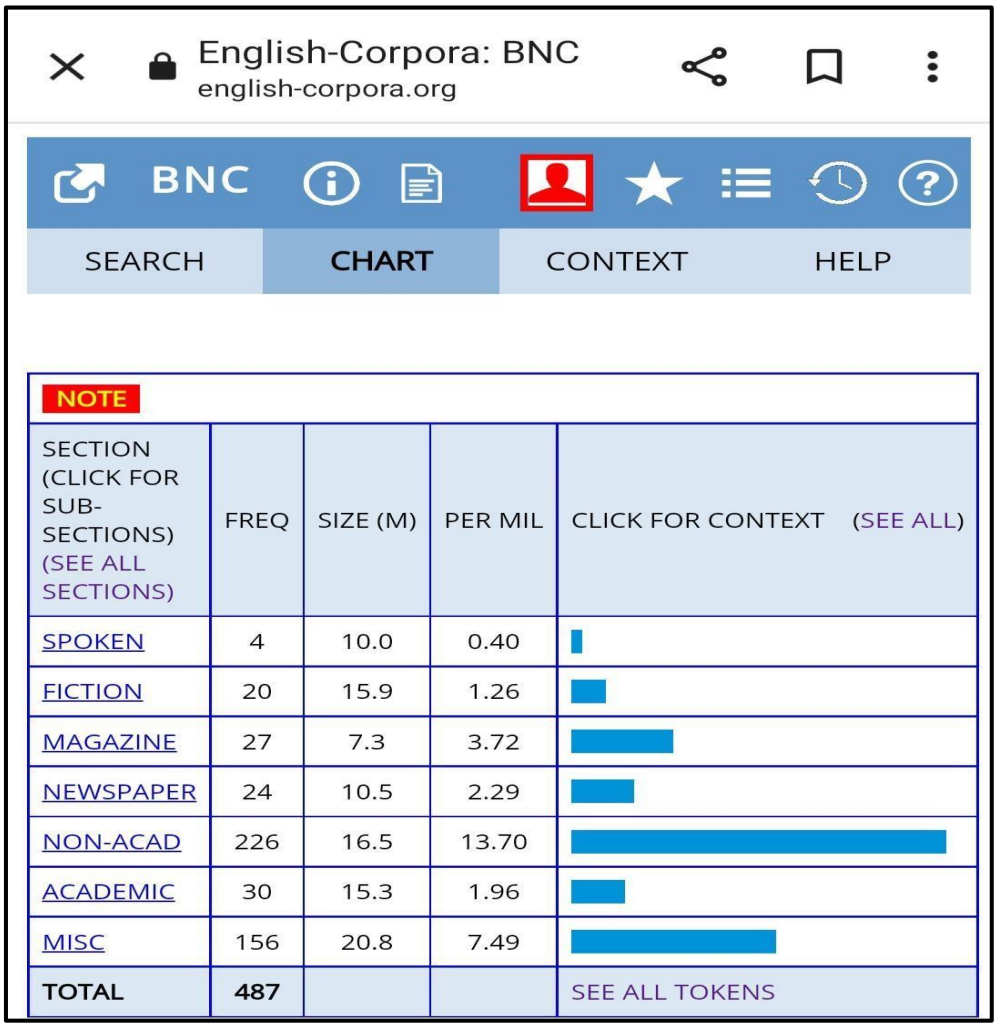


Table 01: the frequency of the word "Algeria" on corpus

Resultsseem to validatethat the word “Algeria” canbefound (487 times) in the corpus. As shown in the table.

2.2. Frequency of the word “Algeria”



SECTION (CLICK FOR SUB-SECTIONS) (SEE ALL SECTIONS)	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
SPOKEN	4	10.0	0.40	
FICTION	20	15.9	1.26	
MAGAZINE	27	7.3	3.72	
NEWSPAPER	24	10.5	2.29	
NON-ACAD	226	16.5	13.70	
ACADEMIC	30	15.3	1.96	
MISC	156	20.8	7.49	
TOTAL	**487**			SEE ALL TOKENS

Table 02: Shows results for the frequency of the word “Algeria” across different genres

According to the British National corpus (BNC) website and as the chart clearly shows, the word “Algeria” variedfrom one type of text to another. It has the mostcommon occurrence in the non-academic genre followed by MISC. The categories' fiction', 'magazines,' 'newspapers', and 'academic' have averagefrequencies and maybealmost non-existentif we compare them to the 'non-academic' category, which has the highestfrequencies. Whereas the 'spoken' category

has the lowestfrequency. This exampledemonstrates the benefits of aggregateswhen associating items with categories.

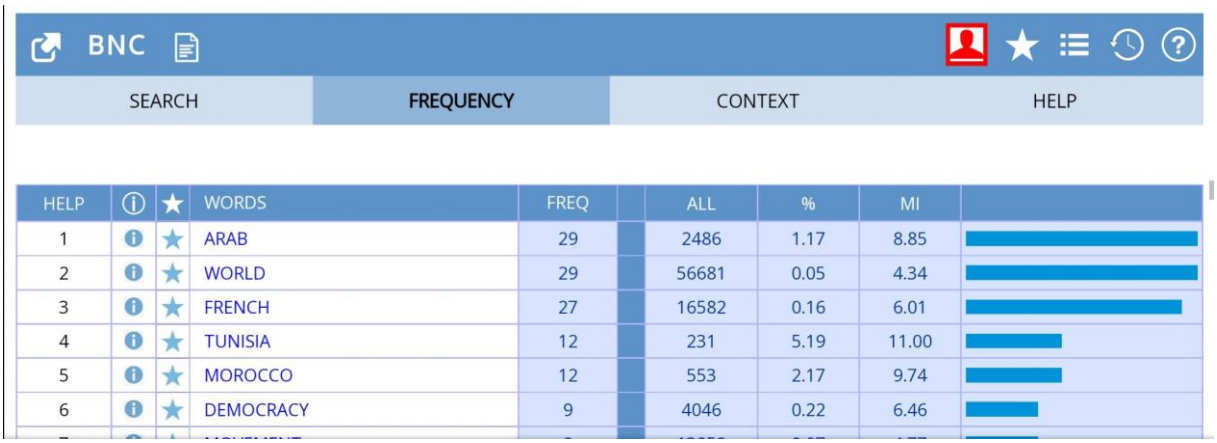
2.3. Collocations of the word “Algeria”

HELP	?		FREQ	ALL	%	MI
1	☐	MOROCCO	32	553	5.79	10.48
2	☐	TUNISIA	31	231	13.42	11.70
3	☐	LIBYA	30	689	4.35	10.07
4	☐	ARAB	30	2486	1.21	8.22
5	☐	WORLD	30	56681	0.05	3.71
6	☐	FRENCH	29	16582	0.17	5.43
7	☐	FRANCE	14	11969	0.12	4.85
8	☐	MAURITANIA	9	108	8.33	11.01

Table 03: collocation of the word Algeria words that occur one to the left

From the table, we see that the countries that occupy the first threeranks in the results table (Tunisia, Libya, and Morocco), becausetheysharegeographicalborderswithAlgeria, and they have political, economic, and social relations. GiventhatAlgeriahastense relations withMorocco, Moroccoranks first in the results table, due to the conflictbetween the

twocountries since the 19th century. France occupiesboth the sixth and seventh place, unlikeotherEuropean countries, becauseAlgeriawas a colony of France in the olddays.



HELP	WORDS	FREQ	ALL	%	MI
1	ARAB	29	2486	1.17	8.85
2	WORLD	29	56681	0.05	4.34
3	FRENCH	27	16582	0.16	6.01
4	TUNISIA	12	231	5.19	11.00
5	MOROCCO	12	553	2.17	9.74
6	DEMOCRACY	9	4046	0.22	6.46

Table 4: collocations of the wordAlgeriawordsthat occur one to be to the right

According to the table, the wordsthat occupy the first three ranks in the results table (Arab, world, and French), the wordArabranks first becauseAlgeriaisconsidered an Arab country, and Arabicisits first official language.

3. Web corpus

3.1 Concordance lines

The website of the web corpus reveals to us the occurrence of the wordAlgeria in a plethora of formal websites. Five of suchwouldbechosen to be used as a model to illustrate the concordance lines in which the wordAlgeria has appeared.

1) https://www.gov.uk/foreign-travel-advice/algeria

Text,Wordlist,text/html, UTF8 (Content-type), 2022-03-17 (Body near 'Last Modified')

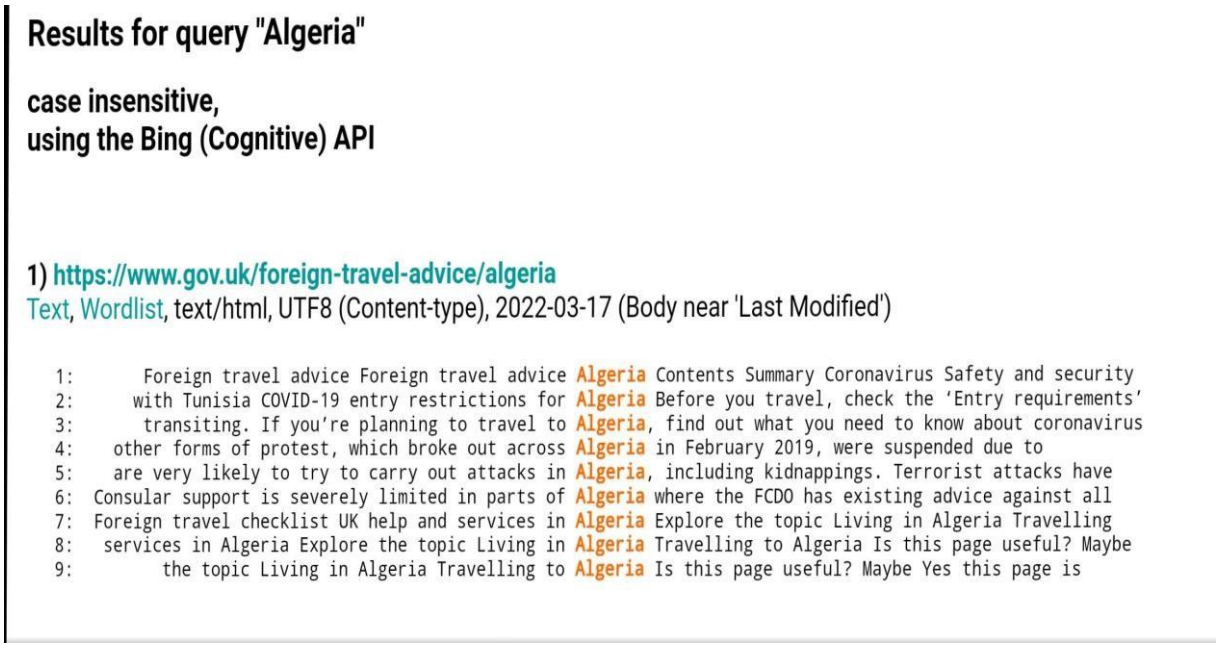


Figure 1: Selected Concordance Line from (uk.travel advice)

Selected concordance linesfrom (uk.travel advice) are concernedwithtwo major events: one is global and the other one is local. The first one is the pandemic therefore, concordance lines in which the wordAlgeriaoccurred are safety measures against the spread of corona virus. The following concordance lines focus on traveladviceconcerning the latestprotestthatbroke out in Algeria in the year of 2019.

2) https://www.bbc.com/news/topics/cp7r8vglgyet/algeria

Text,Wordlist,text/html, UTF8 (Content-type), 2021-04-08 (Body near 'Last Modified') :

Selected concordance lines from (Wikipedia) Shows the word Algeria occurrence in variouspolitical and geographical aspects that are known to be the primal features that defines Algeria as a country.

Figure 2: Selected Concordance line from (Wikipedia)

3) **https://www.britannica.com/place/Algeria**

Text,Wordlist,text/html, UTF8 (Content-type), 2022-03-15 (Body near 'Last Modified' :

Figure 3: Selected Concordance line from(cyclopediaBritannica)

Selected Concordance linesfrom (cyclopedia Britannica) shows varies occurrences of the wordAlgeria as expected in topics revolving around Historical events such as the French

conquest of Algeria and world war two. The focus then shifts on modern social media platforms and methods of transmitting information.

4)**https://www.bbc.co.uk/news/world-africa-14118852**

Text,Wordlist,text/html, UTF8 (HTML source), 2019-12-23 (Body near 'Last Modified')

Figure 4: Selected concordance lines from (BBC. News)

Selected concordance lines from (BBC. News) reveals that: Algeria is mentioned among various foreign and neighboring countries and it’s. Geographical importance. Is highlighted. Afterwards, the occurrence of the wordAlgeriaisstrictlyrelated to stabilityconcerns in the area. And reports about. ExtremismRelatedactivities.

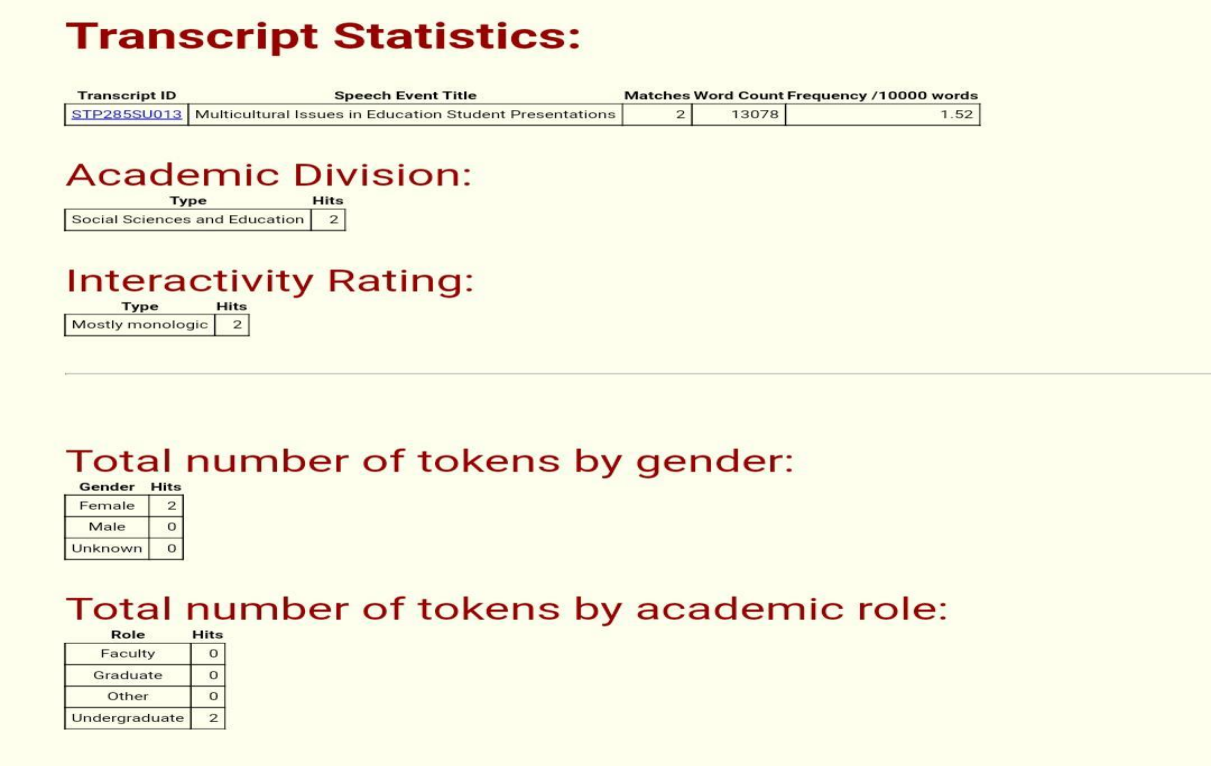
5)**https://www.cia.gov/the-world-factbook/countries/algeria/Text,Wordlist,text/html, UTF8 (HTML source), 2022-04-14 (Server header):**

Figure 5: Selected concordance lines from (cia.gov\the world fact book)

Selected concordance lines from (cia.gov\the world fact book) shows the occurrence of the wordAlgeriain multitude of seeminglyrelated topics. The French occupation of Algeria and itseffects, the establishment of the multi-party system in 1988 and finally the modern Algerian government 2020

4. Michigan corpus:

Michigan the Corpus of Academic Spoken English contains data from a variety of speech events and locations across universities (including lectures, classroom discussions, lab sections, seminars, and advising sessions). We look into the occurrence of the word Algeria in this corpus. The outcomes were as follows:



Transcript Statistics:

Transcript ID	Speech Event Title	Matches	Word Count	Frequency /10000 words
STP285SU013	Multicultural Issues in Education Student Presentations	2	13078	1.52

Academic Division:

Type	Hits
Social Sciences and Education	2

Interactivity Rating:

Type	Hits
Mostly monologic	2

Total number of tokens by gender:

Gender	Hits
Female	2
Male	0
Unknown	0

Total number of tokens by academic role:

Role	Hits
Faculty	0
Graduate	0
Other	0
Undergraduate	2

Figure 6: Transcript statistics of the appearance of the word Algeria in Michigan corpus

According to the statistics in the Michigan corpus of academic spoken English, Algeria is a low-rated country in academic divisions, with only (02) hits in Social science and education. Only two females were to have monologic interactive hits. We believe this is due to a lack of i nnovation in Algeria's educational system. This is due to a lack of appropriate capabilities and development means, despite the presence of Algerian intellectuals and thinkers in the educational field.

Conclusion

This chapter had discussed the presence of the word Algeria in the corpus. The obtained results were analyzed and discussed in order to make an appropriate assessment. According to the tables and analyses related to them, the word Algeria is rarely mentioned in English, particularly in academic field , for a variety of reasons, one of which being French

colonialism, which was a major role in Algeria's growth delay. We are still a developingcountry attempting to improve economically and scientifically.

Chapter five, General discussion

1.Discussion

The present work seek to analyze the use of "Algeria" in real world. The object of this research was clauses that were taken from British National Corpus, taking into consideration that the approach of this research is semantics, in which we should put in mind that by following this approach means the primary objective is disentangling of the meaning of individual words. In so doing, a word gets defined by relating its meaning to the meaning of other words, and sense is the result of recognizing the relationship between words.

According to the results the word Algeria varied from one type of text to another. It has the most common occurrence in the non-academic genre followed by MISC due to the historical background Algeria has with Morocco and France, where the results shows that France has occupied so much of space unlike other European countries, because Algeria was a colony of France in the old days and its effects as well as the establishment of the multi-party system in 1988 and finally the modern Algerian government 2020.

One of the direct results of the French occupation of Algeria was the low level of income and living for the vast majority of Algerians, so that huge numbers of them were deprived of enjoying public services, such as health and education, which were available to European immigrants,andinfact,alltheattentionofthecolonialadministrationwaslimitedtoproviding services to settlers until Even if it leads to neglect of the national education of the Algerian taxpayers.Consequently,Algeriadoesnotoccupysuchaspaceinacademicaspectsbecauseof the lack of materials, failure of the educationalsystem.

ThewordAlgeriaoccurrenceinvariouspoliticalandgeographicalaspectsthatareknownto be the primal features that defines Algeria as acountry.

In defining the meaning of words, semanticists tend to first locate an item within a general category. McEnery and Wilson remark that "psychological works on

categorizationshavesuggested that typically cognitive categories are not hard and fast so that it is not so much a question of whether or not a given item belongs in a given category as of how often it falls within that category as opposed to another" (McEnery and Wilson, 2001, p. 113). This has important applications in the link between semantic boundaries and corpus linguistics. The results seems to validate and isolate the meaning of Algeria in two aspects of (Geographical, political) by contribution of Wikipedia.

2. Advantages and Disadvantages of Corpus Linguistics

2.1. Corpus Linguistics Advantages

Today as a methodological tool CL is an established field in language studies and it provides data to studies and its value is appreciated in the field. However in addition to its benefits, CL also has some limitations. Firstly CL provides data for language studies by supplying huge amounts of examples about a specific structure and linguists can work out how a structure is used looking at these occurrences. In this way corpus is a really good tool for gathering data about language studies and it provides huge amounts of data which would take a lot of time with traditional data collection methods. Secondly, in corpus-led language studies, comments and evaluations about language use is not regarded as reliable evidence as they are totally subjective and it is difficult to test them empirically (Kennedy, 1998). In this way, empirical data is at the core of CL and unlike qualitative methods, it abstains from subjective comments or ideas in analysis. The data obtained from corpora is statistically analyzed and concrete evidence is used. Therefore CL depends on objective and empirical data and the data obtained from CL is scientific and it is open to objective verification. The third advantage of CL is that it provides holistic information about language structures. Thanks to statistical software used in CL, frequencies and statistical informationare obtained about certain structures in a study. The final advantage of CL is that corpora include essential information about the data it offers. Most of the modern corpora include information about the

socio-economic status, educational background, gender or place of birth of speakers. Moreover,corpora can provide data according to subject, field or genre. For instance, a researcher studying bilinguals can obtain data about bilingual speakers looking at their background or a sociolinguist studying sociolinguistic factors in language can obtain data by looking at the socio-economic status of the speakers in a 142 corpus. This is anothersignificant advantage of CL because researchers can get relevant data thanks to the information included about the speakers or subjects in a corpus.

2.2. Corpus Linguistics Disadvantages

As for the disadvantages of data analysis by CL, firstly CL is strictly quantitative and it depends on thedescription of language via numbers and statistics. Although holistic statistics can be valuable in language research, CL misses some of the most important components of language: the context and structure of the talk among speakers. In addition, CL disregards intuition as a tool for data gathering on the ground that it is too subjective and it is not testable. In this respect CL has a point but in language research competence as well as performance is significant. Finally, although CL focuses on large scale analysis and it is good at giving descriptive statistics, it is weak atoffering detailed analysis about individual users (Hunston, 2002). What matters the most for CL is what the 'whole' data suggests rather than what individual speakers do. This is of course a problem for CL because it cannot account for individual variation in language use. CL can give information based on the background of speakers and it can make generalizations about certain groups, though. However this is again at group level and it does not look at individual contexts.

3. The Implications

This study is important and unique in its context because it investigates the use of the word 'Algeria' in English according to the CL. Treating this topic using the BNC, MICASE, the Web Corpus to achieve the collocations of both spoken and written natural words and text

stored in this electronic databases. This study showed their purpose which is to construct a balanced and representative sample of current British English. The word ‘Algeria’ is not popular academically and the most occurrence in political and geographical aspects that are known to be the primal features that defines ‘Algeria’ as a country.

4. Recommendations

The study provides a significant insight on the word Algeria and its classification. The observation provided by the corpus goes as follows: There is a noticeable absence of the word Algeria in the academic classification because of the lack of contribution in the field of scientific research. On the other hand, the word Algeria has a prominent appearance in report matters linked to politics and history due to major historical events.

This imbalance can be worked on through prioritizing academic research and dedicating efforts and budget for the purpose of enriching Algeria’s scientific and academic contributions.

The General Conclusion

For all intents and purposes, corpus linguistics uses vast collections of both spoken and written usually natural words and texts, which are essentially preserved on computers. Corpus linguists may undoubtedly examine diverse concerns regarding language use which is pretty crucial by using a number of computer-based technologies via web data. Learners are an extremely powerful tool for analyzing fairly natural language, and they can provide tremendous insights into how language use varies in various situations, such as spoken and written issues represented by electronic data bases such as the British National Corpus (BNC), the Michigan Corpus (MICASE), and the Web Corpus. This research looks into the meaning of the word Algeria as it appears in the corpus.

The goal of this project is to examine how "Algeria" is used in the real world. The subject of this study was clauses from the British National Corpus, with the understanding that the approach used in this study is semantics, with the primary goal of disentangling the meaning of individual words. A word is defined by linking its meaning to the meanings of other words in this way, and sense is the outcome of understanding the relationship between words.

According to the findings, the word Algeria varies depending on the type of writing. Due to the historical background Algeria has with Morocco and France, where the results show that France has occupied so much space unlike other European countries, because Algeria was a colony of France in the past, and its effects, as well as the establishment of the multi-party system in 1988 and finally the modern Algerian government in 2020, it has the most common occurrence in the non-academic genre, followed by MISC. One of the direct consequences of the French occupation of Algeria was the low level of income and living for the vast majority of Algerians, resulting in a large number of them being denied access to public services such as health and education that were available to European immigrants, and in fact, the colonial

administration's attention was focused solely on providing services to settlers until Even if this leads to the neglect of Algerian taxpayers' national education. As a result of a lack of materials and a failing educational system, Algeria does not occupy such a position in academic terms. The word Algeria appears in a variety of political and geographical contexts that are considered to be the fundamental characteristics that characterize Algeria as a country.

When determining the meaning of words, semanticists usually start by placing the object in a broad category. "Psychological works on categorizations have suggested that typically cognitive categories are not hard and fast, so it is not so much a question of whether or not a given item belongs in a given category as it is of how often it falls within that category as opposed to another," McEnery and Wilson write (McEnery and Wilson, 2001, p. 113). This is useful in the connection between semantic boundaries and corpus linguistics. The findings appear to corroborate and isolate Algeria's meaning in two aspects of (geographical political) by contribution of Wikipedia.

REFERENCES

Allan, K. (2013). *The oxford handbook of the history of linguistics.* Oxford : Oxford University Press.

Andor, j. (2004). The master and his preformance. (N. Chomsky, Interviewer)

Atar, C., & Erdem, C. (2019). *The advantages and disadvantges of corpus linguistics and conversation analysis in second language studies.* Retrieved from Online submission.

Burnard. (2009, January). *British national corpus.* Retrieved from Oxford Text Archive : http://www.natcorp.ox.ac.uk/corpus/

Burnard, L., & Aston, G. (1998). *The BNC handbook* . Edinburgh: Edinburgh University Press.

Castello, D. (2014). *A corpus Study of strong and powerful*. Retrieved from Academia: http://www.academia.edu

Cherief, H. (2019, November). *The essential contributions of corpra in language research.* doi:10.24819/netsol2019.08

Devitt, M. (2020). Instuitions in linguistics. *The British Journal for the Philosophy of Science*.

Guy, C. (2003). *Applied linguistics.* new york: Oxford University Press.

Hunston, S. (2002). *Corpra in applied linguistics.* Cambridge: Cambridge University Press.

Kennedy, G. (1998). *An introduction to corpus linguistics.* London: Pearson Education Llimited.

McArthur, T. (1992). *The oxford companion to the english.* New york: Oxford University Press.

Michigan corpus of academic spoken English (MICASE) data base. (n.d.). Retrieved from Michigan corpus of academic spoken English: https://quod.lib.umich.edu/m/miscase/

Römer, U. (2010, August). *Applying corpus methods to writing: Exploration of MICUSP.* doi:10.17239/jowr-2010.02.02.2

Sinclair, J. (1991). *Corpus concrodance collocation oxford.* New york: Oxford University Press.

The british national corpus (BNS) data base. (n.d.). Retrieved from The british national corpus : https://www.english-corpora.org/bnc/

Thompson, P., & Nesi, H. (2001). *Language teaching research* . Birmingham: Birmingham University.