People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
University of Ibn Khaldoun Tiaret
Faculty of Letters and Languages
Department of English

A CORPUS LINGUISTIC STUDY ON BRIDGING GENERATION LINGUISTIC DIVIDES IN TIARET'S CULTURAL LANDSCAPE

A Dissertation Submitted to the Department of English in Partial

Fulfilment of the Requirements for the Degree of Master in Linguistics

Submitted by:
Mohammed KADARI
Hind Karima ALLALI

Supervised by:
Dr. Moulai Hacene Yacine

Board of Examiners

Dr. Benamor Youcef	**Chairperson**	**University of Ibn Khaldoun Tiaret**
Dr. Moulai Hacene Yacine	**Supervisor**	**University of Ibn Khaldoun Tiaret**
Dr. Bouguessa Amina	**Examiner**	**University of Ibn Khaldoun Tiaret**

Academic Year 2023/2024

Dedication

To our parents , with deep gratitude for your constant encouragement and unwavering support we dedicate this work to you. Aiming to make you proud and to express our sincere appreciation.

To our friends with no exception

To the memory of my deceased best friend

&

To our supervisor Dr. Moulai Hacene, who has been incredibly supportive throughout this journey, offering valuable guidance and patience. His assistance has been crucial in navigating research challenges and refining our ideas. We are truly grateful for his unwavering support

Acknowledgements

We extend our heartfelt gratitude to every individual who stood by us, supported us, believed in us and contributed in any way during the challenging months leading up to this moment. Your support has been invaluable and deeply appreciated.

Primarily,We express our gratitude to our esteemed supervisor ‘Dr. Moulai Hacene’, whose consistent support, patient guidance, and insightful feedback have been indispensable throughout the creation of this work. Thank you for always believing in our capacities and being our mentor during this experience.

Moreover , an exceptional acknowledgement to the board of examiners, namely Dr. Benamor Youcef and Dr. Bouguessa Amina for agreeing to participate in the examination process of this work and dedicating their valuable time and expertise to review and asses it.

For last , our deepest appreciation is addressed to all the teachers whose dedication and guidance have shaped our academic journey over the past five years.

Abstract

This research paper aims to document and preserve the linguistic features of the speech variety of Tiaret. Due to many social factors, the linguistic heritage of Tiaret is progressively diminishing over time. This necessitates immediate preservation efforts. Moreover, in an effort to preserve the language, we adopted a corpus linguistic approach , in which our primary source of data were elderly inhabitants of tiaret and the surrounding areas , aged 55 to 75 chosen for their rich linguistic heritage. Furthermore , we conducted unstructured interviews, carefully documenting their words and expressions. In this respect, the gathered data was then carefully organised, with explanations and translations provided , preparing it for thorough analysis and digitization by a professional for the compilation of the corpus database .The research findings highlighted significant linguistic differences between the older and younger generations. However, our efforts encountered a major setback during the final step of digitization and corpus compilation. It became apparent that current corpus linguistic tools and software were inadequate for accommodating our linguistic variation. Despite this setback, we view it not as a dead end but as an opportunity for further exploration in this specific area of study. Finally , this setback serves as an invitation for continued research in corpus linguistics, aspiring to develop tools that can effectively capture and preserve diverse linguistic varieties such as that of Tiaret.

List of Tables

List of Figures

Table of Contents

CHAPTER TWO: CORPUS LINGUISTICS DATABASES AND FRAMEWORK

CHAPTER THREE : SCOPE AND METHODOLOGY

GENERAL INTRODUCTION

Throughout history , humans have always sought ways to preserve language. Early on, it started with cave walls and other surfaces.These methods evolved over time into more modern techniques like writings and digital technologies.One modern approach, corpus linguistics, significantly contributes to language preservation efforts by systematically analysing and documenting linguistic data.

In this research, we are concerned with the field corpus linguistics as a tool of language preservation. Corpus linguistics has been utilised in various ways to achieve diverse objectives. However, one of the indirect outcomes of these studies is always the preservation of language for future inquiry. The topic is relatively unexplored, as corpus linguistics, research has not been adopted to its full potential in Algeria.

This study aims to explore linguistic differences between the older and younger generation, identifying any gaps that may exist. Additionally, it seeks to propose methods to bridge these intergenerational linguistic disparities.

The main reason for selecting this topic was its underutilization in Algeria as a preservation method, coupled with our desire as members of the Tiaretian community to contribute to it based on our area of study. We initiate our research by asking the following question:

- How can Corpus Linguistics effectively contribute to preserving the language of Tiaret, and what are challenges and opportunities that arise when compiling the corpus?

In order to delve into the aforementioned research question, we've developed a set of sub-questions to thoroughly investigate various aspects of our inquiry. The sub research questions are as follows:

1. How can the data collected from the older generation be effectively organized and analysed to create a corpus that accurately represents the language?
2. What are the potential applications of the corpus for language preservation efforts in Tiaret?

In light of the previous questions, we have made the following hypotheses:

1. The potential challenge in the progress of this research is the complexity associated with the process of compiling the corpus.
2. The utilisation of corpus linguistics in language preservation efforts will reveal and show generational shifts and changes in language usage within the older and younger Tiaretian community.
3. The compilation of a corpus database of Tiaretian language will facilitate the analysis of linguistic trends and patterns, offering valuable insights into language usage and evolution within the Tiaretian community.
4. Significant language loss within the Tiaretian community may be revealed, highlighting the impact of external influences and socio-cultural changes on the community's language heritage.

In order to validate or invalidate the above mentioned hypotheses , the dissertation is split into two parts , theoretical and practical. The theoretical will be conducted in the first chapter and the second chapter. The practical part will be included in the third chapter.

Chapter one entitled “The Scope of Corpus Linguistics” offers an introduction to corpus linguistics, presenting essential key terms and terminologies crucial to the understanding of the field. It delves into the historical background, tracing the evolution of corpus linguistics over time and acknowledging the contributions of individuals who have been a part of its development. Moreover, the chapter underscores the significance of corpus linguistics by emphasising its importance in linguistic research. Additionally, it explores related fields that intersect providing a broader context for understanding its applications and implications.

The second chapter , entitled “Linguistic Databases and Framework“ We begin with a description of recognized databases resulting from the application of Corpus Linguistics ,showcasing their linguistic contributions to the field of linguistic research. Following , this we dive into key theories that complement the analytical approaches used in studying language patterns with the corpora. We will also introduce prerequisites ne required for effective implementation of Corpus linguistic methodology. The limitations of this field will be addressed alongside its application setting the stage for a critical evaluation.

The third chapter , entitled “Scope and Methodology “ We transition to a more practical focus detailing the research methodology employed in conducting the study we discuss the approach to data collection which involves gathering information from participants , specifically and exclusively from the older generation of the Tiaretian community. The data collected served as a valuable source to the compilation of the corpus. We then proceed to outline the data analysis including the methods used to analyse the collected data throughout this chapter. We highlight the challenges encountered during the creation of the corpus. Ultimately we conclude by reflecting on the journey from data collection to Corpus creation acknowledging the challenges faced in order to finalise our findings.

CHAPTER ONE: THE SCOPE OF CORPUS LINGUISTICS

Introduction

Language is the medium through which humans socialize. As easy and simple as it may seem to some people, it is a rather complex task. Linguistics, the scientific study of language, seeks to reveal what language is and how it is actually used through various ways of linguistic analysis. Corpus linguistics (CL) is a means through which linguists investigate and analyse the human language. In this regard, the current chapter serves as a preamble to corpus linguistics, consequently, it will provide a general understanding of corpus linguistics. To begin with, the researchers presents some key terms and terminologies; this is to acquire readers the fundamental concepts that may be used frequently throughout this study. After that, a brief historical background of CL is provided for the purpose of getting the reader to know how it has evolved through time. Then, we move to the significance and importance of CL aiming to highlight its major contributions in linguistic enquiry. Furthermore, we will tackle some related fileds and domains to show how it is utilized. Also, it will shed light on some of the most known theories to delve deeper in the use and application of CL. At last, it will deal with use of corpus linguistics in Arabic language and it's varieties which will lay the foundation for the second chapter.

I. KEY TERMS AND TERMONOLOGIES

Before diving into the exploration of corpus linguistics, Understanding the dynamic character of language studies in the humanities is crucial. In contrast to the natural sciences, which frequently need empirical precision, the humanities provide a more flexible and adaptive method of comprehending social and human events. Diverse approaches have been embraced

by academics in various domains, enabling sophisticated studies of language and culture. The foundation for corpus linguistics, a methodology that uses real-world language data for in-depth study and theoretical advancement, is laid by this interdisciplinary viewpoint.

1. Corpus Linguistics

Unlike natural sciences, which are characterized by exactness and rigidity, humanities on the other hand are characterized by a kind of fluidity and flexibility which has given the chance for scholars in human and social sciences to approach topics differently. Corpus linguistics has been defined by multiple scholars since it has emerged.

The central idea of corpus linguistics is studying language using real life examples, that is to say the linguist brings pieces of texts from a given language and submits them to linguistic analysis for the sake of language description and to come up with theories on language (McEnery &Wilson, 2001, p.3). what "McEnery and Wislon " mean by "pices of texts" is corpus, in which the origins of the term corpus could be traced back to late 14th century, it was used to refer to "matter of any kind", literally "body",(plural corpora) originating from the Latin word "corpus" meaning "body"(see corporeal). The term evolved to denote the "body of a person" (mid-15th century in English) and "collection of facts or things"(1727in English),reflecting its diverse meanings rooted in Latin(Etymology Online, 2024). Additionally, according to Oxfford dictionary, corpus /ˈkɔːpəs/ (British English), /ˈkɔːrpəs/ (American English) with plural corpora /ˈkɔːpərə/ (British English), /ˈkɔːrpərə/ (American English) refers to "a collection of spoken or written texts" which our main concern in this context. Moreover, according to Bennet (2010) corpus linguistics uses corpora- a term that refers to huge, organized collections of electronically-stored natural language samples-to approach the study of language in use. In addition, the primary focus of corpus linguistics is on

the linguistic patterns linked to lexical and grammatical aspects of language, as well as how these patterns differ throughout varieties and registers.Nonetheless, corpus linguistics, according to Bennet (2010), is unable to detect what is grammatical and what is not in a language nor can it justify the absence of a manner or a linguistic item, but rather can tell what is or is not present in a corpus and that if a manner or a linguistic item is not used in corpus, then it simply means that they may not be quite common in that language. Also, corpus linguistics' main concern is not to explain why things are the way they but only to show what they are; introspection, however explains the why. Besides, corpus linguistics regardless of the huge size of the texts provided in a corpus cannot describe all the language and a corpus remains but a sample of it (Bennet,2010).

Kennedy (1998) is of the same mind of the above mentionned scholar where he views that corpus linguistics is a research field in which texts are major source of evidence providence for linguistic description and argumentation. This means that corpus linguistics is an empirical research domain that uses bodies of texts to investigate a given language.

By analysing these large collection of language texts "corpus" the linguists aim to provide a detailed analysis of all language properties ; this has allowed linguists from other research areas a greater chance for best and quick results such as the descriptive study of language, language education, and lexicography.

In principle, This approach seeks to thoroughly analyse any significant quantity of electronic data regarding actual language use, including sets of literary and none literary text samples to consider a language's synchronic and diachronic features. Furthermore, the ways in which corpus linguistics uses contemporary computer technology to gather language data, process language databases, information and data retrieval, and application of these diverse language-related research and development activities are what make it unique and favoured by

scholars from 1950s to this day. After the advent of electronic corpus, the fundamental tenets of corpus linguistics has become 1-our cognitive need to understand language use in everyday conversation and 2- the possibility of developing intelligent systems that have the potential to establish an effective communication with humans.On top of that , computer scientists and linguists have worked together to compile a language corpus for the benefit of the language community at large; one that can only be used to design intelligent systems (such as computer-aided instruction, speech recognition, machine translation, language processing, , and text analysis and comprehension technologies), but it will also support the preservation of people's cultural identities and endangered languages and traditions (McEnery ,2008).

Finally, the information gained by corpus analysis is applicable to all branches of language technology and linguistics. As a result, the description and analysis of linguistic aspects from a corpus has become crucial in all areas of human knowledge (Dash, 2010).

2. Fundamental Concepts in Corpus Linguistics

In order for the readers to delve into the daomain of corpus linguistics, they must be familiar with the most frequently used fundamental concepts which are to be addressed in the subsequent section

A. Corpus

As previously mentioned, corpus is the essence of corpus linguistics as the name suggests. Its origin relates back to Latin corpus which means “body”. Currently, it refers to a collection of spoken or written texts that are used in linguistic research to highlight a specific language, dialect, or other subset of a dialect. A more accurate definition is provided by Dash (2010) where he states that “ In finer definition, it refers to (a) (loosely) anybody of text; (b) (most commonly) a body of machine-readable text; and (c) (more strictly) a finite collection of

machine-readable texts sampled to be representative of a language or variety" (McEnery and Wilson, as cited in Dash, 2010)

The corpus is made up of a substantial number of representative text samples that span a range of language types and are utilised in different contexts for linguistic interaction. In theory, The corpus is a potentially limited set of texts that can be used to represent ideas. It is computer-based, useful for study and application, representative of the target language, processable by both humans and machines, offers an infinite amount of data, and methodical in its development and representation (Dash, as cited in Dash,2010, p.1).

B. Corpus Features

In contemporary linguistic research, corpus has emerged as an invaluable tool for human language analysis and understanding. Like any other methodology or tools used in this process, corpus has some characteristics. In this regard, Dash (2010) has presented some of features which he conceives as being of paramount importance. First, the quantity of a corpus should be substantial so as it covers large amount of either spoken or written form of data. Its body is made up of all of its constituent parts, which together make up its size.

Second, the quality of the corpus needs to be taken into account, every text should be extracted from real-world speech and writing examples. Here, the linguist's function is crucial; this means that he has to confirm that linguistic data is gathered from real-world communication rather than from controlled or experimental settings.

Third, representation is one important feature as it takes a part from the corpus definition, a variety of text samples ought to be included. Since any further research that is developed on it would require the verification and authentication of data from the corpus of a language, it should be balanced to represent the maximum amount of linguistic diversity across all areas of

language use. Another element is simplicity, a corpus should have a straightforward language in an easy-to-read manner. This means that without any further linguistic information marked up inside texts, it is expected to have an uninterrupted stream of characters (or words). Any kind of annotation containing different kinds of linguistic and non-linguistic information is contrasted to a basic plain text.

Equality is also a crucial feature, in which the corpus' samples ought to be uniformly sized. This is a contentious matter, though, and it won't be accepted everywhere. To increase the representativeness and multidimensionality of a corpus, the sampling model may be modified significantly, Moreover, retreavability is as equally important as the quality feature, the end- users should be able to readily retrieve data, information, examples, and references from the corpus. This focuses on methods for maintaining computer-based language data in electronic format. Current technology enables and supports the creation of corpora on PCs and their preservation in a way that facilitates the easy retrieval of material when needed. Furthermore, any form of empirical verification should be able to access the corpus. In other words, data from corpora can be used for any type of verification; this puts the intuitive method of language study behind corpus linguistics. Also, consistent increases in corpus are necessary. The corpus will then be 'at par' to record linguistic changes that transpire throughout time in a language. A corpus gains historical dimension for diachronic studies and for exhibiting language cues to arrest changes in life and society over time by the addition of fresh linguistic data. Last but not least, documentation is also a salient feature a corpus needs to have. The text itself should be maintained apart from all component information. It is usually preferable to have a short header with a reference to the documentation and to keep documentation material apart from the text. With minimal programming work, this enables the efficient separation of plain texts from annotation in the context of corpus management.

C. Types of corpora

It is of paramount importance for the readers to be aquainted with some of the corpus types to be able to understand its design and compilation criteria. Some of its types are addressed in the subsequent section.

1) Specialized Language Corpora

In theory, special corpora differ from a corpus that includes a diversity of natural, real language. Neither a conversation corpus nor a text corpus from a single newspaper qualify as unique corpuses. A distinction is drawn between varieties that, for whatever reason, depart from the general core and variations that fall within the bounds of what is fair to expect from the type of language that a significant number of native speakers use on a daily basis. The special corpora are those that don't add anything to the description of ordinary language, either because they are largely composed of uncommon traits or because their sources cannot be trusted as accounts of typical behavior (EAGLES 1994). Their use derives from the belief that specialized language will provide a more ordered, if constrained, representation of the idiosyncrasies of real-world usage within the framework of an approach that seeks to standardize and simplify the chaotic nature of natural language. From this angle, a corpus with a greater diversity of texts could make training these kinds of tools more difficult. In contrast, the unique corpus can be utilized as a training corpus for annotation**.**

2) General Purpose Corpora

In this regard, it is crucial to keep in mind that corpus work should always be comparative, meaning that data from a general purpose corpus should be compared with data from a specific-domain corpus, regardless of whether the focus is on LSP, translation, the learning process itself, or another topic. Because the corpus may provide students with an approximate

representation of first-hand language experience, this is especially crucial for teaching language to students. Obviously this also raises the issue of compiling general-purpose corpora that are thought to be representative of the language in its whole. The core of corpus linguistics is the problem of corpus representativeness (Biber 1990, 1994). When putting together a corpus, factors like sampling and target population definition become crucial. The majority of these broad corpora are too big for one person to handle because they are 300 million words or more in size and growing every day. On CD-ROM, though, or through remote access they are typically made available.

3) Learner Corpus

A learner corpus is a different kind of corpus that focuses only on the teaching process and it is very helpful for error analysis (Granger 1994, 1996, 1997). This can be applied to recognize recurring themes in student writing. Fan et al. (1999) go through the primary advantages of building and using a corpus of language learners. They point out that it can be used to identify characteristic errors made by a single student participating in a particular activity and is a helpful diagnostic tool for educators and students alike. It provides the option to select the text type and/or subject area. Where mistakes are most likely to happen, allowing the instructor to pre-teach and so avoid common mistakes.

It promotes and strengthens the movement toward learner autonomy and, with the right support, empowers language learners to take on the role of language researchers acquire the abilities needed to recognize, clarify, and correct repeated mistakes. Learning activities and resources can be derived from a learner corpus. With access to extensive databases, newspaper collections, and other resources, students are able to evaluate and contrast usage patterns between native and non-native speakers across various corpora.

4) Language for Specific Purposes LSP

In the field of LSP, a large number of restricted variety corpora are being assembled for educational purposes in direct alignment with the trend towards vocational language training, in which economics students, for example, are exposed to the language of economic texts, journals, and/or spoken negotiations, preferentially. These days, getting access to a newspaper or magazine's CD-ROM is rather simple. With relatively little work on the part of the individual scholar or instructor, a collection of these, for example, of the Economist or single 24 hours that includes all published material within a specified time frame or according to a particular theme is a significant resource.

When it comes to LSP, it is crucial to identify certain terms. It has been demonstrated that, in this instance as well as in other domains, the formalization of contextual patterning can be quite beneficial for meaning recognition. For instance, Pearson (1998) contends that by concentrating on the meta-linguistic patterns that are typical of particular kinds of specialized material, a domain-specific corpus can provide a way to identify semi-technical terminology. Formally, these patterns may be recognized in relation to these terms, and they can greatly assist the student in creating terminological definitions.

II. Historical Background

With that being said, we can trace corpus linguistics history into two crucial eras, the first one being the pre electronic era and the second being the post-electronic one. The field of corpus linguistics has a lengthy history where in its earliest days of emergence, it was not considered as reliable and practical; however, with the appearance of computers it had reclaimed its usability as a more efficient tool.

1. Pre-Electronic Corpus Linguistics

Early corpus linguistics was the term given to corpus linguistics by McEnery and Wilson (2001) prior to the contributions of Noam Chomsky, meaning the time of the late 1950s. Although many of the linguists at that time, such as Bloomfield, have used this method in their work, it was only labelled as corpus analysis in a much later time. In its early stages, corpus linguistics was predominantly focused on the study of the English grammar that is according to (McEnery and Wilson) 2001. Yet it was not only limited to the English Language as it was used by many other languages. We will encounter illustrative examples of this within this chapter.

Its beginnings was with a large collection of utterances of language (Harris) 1993, it is hardly astonishing that linguists have always shown interest in how people use language thus, research predates computers and everything electronics.

Examples of work done that was corpus based and manually collected at that mentioned in (McEnery and Wilson, 2001.p.03-04)

- *In Language Acquisition:*

- Studies of first language acquisition based on parental diaries (stern 1924)

Corpora of children's utterances (McCarthy 1954).

- *Syntax and semantics:*

- Descriptive grammar of English (Fries 1952).

- *Foreign Language (spelling) :*

- Corpus of German to study frequency distribution and sequences of letters in German (Käding 1897).

Despite the very widely common use of corpus linguistics by many linguists, it was highly criticized and it has endured a period of diminished popularity due to some of the constraints it has encountered (McEnery and Wilson) 2001.

Over time, Corpus Linguistics has evolved and reclaimed its popularity in the domain owing to many factors, one of which being the emergence of computers.

2. Post-Electronic Corpus Linguistics

In the words of (Wonner, 2001.p.04)), while describing the evolution of corpus linguistics with a set of adjectives, from impossible, to marginally possible but lunatic, to quite possible but still lunatic, to now, very possible and popular. with the development of technology and computers corpus linguistics became more accessible and less time consuming even when using a huge amount/ collection of texts , as the advancements in the technological field have facilitated the whole collection ,maintenance and analysis process all the while enabling researchers to dive into language patterns with greater depth and increased efficiency.(Kenndy,1998 p.05)

Corpus linguistics has been considered as a complex field because it depends on specialised tools and Technologies, but thanks to modern electronics, all of these tools can now be used on a single computer making the field more accessible and easier to handle. This means that data collection storage and Analysis have all become much simpler and more efficient. (Kennedy) 1998.

III. SIGNIFICANCE AND IMPORTANCE

As a result of this revolution in corpus linguistics, many electronic corpuses took place including the brown corpus which was a huge contribution to not only the field of linguistics as

a whole but to many other subfields, saying that the Brown corpus conducted by Francis and Kűcera (1964) merely shaped the modern corpus studies would be an understatement (Kennedy, 2001). Computers have given corpus linguistics a huge boost by reducing much of the text-based linguistic description and vastly increasing the size of the databases used for analysis (Kennedy, 1998.p.23).

Furthermore, Even if someone has not personally engaged with a corpus, they have likely relied on dictionaries and grammar books built upon data derived from corpora, particularly if English is not their native language; this was the view of (Hamzaoui, 2020.p.55).

In general the study of corpus linguistics offers an basis, for exploring different elements of language structure, usage and development. This research greatly contributes to theory and its practical applications, in linguistics and related areas. (from Bseln.com).

I.V. MOST KNOWN THEORISTS IN CORPUS LINGUISTICS

There were many influential figures in the development of corpus linguistics, some contributions were direct and others indirect.

1. John Sinclair (1960-1970)

During 1960 and 1970, a number of researchers, including John Sinclair and Geoffrey Leech, who are now seen as the field's greatest representatives, carried on the work that Noam Chomsky started.

John McHardy Sinclair was a first-generation modern corpus linguist who founded the COBUILD project, which aimed to build corpus-driven lexicons for foreign learners of English. This project was a partnership between the University of Birmingham and the Collins publishing house, more known as the COBUILD English Dictionary (Sinclair) 1987. He revolutionized lexicography and took it to a whole other level in the 1980s by proposing a new

Kind of dictionary for advanced learners of English. (Lavid, 2007.p.09). Robin Fawcett says in this respect: “I remember that, when I first turned the pages of my copy, I realized that it would be a research tool that I had to ensure I had by my desk for the rest of my working “ as cited in (Lavid,2007.p.10) , this praise was very much shared by many as the COBUILD project emphasis on corpus based analysis and its practical application in language teaching and learning made it a valuable treasure of a source for many in the field. According to the Collins Cobuild (1996), the techniques used to compile the dictionary are new and use advanced computer technology. For the user, the kind, the quality and the presentation of information is different.

2. M.A.K. Halliday (1960)

Professionally known as M.A.K. Halliday, after his full name: Michael Alexander Kirkwood Halliday is the father of Systemic Functional Linguistics.

In contrast to Chomsky’s views, Halliday believed in the functionality of language and its study in a social context. He believed that language is not necessarily wired into our brains but more so a way that we can express ourselves with saying that language is a function in the brain, Chomsky sees that this function is built in and innate, and an instinct that pre-shapes our understanding of language, with a set of formal rules. Holliday on the other hand considers a function as one that is mouldable by us; we can shape and use it. (Jiří Lukl, 2019.p.99).

If someone were to inquire about the purpose or reason behind humans having or learning language, the most common response would probably be "to communicate “, the ultimate aim of acquiring a language isn't merely about mastering correct linguistic structures and forms. It is about effectively communicating meaning and achieving specific social objectives through language (Endarto, 2017.p.01-02).

Systemic functional Linguistics is a theory founded by halliday in the 1960s with a functional and semantic orientation. This Theory works to examine and show how language is used to fulfil certain social purposes and how these purposes vary according to many factors including individuals and social contexts. This theoretical framework helped provide an orientation towards Corpus linguistics, recognizing it as a reliable tool to be used in similar works. (Study Smarter UK website)

3. Anthony McEnery (1980)

As described in the Lancaster University official website (2024) Tony McEnery is a Professor in English Language and Linguistics in the University of Lancaster and an Adjunct Professor in the University of Limerick. He is a leading figure In Corpus Linguistics research , as he has significantly advanced the field of corpus linguistics and its applications in linguistic and related fields by working on projects that examine language change And providing valuable insights Into the evolution of language , He has Authored and co-authored numerous books, edited books, Journal articles and research papers Related to Corpus Linguistics and the application of corpora , two of which are (" Corpus linguistics : and introduction " : With Andrew Hardie) , ("Corpus linguistics : method, Theory , and practice " (with Andrew Hardie and Wendy Anderson).

"I am still active in teaching. Each summer I teach on a summer school in corpus linguistics at Lancaster University. I also run a MOOC on the Futurelearn platform, through which I have introduced corpus linguistics to tens of thousands of people." (Tony McEnery, Lancaster University 2024)

To this day, he still has an active role in CL.

V. REACTION TO CORPUS

Corpus linguistics has been a subject to varying definitions and interpretations within the linguistics community over the past years, it has undergone a period of repulsion before it became all about data and computation. Many linguists have contributed to the development of corpus linguistics. While Noam Chomsky is not necessarily considered as a pioneer of corpus linguistics as he argued against the use of corpora, his criticism towards this last significantly elevates its existence in the field of linguistics within scholars (McEnery and Wilson,2001.p.5-12). When asked about corpus linguistics Chomsky have always had reservations about it. He believed that studying the rules of was more valuable than studying how people use language in practice. He rejected it with a statement given in an interview: "corpus linguistics doesn't mean anything. It's like saying suppose a physicist decides, suppose physics and chemistry decide that instead of relying on experiments, what they're going to do is take videotapes of things happening in the world and they'll collect huge videotapes of everything that's happening and from that maybe they'll come up with some generalizations or insights. Well, you know, science does not do this "(Andor, 2004.p.97) as cited from Guillaume Desagulier's article on corpus linguistics (2017), in simple words, Chomsky emphasized on the importance of theoretical linguistics rather than empirical data analysis.

Although Noam Chomsky has not directly contributed to corpus linguistics, his work has reopened debates and sparked discussions about theory and data in linguistics. Researchers started to search to find ways on how to include chomskyan ideas, in other words they reached to find a way using both theoretical and empirical methods. Therefore, McEnery and Wilson (2001) his criticism ended up being more constructive than anything is, even if it had not been perceived as such at the time being. It should be noted that Chomsky's criticism of the utility of corpus linguistics is an ongoing one as his recent interviews have shown. (One of which "The

Galilean Challenge: Architecture and Evolution of Language" On November 24, 2016 at Paris University).

V.I. ARABIC CORPUS LINGUISTICS AS AN EMERGING FIELD

Arabic, a language as rich in history as it is in complexity, stands as one of the most influential and widely spoken languages in the world. With a legacy spanning millennia. Arabic is not just a way to communicate; it is also a big part of culture and religion for many people. Originally, Arabic started in the Arabian Peninsula, where it was spoken by ancient Arab tribes. As time passed, it spread to other areas, especially after the rise of Islam in the 7th century. Nowadays Arabic is spoken by but by hundreds of millions of people across northern Africa and western Asia, and more broadly around the world. Despite being one of the six official languages of the UN and a major language spoken natively by almost 300 million people, Arabic has not been a primary focus of corpus-based study, either in the early days of corpus linguistics or in the later flourishing of multilingual work conducted using corpora. This is unexpected, particularly considering that the Arabic grammatical tradition is rooted on what Ditters (1990, P. 130) refers to as a type of corpus linguistics.

The most notable is that the Persian linguist Sībawayh (c. 760–96 ce; see Carter 2004) built the earliest written Arabic grammar, known as al-Kitāb, using attested language. Similarly, al-Ḫalīl, the tutor of Sībawayh, wrote al-Kitāb al-ʿAyn, or "The Book of (the letter) ʿayn," which is the first Arabic dictionary. According to Brustad (2016: 148–9), the Arabic corpus included "formal speeches, tribal combat (ayyām) material, and pre-Islamic poetry, which grammarians and others refer to as "kalām al-ʿarab." (literally, "talk of the Arabs"); Ditters and Brustad argue on whether daily speech was included in this corpus. For example, this made it possible for early Arabic grammarians to recognize characteristics common to Bedouin usage (Ditters 1990, P.129). There are moments when the terminology used by al-Ḫalīl and Sībawayh

to reference evidence from the corpus are remarkably contemporary, based on Brustad (2016, P. 150), Al-Ḫalīl expresses opinions regarding the inclusion of specific phrases and patterns as "part of the corpus" (min kalām al-ʿarab). For several of the structures he discusses, Sībawayh provides frequency judgments using terms like "this occurs more than I can describe for you in the corpus" and "this rarely/often exists in the corpus." Both scholars have repeatedly said that a specific phrase or construction is either absent from the kalām al-ʿarab or the [ʿArabiyyah, or it occurs either infrequently or frequently.

The Arabic language tradition's natural propensity for corpus-based analysis may indicate that Arabic speakers should embrace contemporary corpus linguistic approaches with enthusiasm and speed. Even yet, acceptance remained somewhat muted even after the technical difficulties associated with computerizing Arabic script were resolved between 1990 and 1995. Compared to languages without historical figures similar to Sībawayh, Arabic has seen comparatively less corpus linguistic work done on it. A search on Google Scholar for 'arabic corpus' in twentieth-century publications yielded 90 results, while 'english corpus' returned 1,940 papers, indicating a significant disparity. Even in the twenty-first century, 'english corpus' returned 15,600 results, while 'arabic corpus' yielded only 1,800. Though this informal test has its limitations, it underscores the underrepresentation of Arabic corpus linguistics relative to its linguistic and cultural significance. Rather than viewing Arabic's underrepresentation as a deficiency, it can be seen as an opportunity for the field's emergence. There are promising signs of momentum and prominence within Arabic linguistics, such as the presence of Arabic corpus linguistics papers in academic symposiums. The vitality of this emerging field is evident from the number of results found on Google Scholar.

The persistent lack of focus on Arabic in corpus-based studies raises questions about its implications. Ultimately, attested language use serves as a reliable guide for linguists,

mitigating biases inherent in intuition-based analyses. An illustrative anecdote from the debate between Sībawayh and al-Kisaʾi highlights the importance of attested language data in linguistic analysis, demonstrating how biases can misguide linguistic interpretations. Corpus analyses offer a window into real language usage, providing relatively objective insights and resolving linguistic debates effectively.

Conclusion

To summarize what has been dealt with, this chapter was to establish the foundational principles of corpus linguistics. It may come across as surprising but a considerable amount portion of linguistics students lack awareness about the field , despite its well-known contributions to the wide scope of linguistics and its inclusivity to all languages around the globe.Now , as previously discussed corpus linguistics initially has a shared definition amongst linguists , but the same cannot be said about its interpretation , as many do hold contracting viewpoints and scepticism about it as a field of reliability .Moreover , many key points have been revealed in corpus linguistics and its development that was highly influenced by linguists alongside technology and the rising of it. not only is the term a controversial one but also a multidisciplinary one , so much can be said about it and so much can be learnt in the process of trying to get familiar with what a corpus undergoes for it to be constructed. More will be provided in the next chapter in regards of the understanding of this las

CHAPTER TWO: CORPUS LINGUISTICS DATABASES AND FRAMEWORKS

Introduction

In this chapter, we will delve deeper into the field of corpus linguistics exploring the usability of corpus. We will first start by presenting some of the major corpus databases, the corpus linguists data sources, and see how they contribute to linguistics research. Then, we discuss the prerequisites for using corpus linguistics including theoretical knowledge, technical skills, tools, and software required for corpus analysis. This is to help readers track essential tools and methodologies needed to work with large-scale linguistic data. Next, we delve into the application of corpus linguistics in other disciplines, highlighting the diverse ways in which corpus can be used in domains such as pragmatics , semantics, sociolinguistics and how it contributes to the study of language in real-life settings. Despite its many advantages, corpus linguistics faces certain limitations and challenges in which we aim to include in this section. This is to help researchers navigate the complexities of corpus-based studies.

I. MAJOR CORPUS LINGUISTICS DTABASES

In the field of linguistic research Corpus databases serve as essential resources. Playing a crucial role in advancing our understanding of language these databases help in exploring linguistic phenomena from various angles by providing access to rich and diverse collections of language data.

1. The British National Corpus

The British national corpus is regarded as one of the largest corpora, with over 100 million words collected from both spoken and written samples that represent British English from the later half of the 20th century. The BNC Consortium, an academic and industrial consortium headed by Oxford University Press, started building the corpus in 1991 and finished it in 1994 (the BNC official website).

Since the project's completion, two sub-corpora containing content from the BNC have been made available: the BNC Sampler, which is a general collection of one million spoken and one million written words, and the BNC Baby, which is made up of four one-million word samples from four different genres. (The BNC official website).

10% of the BNC is spoken language, which consists of informal, spontaneous conversations, and the remaining 90% is written language, which consists of books, periodicals, as well as some written to be spoken, like political speeches (The BNC official website).

Despite being regarded as monolingual some non-British English and foreign language words do occur in the corpus. It is a synchronic corpus and not a diachronic one , meaning language used from the late 20s is represented as previously stated , and not limited by any subject field in particular, it includes a variety of different styles. (the BNC official website).

2. The Corpus of Contemporary American English

Developed in 2008 by Brigham Young University's retired corpus linguistics professor Mark Davies, The only comprehensive and well-balanced corpus of American English is the Corpus of Contemporary American English (COCA), which has over one billion words of text in it. It contains a wealth of information spanning the last 30 years, with 20 million words

added annually between 1990 and 2019 (all years maintaining the same genre balance). (The official COCA website)

Properly distributing across a wide range of genres, including spoken word, fiction, popular magazines, newspapers, and scholarly journals , it allows users to browse through offering powerful search features which are as follows: (from the official website)

- Searching for phrases and strings with optimized speed.
- Browsing a frequency list of the top 60,000 words
- Access the Academic Vocabulary List (AVL).
- Searching for individual words
- Analysing entire texts by inputting them into COCA for detailed word and phrase information.
- browsing through randomly-selected "Words of the Day,"

COCA is available at English corpora.org, which is the most widely used corpus website in the world. as the creator Mark Davies states , it is useful for English learners and teachers as it allows for the distinction between formal and informal language , this is an important part of becoming fluent in any language , one can get as many contexts as they please for further understanding of any word searched , including pictures , videos , synonyms and translations.

3. The International Corpus of English

The Primary goal of the International Corpus of English (ICE) since 1990 has been to gather data for global English comparative studies. Together with linguists and researchers from numerous institutions and nations throughout the world, this project is a team effort as opposed to an individual one.Twenty-six research teams from various countries, including Australia, the Bahamas, Canada, East Africa, Fiji, Ghana, Gibraltar, Great Britain, Hong Kong, and India, are creating electronic corpora of their own national or regional varieties of

English. An English spoken and written word produced after 1989 makes up one million words in each ICE corpus. (The official website of ICE).

There are 500 texts in each component corpus, with an approximate word count of 2,000, for a total of about one million words.

The corpus encompasses a broad spectrum of age groups and includes samples of writing and speech by men and women in a variety of genres and registers. (The official website of ICE)

Some ICE corpora may be freely accessible online, while others may require registration or subscription for access according to the official website.

II. KEY THEORIES IN CORPUS LINGUISTICS

A corpus is a collection of text specific to a particular topic or language providing data for language research. Corpus Linguistics focuses on analysing and interpreting linguistic data from corpora, emphasising the quantitative analysis of large and structured collections of text or spoken data. It investigates language properties such as word frequency collocations and patterns of language use , to gain insights into the structure and nature of language. Corpus Linguistics relies on the presence of corpora for analysis and the Corpus gains value through linguistic analysis making them complementary. This field is characterised by specific theories and methodologies that transform a simple collection of text into a corpus suitable for detailed linguistic analysis.

1. Collocation Theory:

The theory of collocations finds its roots in the pioneering research of British linguist Herald Edward Palmer, Palmer's investigations into collocations prompted a fundamental revaluation of vocabulary. Later advancements in computer assistant Corpus analysis allowed

researchers to build upon Palmer's work. John Firth was among the first to delve into collocations to find them as actual words in habitual company as Kennedy (1989) stated.

In short , the term collocation denotes the idea that important aspects of the meaning of a word (or another linguistic unit) are not contained within the word Itself or considered in isolation, but rather subsist in the characteristic associations that the word participates in, alongside other words or structures with which It frequently co-occurs (McEnery, Hardie,2011p.123) , This means that when we talk about collocation we are highlighting that a word's meaning isn't just about the word itself , it's also about the company it keeps in other words it frequently appears with, these word partnerships give additional context and layers of meaning to individual words showing how language works beyond just individual units.by studying collocations within corpora , linguists can uncover patterns of word associations and understand typical language usage and gain insights into language variation and change which helps understand the nuances of meaning conveyed through word combinations that consequently improves the understanding of language structure usage and communication . (McEnery and Hardie ,2011).

2. Frequency analysis theory

Frequency refers to the number of times a particular word appears in a corpus, a frequency list shows you the words in the corpus along with the count of their occurrences. Depending on the purpose of your analysis you may wish to generate frequency lists with different subsets of the Corpus or particular fields. These lists can be tailored to analyse specific genres, time periods ,or authors providing insights into word usage patterns and helping to understand various linguistic phenomena.

Frequency analysis coupled with lemmatization offers a powerful method for understanding word usage patterns in a corpus by standardizing words to their base forms (lemmas) ,variations of the same word are grouped together providing a clearer picture of the word frequencies. For example variations like talk talks and talking are all counted under the same lemma talk. This ensures that even if the exact word talk is not repeatedly used in its root form it is counted for in the frequency analysis leading to more accurate linguistic insights. (Lu, 2014,p.69).

3. Semantic Analysis Theory

Unlike quantitative methods as the ones mentioned above that focus on numerical data, quantitative methods like semantic analysis explore meaning and context providing a deeper understanding of language use, words can have varied meanings depending on the context and semantic analysis help us make distinctions between those variations.

Semantic analysis dives into the meanings of words within their respective contexts, revealing the intricate layers of language usage. words while having inherent definitions can take on various shades of meaning depending on the context in which they are used , semantic analysis helps unravel these nuances offering a deeper understanding of language within the corpus by examining semantic relationships and contextual meanings, researchers can gain valuable insights into how language conveys ideas emotions and nuances within different textual environments. (Kennedy ,1998.p.225).

IV.PREQUISITS FOR USING CORPUS LINGUISTICS

In order for the researcher to adopt a corpus based approach, he must be aquainted with both theoretical and practical knowledge necessary. By theoretical knowledge we are referring to a strong foundation in theoretical linguistics (Bauer, 2007; Cruse, 2004; Ladefoged, 2006; Radford, 2011) to understand language structure and meaning, as well as

familiarity with corpus search tools and techniques. The practical knowledge includes familiarity with software and tools used for analysis and other skills that help researchers navigate the digital landscape of corpus linguistics.

III.Skills and Equitment

The process of building a corpus urges the researcher to have certain skills and materials without which this achieving this task is impossible.in the subsequent section we aim to address some of these skills and material.

1.Corpus Access

The cornerstone of corpus linguistics is access to relevant corpora. Thankfully, numerous online platforms offer free access to a variety of corpora, such as the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC), these platforms often provide user-friendly interfaces for searching and analyzing the data.

2.Corpus Analysis Tools

Once the researcher has identified a suitable corpus, he will need specialized software to unlock its potential. Popular choices include WordSmith tools and AntConc, These tools offer powerful features like concordance lines (showing how a word appears in context), word frequency analysis, and collocation identification (examining frequently co-occurring words) (Hoover et al., 2008).

V.APPLICATION OF CORPUS LINGUISTICS IN SOCIAL CONTEXTS

In language studies, the significance of corpora is intimately linked to the value of empirical data in general. In contrast to subjective assertions that are based on an individual's internalized cognitive perception of the language, linguists can make objective statements based on empirical data about language as it actually exists.Additionally, by using empirical

data, language variations that may not be amenable to a rationalist approach such as dialects or languages from earlier eras can be studied.

It is possible to conduct empirical language research without utilizing a corpus, though. While a lot of researchers will call their data a corpus, a lot of the time these data do not meet the definition of a corpus as we, along with many other corpus linguists, have used it in this book to refer to a collection of carefully sampled text that is maximally representative of a language or language variety.It would be more appropriate to think of this additional data as collections of texts. Therefore, corpus linguistics proper should be viewed as a subset of the activities that make up an empirical approach to linguistics: an empirical approach is a prerequisite for corpus linguistics, but using a corpus is not a requirement for empirical linguistics.

A.Corpora in Pragmatics and Discourse Analysis

According to Fairclough (1993, p.226) discourse analysis is another field where the "standard" corpora have not been used very much. This is mostly due to the fact that discourse analysts typically have an interest in examining the discursive practices connected to specific social practices. There exist significant areas of intersection between discourse analysis and corpus linguistics.First, in some discourse analysis traditions, the use of computer-aided corpus analysis techniques has a long history. For instance, Michel Pécheux used a fairly advanced form of automatic parsing as a supplement to a Marxist theory of discourse as early as the 1970s. A key component of his methodology was breaking down a corpus of sentences into sets of more manageable structures, after which he used distributional procedures to find recurring patterns of substitution and equivalency.Additionally, there is a lengthy history in France connected to a group of academics at the St Cloud of analyzing political text discourse by word-frequency statistics and multivariate statistical analysis, or "lexicometry" (Bonnafous and

Tournier 1995).As a last example, consider the growing interest in English linguistics in employing computer programs for concordancing and collocation analysis to support discourse analysis (cf. Hardt-Mautner 1995). Thus, there is a long history of methodological symbiosis between the two fields.

Second, the standard corpora have significant promise as control data in discourse analysis, while being employed in this manner rarely at the moment. Discourse analysts may legally inquire if certain features that they consider significant in a given corpus of texts are genuinely connected to the particular social practices at hand, or if they emerge from broader social practices that give rise to genres.Standard corpora can be used in conjunction with specialized discourse analysis corpora to determine the extent to which particular aspects are unique to the discourse under study and the extent to which they occur elsewhere in the language as a whole. This is because standard corpora encompass a wide range of genres and text types. For example, Myers (1991b) examined cohesion devices in a selection of scholarly and popular molecular genetics papers.He discovered that the two text forms tended to employ quite different kinds of devices: professional pieces used lexical cohesion far more frequently than popularizations did.

B.Corpora and Semantics

According to Mindt (1991), the corpus's primary useful function is to offer objective standards for determining the meanings of linguistic objects. Mindt notes that the majority of the time in semantics, the linguist's own intuitions—that is, what we have classified as a rationalist approach—are used to describe the meanings of lexical items and language structures. He continues, however, by arguing that semantic distinctions are, in fact, linked to distinct observable contexts in texts, such as syntactic, morphological, and prosodic ones. As a result, an objective empirical indicator for a given semantic distinction can be determined by taking into account the environments of the linguistic entities.Mindt supports this assertion with three brief studies in semantics. As an example, focusing only on the concepts of specification

and futurity. In this case, Mindt is interested in demonstrating the empirical validity of our understanding of the inherent futurity of verb constructions denoting future time. Specifically, he wants to know how much the sense of futurity seems to depend on co-occurring adverbial items that offer distinct indications of time (what he terms "specification") and how much it seems to be present in the verb construction itself on its own. Mindt examined the following four temporal constructions: will, be going to the simple present and the present progressive in two corpora the Corpus of English Conversation and a corpus of twelve modern plays and looked at how frequently the four different constructs were specified. He discovered that the simple present had the highest frequency of specification in both corpora, with the present progressive, will, and be going to following in that order. The two present tense constructions, which are frequently modified adverbially to intensify the sense of future time, are at one end of the scale according to the frequency analysis, and the two constructions that are naturally future-oriented are at the other end, with a much lower incidence of additional co-occurring words indicating futurity.Therefore, Mindt was able to show that objective indicators for intuitive semantic distinctions can be obtained through the empirical examination of language contexts in this case: it was found that intrinsic futurity was inversely connected with the frequency of specification.

The second significant role for corpora in semantics is the idea of fuzzy categories and gradience have been solidified, Traditionally, categories in theoretical linguistics have been thought of as hard and fast, meaning that an item either belongs in a category or it doesn't. It has been suggested by psychological research on categorization, however, that cognitive categories usually have fuzzy boundaries rather than hard and fast ones. This means that the question of whether a particular item belongs in a given category or not is less important than how frequently it does so compared to other categories.

This has significant ramifications for our comprehension of how language functions. For example, it implies that decisions about how to phrase things are driven by probabilities rather than rigid categories, as a model of language would imply. When natural language in corpora is examined empirically, it becomes evident that this "fuzzy" model better describes the data: gradients of membership that are correlated with inclusion frequency rather than simple inclusion or exclusion are frequently present instead of distinct category borders. When attempting to ascertain whether or not such gradients exist, corpora are a priceless resource. The subjects of verb formulations with future time reference attracted Mindt's interest in this case, particularly the distinction between subjects that entail conscious human agency and those that do not—a distinction that theory has previously recognized as crucial.Mindt calculated the frequency of personal and non-personal subjects of the four future time constructs in the same two corpora mentioned above as a rough correlate of this differentiation.Contrary to earlier theoretical assertions, he discovered that personal issues happened more frequently with the present progressive and less frequently with the simple future.Will and be going to had the same rank order for both corpora, with only a slight bias (about 2-3%) for personal subjects.

C.Corpora and Sociolinguistics

sociolinguistics is an empirical discipline of study that has historically focused more on the gathering of data relevant to the subject at hand than on broader corpora. However, these kinds of data are frequently not thoroughly sampled because they are not meant for quantitative research. In addition, solicited data can occasionally be used in place of naturalistic data. A corpus is significant because it can offer something that other types of data cannot: a quantifiable representative sample of naturalistic data.There is evidence of growing interest in the use of corpora in sociolinguistics, despite the fact that this subject has not yet made extensive use of them.

The majority of corpus-based sociolinguistic efforts to date have been comparatively straightforward lexical studies focusing on language and gender.The study by Kjellmer (1986), which examined masculine bias in American and British English using the Brown and LOB corpora, is a prime example of this. In particular, he examined the frequency of both masculine and feminine pronouns as well as the lexical elements man/men and woman/women.Kjellmer discovered that although feminine forms were more common in British English than in American English, the frequency of female items was significantly lower than that of male items in both corpora. He also discovered that the male to female ratios varied by genre. Generally speaking, women were more prevalent in imaginative prose than in instructional writing, with romantic fiction, predictably, having the the highest proportion of women. Though these frequency distributions did not surprise Kjellmer, he discovered that women and men actually had similar subject/object ratios, refuting his other theory that they would be less "active," or more often the objects than the subjects of verbs.

Holmes (1994) has examined the technique of these types of investigations in further detail.Holmes raises two crucial methodological arguments by focusing on three gender-related vocabulary issues: the usage of generic man; the frequency of Ms compared to Miss/Mrs; and the use of "sexist" suffixes. She first demonstrates the significance of considering the context and the availability of viable alternatives when classifying and counting occurrences. For example, the term "policeman" or "policewoman" has a non-gender marked counterpart, "police officer," yet the term "duchess of york" does not have such an option.It is therefore appropriate to omit the latter from counts of "sexist" suffixes when examining this type of gender bias in writing. Second, Holmes highlights the challenge of categorizing a form that is experiencing semantic change in real time.She uses the term "man" as an example.She contends that although a phrase like "A 35-year-old man was killed" clearly refers to a single male referent, and phrases like "Man has engaged in warfare for centuries" are clearly generic (referring to mankind), it is

difficult to determine whether a phrase like "We need the right man for the job" refers to a male person exclusively or if it is not gender specific and could just as easily be substituted by "person." The trend towards "non-sexist" writing makes it more difficult to determine whether a usage is generic as there is now a choice that was not there before.These straightforward but significant objections from Holmes need to encourage a more cautious attitude toward data classification in future corpora-based sociolinguistic research, both inside and outside the field of gender studies.

Yates (1993) conducted studies using a corpus of computer-mediated communication (e.g., e-mail) to examine the nature of this new genre from a sociolinguistic perspective, focusing on issues such as the presentation of knowledge, literacy practices, and the presentation of self. Yates has demonstrated how quantifiable measures can be generated from the Hallidayan theory of language as social semiotic.The operationalization of sociolinguistic theory into quantifiable categories that can be applied to corpora, the lack of sociolinguistic information encoded in current corpora, and the absence of sociolinguistically motivated sampling appear to be the only three practical issues impeding the greater expansion of sociolinguistic corpus work.Yates's work serves as a case of how to approach the first issue.Regarding the nature of the corpora, things are also evolving. Sociolinguistic characteristics including socioeconomic status, the writer's sex, and educational background are now being encoded in historical corpora. Such data are encoded in the Lampeter Corpus of Early Modern English Tracts and the Helsinki diachronic corpus.Modern language corpora, like as the Longman- Lancaster corpus, already have header categories for the writer's gender. In addition, the spoken portion of the BNC was gathered using demographic market research methodologies that took into account the writer's age, socioeconomic status, and geographic region.

V.LIMITATIONS OF A CORPUS

From the 1960s to the 1990s in continuing to the present day, Corpus linguists have emphasized on the compilation structure and size of corpora , this focus has evolved alongside advancements in technology and changes in linguistic theories and methodologies. Thanks to those advancements corpus linguists has been able to compile larger and more diverse corpora ,leading to ongoing discussions about the optimal size composition and structure of corpora. (Kennedy ,1998.p.60).

Corpora are very important for studying language in a variety of ways, no doubt in that. but like any other research tool, they are not perfect nor complete by any means. Some of its limitations as described by Kennedy (1998) include:

1. Limitation in Incompleteness

Languages infinite nature contrasts with the finite capacity of a corpus. One of the primary limitations of Corporal and also a long-standing argument in the Chomskyan Linguistics is the fact that a corpus can never capture language fully regardless of whether the Corpus is dynamic or static it will never achieve full coverage even when focusing on a specific era the Corpus cannot capture all language of that time. similarly and dynamic corporate studying language changes will also fall short in fully documenting with the evolution simply because again language is infinite and a corpus is not. (Kennedy ,1998.p.60-62)

2. Limitation in Size

Corpus size is a matter of discussion and there is no one size fits all answer. The appropriate corporate size depends on the specific research goals. a corpus that is too large

Paul's challenges in terms of management confrontational resources and overwhelming researchers with excessive data conversely a corpus that is too small like the necessary diversity and quantity of data.In essence the Corpus that is too large Maybe simply too large to handle while a corpse that is too small maybe insufficient for Meaningful analysis . (O'Keeffe, McCarthy, & et al., 2010).p.59)

3. Limitation in Computational skills

The requirements for computational skills which may not be within the expertise of all linguists. This dependency on computational skills may cause challenges for linguists who might not have a background in computer science or programming or handling data. Consequently the cooperation with computer scientists or acquiring computational skills becomes necessary introducing an additional layer of complexity to Corpus research. (Dash , 2010.p.21)

4. Limitation in Ethical Dilemma

One of the main things to consider when creating a corpus is whether you have a legal permission to collect and share the data you want to include. advancements in technology and the widespread availability of online content have made it easier for Corpus Builders to access vast amount of data yet the ethical question remains if the documents used, have been obtained and used in a consensual manner, The Dilemma comes when the consent requirements limit the availability of data for corpora as some documents might not be accessible due to absence and lack of consent..In an attempt to address this concern many institutions have developed ethics guidelines and regulations for Corpus construction and data privacy which can vary depending on the research objectives and may differ from institution to institution as well as from country to Country. (McEnery and Hardie ,2011.p.60-69).

Conclusion

In this chapter we explored and delved into the fascinating field of Corpus linguistics that does sometimes go beyond just linguistics, where we thoroughly examined major Corpus linguistic databases such as BNC and COCA. We also explored the essential tools, software and prerequisites required for effectively managing Corpus linguistics. Given its unique blend of language and computer science, CL demands specific prerequisites and software infrastructure. Despite encountering challenges like computational limits it maintains its usefulness to the field on linguistics and many other sub-fields

CHAPTER THREE: SCOPE AND METHODOLOGY

Introduction

This chapter presents the scope of our research and the methodology utilized, focusing first on the languages spoken in Tiaretian landscape as the scope, Then we deal with the practical aspects of compiling a lexical database (corpus) for Tiaret, Algeria. Our aim is to gather and analyze linguistic data specific to this region. However, we encountered significant challenges due to the inability of existing tools to accurately recognize and process Arabic scripts. This chapter will discuss these challenges and the approaches we considered to address them, highlighting the impact on our research outcomes.

Ⅰ- SCOPE OF THE RESEARCH

1. Historical Background of Tiaret's Landscape

Thus, linguistic situation in the Tiaret speech community is particularly fascinating and can be accurately described as multilingual. Currently, the people of Tiaret use several language varieties, including Modern Standard Arabic (MSA), Classical Arabic (CA), Algerian Dialect (AD), French, and Berber. It can be said that they employ multiple language varieties within the same context to achieve specific communicative purposes.

Classical Arabic (CA) is the language of the holy book (Quran), while Modern Standard Arabic (MSA) is the official language used in administrations and schools. The Algerian Dialect Arabic (ADA) is the dominant language variety spoken by the majority of the Tiaret community. Despite this, MSA and CA are also widely used by most people. Additionally, the minority

Berber group uses their native language variety when communicating with each other.The majority of people speak French as their first language while speaking foreign languages since, even after 55 years of independence, many government officials and administrative personnel still speak French. The French language is becoming more and more noticeable in business signage these days, particularly in Tiaret town's urban public signage.

French is the dominant spoken language used by the majority of inhabitants. Additionally, English, being the first choice of foreign language in many countries around the world, holds the status of a global lingua franca.In Algeria, including Tiaret, English holds the status of a foreign language. It is taught from primary school to secondary school. English is becoming increasingly attractive to the new Algerian generations, as evidenced by its influence on clothing, advertisements, and popular culture choices. Furthermore, English is seen as a pathway to success and job opportunities, with many companies and private schools valuing proficiency in the language. Additionally, it is common for shop owners to code-switch between English, French, and Arabic, indicating that English is a useful language in contemporary times.

A. Brief History about Tiaret

Tiaret is a town of approximately 1,500,000 people, located about 100 miles inland from the Mediterranean seacoast. Known variously as Tiaret, Tahert, or Tihert, it is the main city in the province of Tiaret, an upland agricultural region in the Tell Atlas area of Algeria. The word "Tihert" means "station" in the local Berber dialect, and historically, Tiaret has served as a station or stopping place for travelers, traders, and armies. Situated in a strategic mountain pass, Tiaret has been crucial for any power seeking to control the surrounding land and the lucrative trade routes that pass through it. It was also a key point for funneling slaves from sub-Saharan Africa.

Figure 3.1: The Location of Tiaret in Algeria

Note:Reprinted from Shutterstock 2024, Retrieved May 18 , 2024 from google image *https://www.shutterstock.com/fr/image-vector/tiaret-red-highlighted-map-algeria-1731045568*

With 42 municipalities and 14 Dairas, Tiaret has a total area of 20.050.05 square kilometers. It is surrounded by a number of wilayates, including Djelfa to the east, Mascara and Saida on the western side, Laghouat and El Bayad to the south, and Tissemsilet and Relizane to the north.The Wilaya of Tiaret seems to serve as both a hub for numerous significant wilayates and a point of communication for the North and South.Because of its vastness, the area has a clever underbelly. a region of mountains to the north. the semi-arid regions south of the wilaya and the high plains in the middle. These traits demonstrate the diversity of the landscape and terrain. Tiaret is regarded as a farming region due to its high-quality wheat of several varieties and other agricultural goods that significantly boost the economy of the country.An estimated 7,190,000 sheep, 347,652 cows, and 615,957 goats are provided by the state.The original Arabic horses (the Chawchawa barn) are also known. It looks after 288 horses, of whom 174 are native Arabian horses and 68 are other savage horses.

In 1153, Tiaret joined the Almohad dynasty, also known as Al Mowahidin. In 1253, it became a part of the Ziyani state. Abd Ar Rahman Bin Muhammed ibn Khaldoun started the introduction (Muqadimma) in 1377, laying out his ideas regarding critical historiography and intending to write a history of the modern Maghreb. The Muqadimma quickly expanded into a comprehensive historical theory. or civilizational science. how he described it. Several languages have translations of his work available. In his honor, Tiaret University was established

Figure 3.2: Tiaret's Gaographical Map

Note : Reprinted from Gifex , Retrieved May 18 , 2024 from Google images
https://gifex.com/fr/fichier/quelles-sont-les-communes-de-la-wilaya-de-tiaret/

B. Linguistic Behavior of Tiaretian People

The Algerian dialect, referred to as Arabic or Daridja by speakers, serves as the primary language for mutual communication in Tiaret. It is the mother tongue of 75-80% of the population and used by 95-100% of Tiaretians.

Arabic has been the official language in Tiaret since the constitution of 1963, alongside Tamazight. Approximately 99 percent of Tiaretians speak Arabic and Tamazight, with 72 percent speaking Arabic and 22 percent speaking Berber. Additionally, French is widely used in the Tiaretian state, particularly in cultural fields, media, and education at universities, reflecting the historical influence of French colonization in Algeria. While French is considered a semi-official language in the Tiaret speech community, it is not officially recognized in any state publications.The Arabic dialect spoken in Algeria is distinguished by a number of regional forms that fall into two categories. Urban dialects predate Hillel dialects, which are Bedouin dialects. Algerian dialects are found on the high plateaus near Setif, Tebessa, Biskra, Bordj Bou Arreridj, Mesilla, Djelfa, and Laghouat. They are classified as Eastern Hillalian dialects.

The simplicity with which all current letters can be used characterizes Tiaretian speakers. They are able to mimic every dialect that is spoken in the other areas. For example, a Tiaretian speaker who relocated to Oran would have little trouble mimicking their dialects and getting along with them. The other dialects are in the same situation. The traveler Al Maqdisi discusses the Al Maeqal dialect, which is spoken in parts of western Algeria like Tiaret town, in the same line of reasoning. In contrast to what we have described in the regions, their Arabic language is closed, and they also speak a tongue that is similar to Rumi.

II-Methodology of the Research

1. Research methodology

The primary objective of this research paper is to explore and examine the linguistic disparities between the older and younger generations within the Tiaret region and its surrounding areas. By initiating this investigation, we aim to shed light on the evolving nature of language usage

and communication styles across different age groups. Our ultimate goal is to not only identify these differences and spot them but also to utilise our findings and make use of them. Our findings are going to be used as a key that opens a door of clarifications and closes another door of ambiguity and miscommunication between generations.

This research is of an exploratory nature as it attempts to explore the nature of the factors that contribute to the changes that occur within language during its journey of evolvement , these factors could be of a historical nature , a cultural nature , a technological nature and , or overall social nature. By exploring these factors, we will gain insights into how these changes are illustrated in the linguistics form (language). That is why we adopted the corpus-based approach, which entails collecting a certain amount of data from written or spoken texts (in our research it is spoken). This data is then organised into a corpus, allowing for specific analysis and examination of patterns and changes in language. We chose this approach as it was the most suitable method for studying linguistic phenomena in our context of trying to uncover the older generations' communication styles and how it sits in contrast to the younger generation.

a. Data Collection Tools

During our research, we engaged in personal discussions with elderly residents of Tiaret and its nearby regions, ranging in age from 55 to 75. Through these one-on-one conversations, we gathered valuable insights into the evolving nature of language and observed the distinct communication styles of the older generation. Our focus covered a broad range of linguistic elements, including expressions, vocabulary usage, words, and descriptions of various objects. In order To facilitate these discussions, we employed unstructured interviews as our primary data collection method. By using open-ended questions, we set an environment where conversations could flow naturally, allowing participants to share their thoughts and

experiences freely and spontaneously. This approach was particularly crucial for our research, as it aimed to capture the nuances of naturally occurring language usage.

Throughout these interactions, we took precise notes while actively listening to the perspectives of the participants and engaging with them. This provided us with a deeper understanding of the collaborative connections between language, culture, and generational differences.In addition By involving ourselves in these conversations, we were also able to gain valuable insights into how language evolves over time and how it serves as a marker of generational identity and belonging , as the different communicative style the participants presented was of a distinct nature , as we (a part of the younger generation) were able to detect during the exchange.

b. Participants

Participants in my research were selected based on specific criteria , with the primary focus being on the older generation of Tiaret .Therefore , individuals aged 55 to 75 years , originating from tiaret area ,were chosen to participate in the study .

2. Data Collection and Data Analysis Procedures

A. Data Collection

Prior to the interview, the following data is collected. We are providing data as it is for the research integrity, then we not only are going to provide a meaning for each word but also a transliteration for the pronunciation of each word to facilitate understanding for none Algerian students

1. القْطيفَة -Al-q’teefa

2. البورابَح - Al-buurabah

3. الكْسَى - Al-ksa

4. المَزْوَدْ - Al-mazwad

5. القَرْبَة - Al-qarba

6. الشَكْوى - Ash-shakwa

7. لتّفال - At-tifal

8. الحَمّارَة - Al-hammaara

9. الحرّاق - Al-harraaq

11. القَرْنَنٌوش - Al-garnanuush

12. لبَردعة - Al-bard'a

13. الحلاس - Al-hlaas

14. الخليع - Al-khlee'

15. القديد - Al-qadeed

16. الدَّق - Ad-daq

17. القَسْطْ - Al-qast

18. الرْحَلْ - Ar-rahal

19. الغرارة - Al-gharaara

20. التاتة - At-taata

21. القَرْنجدي - Al-qarnjdi

22. التافغة - At-taafgha

23. القِيز) - Al-geez

24. التّالمة - At-taalma

25. المَريوت - Al-maryuut

26. المَشْطْ - Al-mashṭ

27. الصُوصيا - As-suusiya

28. الكليلة - Al-kaleela

29. العُكة - Al-ukka

30. المزريط - Al-mazreet

31. البياضة) - Al-bayaadha

32. مفلقة تشيشة - Tsheesha mufalga

33. النادر - An-naadir

34. مالطي قمجة - Qamja maalti

35. كانكي - Kaanki

36. الكوك - Al-kook

37. الشواري - Ash-shawaari

38. الخُرْجْ - Al-khurj

39. السْماطْ) - As-smaṭ

40. العْمارة - Al-amaara

41. القَرْويشَة - Al-qarweesha

42. الكَمْبُوش - Al-kamboosh

43. الكْلاحْ - Al-klaaḥ

44. الحَسْكة - Al-ḥaska

45. اللْيان - Al-lyaan

46. الهَرماس - Al-harmaas

47. المَرْوَدْ) - Al-marwad

48. القَرْدَاش - Al-qardaash

49. الهَيْدُورَة - Al-haydoora

50. الذَرْوَة - Ad-dharwa

51. الكَفاتيرة) - Al-kafaatira

52. الطَرْحة) - Aṭ-ṭarḥa

53. الراحْلَة - Ar-raḥla

54. الدير - Ad-deer

55. اللِجام - Al-lijaam

56. الرْكيزَة - Ar-rkeeza

57. القربي - Al-qarbee

B. Data Analysis procedures

The process data analysis of the the Tiaretian old language corpus (Lexical database) involves several procedural steps. Initially, the corpus must be compiled from naturally spoken

language of the original inhabitants of Tiaret. This requires meticulous transcription and digitization of these sources. Following this, the data must be cleaned to remove any inconsistencies or errors. Next, linguistic features such as phonetics, morphology, syntax, and semantics need to be annotated, typically using Natural Language Processing (NLP) tools. This annotation process involves identifying and categorizing linguistic elements within the corpus to facilitate detailed analysis. Subsequently, statistical and computational methods are applied to analyze patterns and structures within the language, providing insights into its unique characteristics. Each step must be performed with precision to ensure the reliability and validity of the analysis.

3. Data Analysis

A. Data translation: Contextual and Cultural Aspects of the Gathered Data

Bellow are the literal and contextual meanings of some of the data, in which we provide a clearer understanding of the data in terms of use and functionality. The reason behind the missing part of the words is our inability to find equivalents in the English culture, we will address this issue in the limitations section.

Table 3.1: *Tiaretian Lexicon: Words, Meanings, and Contextual Descriptions*

Word	Meaning	Contextual Meaning
القَرْبَة	Waterskin	A flexible container made from leather or animal skin, traditionally used for storing and carrying water.

الشَكْوى	Churn (for milk)	A traditional wooden or ceramic vessel used for churning milk into butter or yogurt.
لتّفال	Doum plate (for couscous)	A flat plate made from the leaves of the doum palm tree, traditionally used for serving couscous.
الحَمّارَة	A tripod for hanging meat	A three-legged stand made of wood or metal, used for hanging and drying meat after slaughter.
الحرّاق	A coffee roaster	A traditional tool or container used for roasting coffee beans.
العَسْلُوجْ	An edible herb	A wild plant or herb considered edible and used in cooking or traditional medicine.
لبَردعة	A donkey saddle	A padded saddle placed on a donkey's back for carrying loads or riding.
القَرْنَنْ ُوش	An edible herb (used in salad)	A wild plant or herb considered edible and often added to salads.

الحلاس	A donkey pad	A thick pad placed under the donkey saddle to provide comfort and protection.
الخليع	A preserved fat (used in bread)	Fat from sheep or other animals, preserved by salting or other methods, used in traditional bread recipes.
القديد	Dried meat	Meat, typically beef or lamb, that has been dried and preserved for long-term storage.
الدَّق	A ground herb (for coffee)	A type of herb, often roasted and ground, added to coffee for flavor or medicinal purposes.
القَسْطْ	A container (for preserved fat)	A small pot or container used for storing preserved fat (khlie).
الرْحَلْ	A rack (for carrying sacks)	A wooden or metal frame used for carrying multiple sacks or bundles on camels or donkeys.
الغرارة	A wool sack (for wheat)	A large sack made of wool, traditionally used for storing and transporting wheat.
التاتة	A lizard	A general term for any type of lizard found in the region.

القَرْنجدي	An edible spring plant	A wild plant or herb that grows in the spring and is considered edible.
التافغة	An edible plant (similar to artichoke)	A wild plant or herb with a similar appearance to an artichoke, considered edible.

B. Data Transcription

In the table below we have provided data and its phonetic transcription as part of the data analysis

Table 3.2 *Tiaretian Words and Their Phonetic Transcriptions*

Word	**Transcription**	**Word**	**Transcription**
القْطيفَة	[alˈqtiːfa]	المَشْطْ	[alˈmaʃṭ]
البورابَح	[albuːˈraːbaħ]	الصُوصيا	[aṣˈṣuːṣijaː]
الكْسَى	[alksaː]	الكليلة	[alkaˈliːla]
القَرْبَة	[alˈqarba]	العُكة	[alʕˈukka]
الشَكْوى	[aʃˈʃakwaː]	المزريط	[almaˈzriːṭ]
لتّفال	[atˈtifaːl]	البياضة	[albaˈjaːða]

الحَمّارَة	[alħamˈmaːra]	مفلفة تشيشة	[tʃiːʃa mufaˈlɣa]
لحرّاق	[alħarˈraːq]	النادر	[anˈnaːdir]
العَسْلُوجْ	[alʕasˈluːʒ]	قمجة مالطي	ˈ]qamdʒa ˈmaːlṭi]
القَرْنَنْ ُوش	[algarnanˈnuːʃ]	كانكي	[ˈkaːnki
لبَردعة	[alˈbardʕa]	لكوك	[alˈkuːk]
الحلاس	[alħˈlaːs]	الشواري	[aʃʃaˈwaːri]
الخليع	[alxˈliːʕ]	الخُرْجْ	[alˈxurʒ]
القديد	[alqadˈdiːd]	السْماطْ	[asˈsmaṭ]
الدَّق	[adˈdaq]	العْمارة	[alʕˈmaːra]
لقَسْطْ	[alˈqasṭ]	القَرْويشَة	الحجر)
الرْحَلْ	[arˈraħal]	الكَمْبُوش	[alkamˈbuːʃ]
الغرارة	[alɣaˈraːra]	الكْلاحْ	[alkˈlaːħ]
التاتة	[atˈtaːta]	الحَسْكة	[alˈħaska]
القَرْنجدي	[alqarˈnʒdi]	اللْيان	[allˈjaːn]
التافغة	[atˈtaːfɣa]	الهَرماس	[alˈharmaːs
القيز	[alˈgiːz]	المَرْوَدْ	[alˈmarwad]
التّالمة	[atˈtaːlma]	القَرْدَاش	[alqarˈdaːʃ]

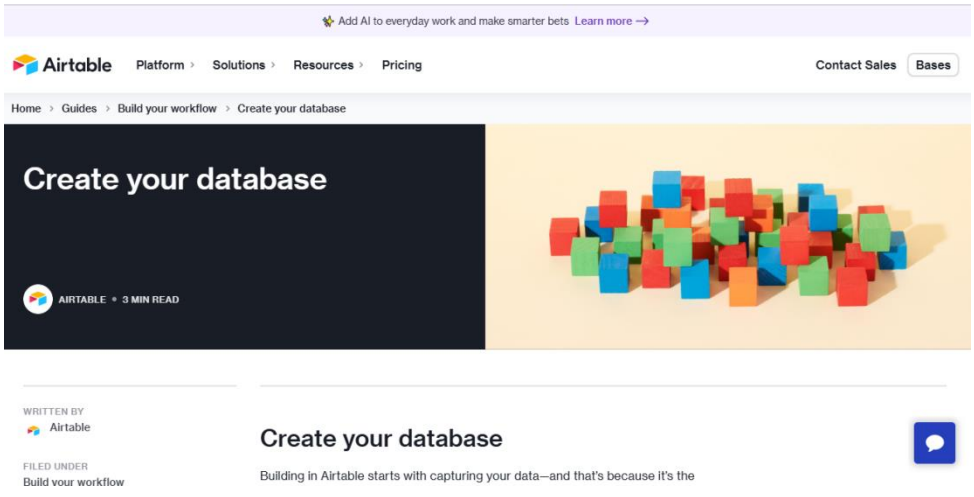
المَريوت	[almarˈjuːt]	الهَيْدُورَة	[alhajˈduːra]
الذَرْوَة	[adˈðarwa]	الكَفاتيرة	[alkaˈfaːtiːra]
الطَرْحة	[aṭˈṭarħa]	الراحْلَة	[arˈraħla]
الدير	[adˈdiːr]	اللِجام	[allijˈjaːm]
الرْكيزَة	[arrˈkiːza]	القربي	[alqarˈbiː]

C. Digitization Process

After a careful selection of databases creation programs and with assistance of the technician we have hired, we have decided to utilize **Airtable** to generate our lexical database. The decision was made after a careful review of the use and functionalities it offers, and also the fact that it requires minimum level of technical knowledge. Therefore, the procedural steps to build the database are as follows:

- Step one : We sign up for a free account on Airtable.

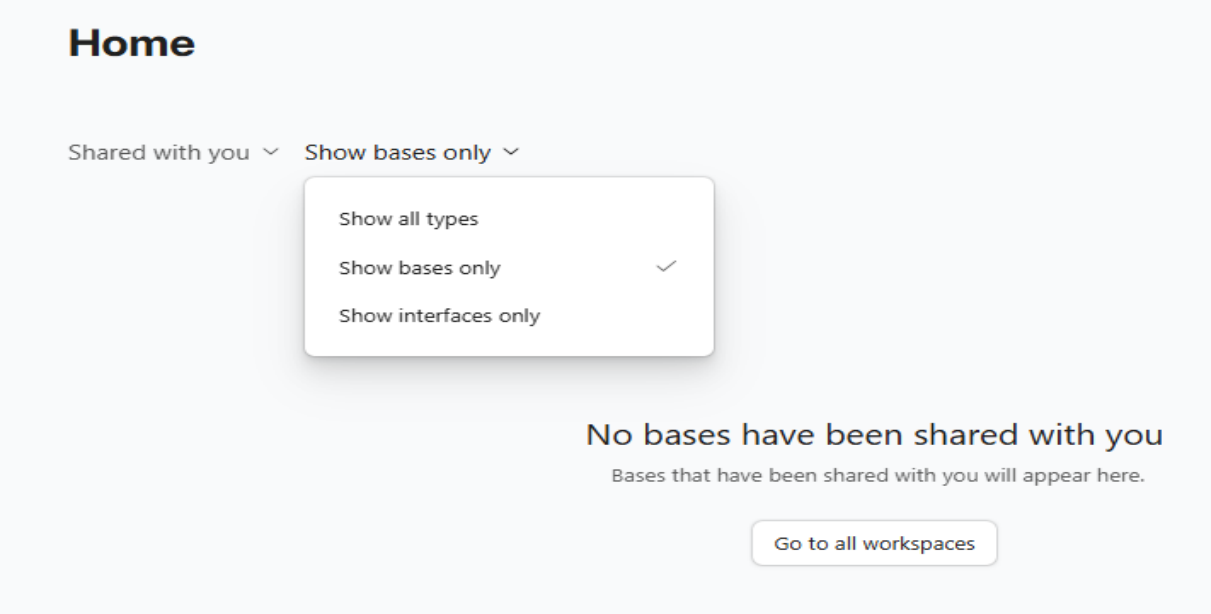
Figure 3.3: A Screenshot of Signing Up in Airtable



.

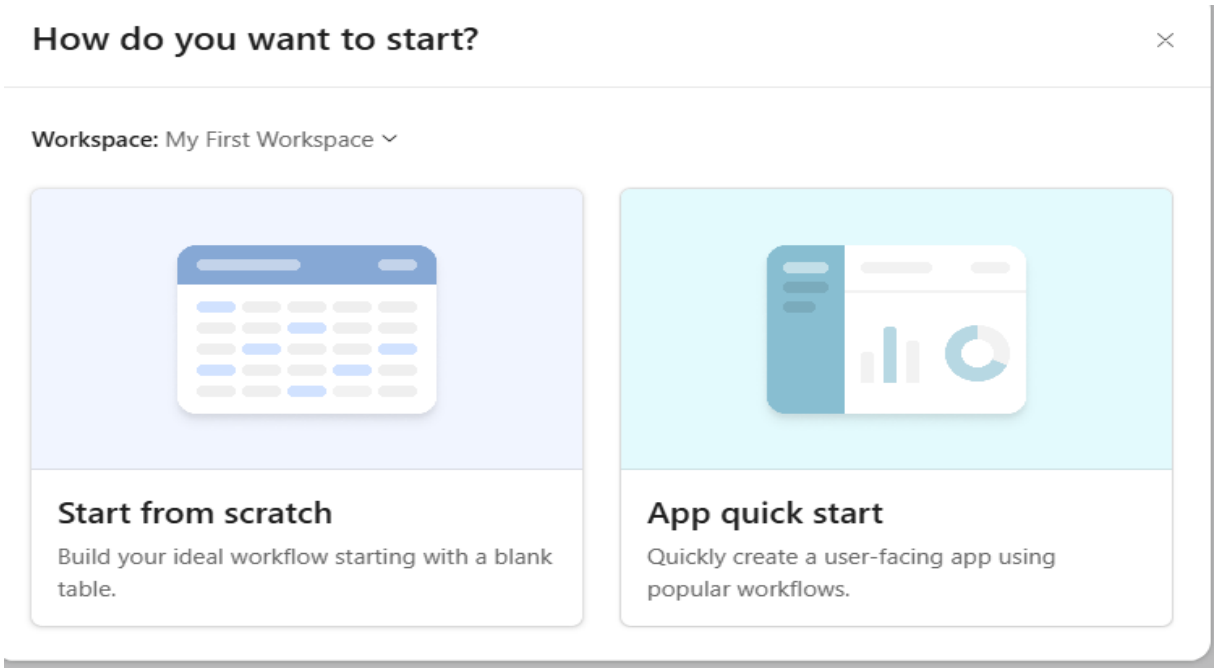
- Step two : Once we are logged in, we click “Add a base” to create a new database.

Figure 3.4: A Screenshot of creating a new database in Airtable



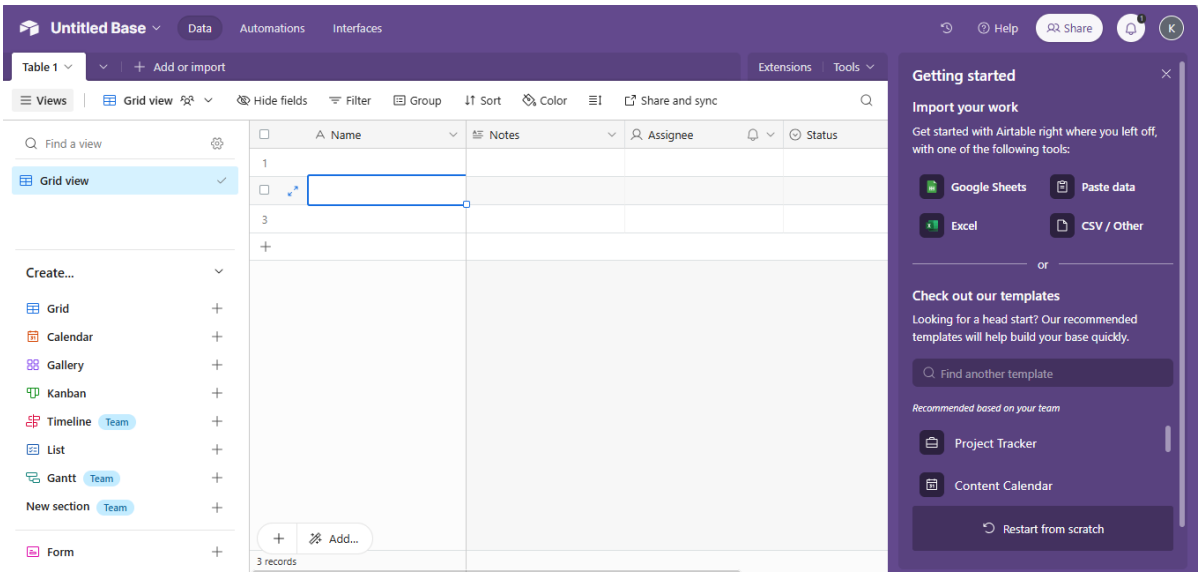
- Step three: Then we choose the “Start from scratch” option, and give our database a name, like "Tiaretian Lexicon."

Figure 3.5 :A screenshot of Initializing Your Database: Naming 'Tiaretian Lexicon



- Step four :Now, we will see a default table with columns labeled "Name," "Notes," etc.
- Step Five : the next step is to rename these columns to match our data: “Word,” “Pronunciation,” “Meaning,” and “Context,”.” Next, we enter the data. We click into each cell to type the corresponding information from our corpus. For example, under “Word,” type the Tiaretian word; under “Pronunciation,” enter how it is pronounced; under “Meaning,” describe what it means; under “Context,”. If the data is in a spreadsheet, we can import it directly by clicking on the dropdown arrow next to the table name, selecting “Import Data,” and following the prompts to upload our file. Once the data is entered or imported, Airtable automatically saves changes. We can now use Airtable’s features to sort, filter, and organize your words easily, creating a well-organized and accessible digital database of your Tiaretian lexicon

Figure 3.6: A Screenshoot of Airtable: Organizing the Tiaretian Lexicon



D. Database Integration

Airtable was chosen due to its simplicity, comprehensive features, and minimal need for technical expertise.. However, we encountered a significant limitation: Airtable does not support our language, Tiaretian, for language detection, it does not support Arabic symbols. This limitation affects our ability to fully utilize certain features.

4. Discussion of findings

The study aimed to build a corpus of spoken Tiaret Arabic, focusing on the dialect used by the older generation in Tiaret, Algeria. Despite the limited data collection, which resulted in only 57 words being gathered, the analysis of these words provided significant insights into the linguistic features and cultural heritage of the Tiaret dialect. The unique lexical items, such as "القَرْبَة" (waterskin) and "الشَكْوى" (milk churn), reflect the distinct cultural and environmental context of Tiaret. This indicate that the Tiaretian dialect contains unique lexical items not found in other Arabic dialects.

Among the findings, it is notable that certain words are still actively used by the youngest generation, underscoring continuity in linguistic heritage despite modern influences. Words such as الكَفاتيرة, البورابح, الخليع, القديد, and الدقcontinue to find currency among younger Tiaretians. These words are likely retained due to their practical utility in traditional contexts or cultural significance that transcends generational shifts. Conversely, the study also identifies words that have fallen out of common usage among the youngest generation, including الليان, الحسكة, الدير, القيز, التافغة, and التالمة. The decline of these words can be attributed to several interrelated factors. Firstly, technological advancements and globalization have introduced new lifestyles and dietary preferences among the younger generation. Modern conveniences like fast food chains offering pizza and tacos have supplanted traditional dietary practices that once relied on (القرنجدي ، القرننوش , العسلوج) which are edible plants . This shift in dietary habits diminishes the need for specific vocabulary associated with traditional food items.

Furthermore, changes in socio-cultural practices and urbanization have diminished the daily relevance of certain traditional tools or practices, thereby reducing the frequency of corresponding vocabulary in everyday speech. Words like (which refer to pad) البردعة , الحلاس and saddle placed on donkey's back. These words have lost currency as Tiaretian society has shifted away from rural agricultural practices towards more urbanized lifestyles.

Based on the study's findings, the hypothesis suggesting that the complexity associated with compiling the corpus would pose a significant challenge was partially confirmed, as the study encountered major challenges in data collection, resulting in a smaller corpus than initially anticipated.

In summary, while one of the above-mentioned hypothesis has been confirmed, the major limitations we have faced hindered our attempt of studying language preservation through

corpus linguistics in the Tiaretian context. The findings emphasize the importance of continued research and proactive measures to document and preserve the unique linguistic heritage of the Tiaretian community amidst ongoing societal changes.

The study highlights the challenges of digitizing and organizing a lexicon for a dialect that lacks widespread digital support. Our Attempt to build a lexical database using Airtable has been aborted because of its inability to recognize Arabic scripts.. This experience underscores the need for more advanced and culturally sensitive digital tools to support the preservation of minority languages and dialects. Despite these technical challenges, the study successfully demonstrates the feasibility of creating a digital lexicon for Tiaret Arabic. The meticulous process of transcription, annotation, and data entry ensures the reliability and validity of the collected data. By integrating linguistic features such as phonetics, morphology, syntax, and semantics, the analysis provides a comprehensive understanding of the dialect's structure and usage. Statistical and computational methods further elucidate patterns within the language, offering insights into its unique characteristics.

The study's findings also suggest avenues for future research and preservation efforts. Continued data collection from the older generation in Tiaret will enrich the corpus and provide a more comprehensive representation of the dialect. Collaborations with linguistic experts and the exploration of alternative digital platforms could enhance the accuracy and usability of the lexical database. Additionally, educational initiatives to raise awareness about the cultural and linguistic heritage of Tiaret Arabic could foster greater appreciation and usage of the dialect among younger generations.

5. Limitations of the study

In our attempt to preserve the Tiaretian dialect through corpus, linguistics, we encountered numerous challenges. Firstly, the dialect's shortage of written and spoken resources necessitated primary data collection within a limited timeframe, whereas corpuses require a large amount of data. Additionally, certain words and linguistic elements relevant to the dialect were fading, requiring careful manual efforts for documentation, yet often without available visual aids. Moreover, the project faced setbacks due to our limited proficiency in computer software, which hindered efficient data management and analysis as both skills in computer science and linguistics are required. Compounding these issues was inadequate access to essential technology such as reliable internet and computing infrastructure, further complicating data collection and storage. Furthermore, existing software tools and platforms did not support the Tiaretian language, making the technical challenges more difficult faced in corpus construction and linguistic analysis. Addressing these intertwined challenges underscores the complexities and urgent needs utilising corpus linguistics for effective language preservation efforts.

6. Suggestions and Recommandations

The preservation of the linguistic heritage of Tiaret is not only a matter of academic research but also a cultural duty, that emphasises the importance of safeguarding cultural heritage and promoting linguistic diversity, particularly for minorities. Our research was an attempt to preserve the Tiaretian linguistic heritage while introducing a corpus based approach , but encountered challenges due to inadequate technological support , highlighting the need for advancements in linguistic tools for minority languages. Here are some recommendations derived from our study:

-The exploration and implementation of more advanced linguistic tools and technologies specifically tailored to support minority languages like the one of Tiaret

-Further studies should prioritise the collection of extensive datasets of the Tiaretian dialect in both written and spoken form.

-Ensuring the availability and easy access to data would facilitate more comprehensive research and analysis of the linguistic features of the Tiaretian speech variety.

-Encouraging researchers in the field to develop computational skills in order to enable them to independently manage and analyse corpus data, reducing dependency on external technicians and enhancing research autonomy.

-developing or adapting software specifically designed to handle diverse linguistic varieties as the one of Tiaret , ensuring compatibility with less researched and less used languages.

-Promoting open access to already digitised data and resources if available , allowing broader academic and public access for research and educational purposes.

- looking into the potential of machine learning and artificial intelligence algorithms to automate aspects of transcription, translation, and analysis, with the aim of accelerating the digitization process.

Conclusion

Finally, in this chapter we took a brief look into the overall history of tiaret and its linguistic and cultural history , before revealing the practical part of the study ,that focused on the preservation of Tiaretian linguistic heritage through a corpus based study . the primary findings of our research highlight a diverse range of linguistic features use within the older generation . Through an exploration of language dynamics with a focus on the older generation , this part of the research illuminated how language serves as a dynamic marker of identity and cultural continuity. It underscored the resilience of certain linguistic traditions

while highlighting the evolution and potential decline of others, due to societal changes influenced by globalisation, technology, and shifting cultural practices.

By documenting and analysing the lexical richness of Tiaretian arabic , particularly from the perspective of the older generation, this study contributes to the preservation and appreciation of Tiaretian cultural identity ,regardless of the fact that our ultimate aim of building a corpus databse for the tiaretian speech varierty was not accomplished ,we were still able to capture and gather useful data and provide information to be used in further studies , and shed light on issues that would open many doors in both linguistic research field and language preservation .

Moving forward, efforts to further document and analyse Tiaretian dialects can enhance our understanding of linguistic diversity and its significance in maintaining cultural heritage alongside the exploration of modern linguistic tools and softwares in favor of minority linguistic speech varieties is highly recommended .

GENERAL CONCLUSION

In conclusion, this study aimed to preserve the linguistic heritage of the speech variety of Tiaret through the application of corpus linguistic methods . To achieve this , individuals of the older generation with the age range of 55 to 75 , originating from Tiaret and the surrounding areas , were the main source of the data collected , considering , they are the beholders of the linguistic and the cultural heritage of the region .

The primary comprehensive evaluation of the data, revealed apparent linguistic disparities between the older and younger generation. Where basic day-to-day words of the older generation were rarely used, to unknown by the younger ones. Through detailed linguistic analysis, including examination of meaning, contexts and consideration of social and generational shifts, reasons for the neglect of these linguistic features by the younger generation were social factors such as shifting cultural and social norms, social evolution and technological advancements.

The linguistic divide identified was just the initial step of our research , with the creation of the corpus database as a preservation tool as our primary aim. Despite the assistance of a professional with the data digitization and compilation , we were unable to go through with the process as the reason that hindered it , was the fact that current corpus linguistic tools and softwares did not support minority language varieties , the Tiaretian variety being one of them. This underscores the need for ongoing advancements and methodologies specifically designed to support less researched diverse linguistic varieties.

Based on the findings of this study , it is recommended that further research should not only focus on the advancements of technological tools and softwares , but also be directed towards the availability of data , spoken and written forms specific to the Tiaretian dialect. The

increased access to such resources will enrich the amount of accessible data and therefore enable computational tools to accurately digitise and analyse the unique features of our dialect.

In summary, this study has shed light on the linguistic intergenerational gap within the Tiaretian speech community and introduced a tool aiming at its preservation. The findings emphasised the necessity for continued efforts to explore the development of modern technological preservation methods, particularly corpus methods. Despite the limitations encountered, the outcome of this research lay the groundwork for future research efforts in the field, emphasising the ongoing importance of preserving linguistic diversity that maintains cultural heritage for future generation

List of References

- Bauer, L. (2003). *Introducing linguistic morphology*. Edinburgh University Press.
- Biber, D., & Reppen, R. (2012). *Corpus linguistics*. Sage.

Chiang , D. (n.d.). *Asian Chapter of the Association for Computational Linguistics - ACL Anthology* (A. Köhn , D. Gildea, & N. Schneider, Eds.). Aclanthology.org. Retrieved June 13, 2024, from https://aclanthology.org/venues/aacl/

- Cruse, D. A. (2000). *Meaning in Language*. Oxford University Press, USA.

editor. (2023, August 30). *Cross-Generational Communication: Bridging the Generation Gap*. https://rcademy.com/cross-generational-communication/

- Gries, S. Th. (2009). What is Corpus Linguistics? *Language and Linguistics Compass*, *3*(5), 1225–1241. https://doi.org/10.1111/j.1749-818x.2009.00149.x
- Imad Zeroual, & Abdelhak Lakhouaja. (2018). Arabic Corpus Linguistics: Major Progress, but Still a Long Way to Go. *Studies in Computational Intelligence*, 613–636. https://doi.org/10.1007/978-3-319-67056-0_29
- Kennedy, G. D. (1998). *An Introduction to Corpus Linguistics*. Longman Publishing Group.
- Ladefoged, P., & Sandra Ferrari Disner. (2012). *Vowels and consonants*. Wiley-Blackwell, Cop.
- Mcenery, T., & Wilson, A. (2001). *Corpus linguistics : an introduction*. Edinburgh University Press.
- Radford, A. (1988). *Transformational grammar : a first course*. Cambridge University Press.
- Soumia Bougrine, Aicha Chorana, Abdallah Lakhdari, & Hadda Cherroun. (2017).

Toward a Web-based Speech Corpus for Algerian Dialectal Arabic Varieties. https://doi.org/10.18653/v1/w17-1317

- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. J. Benjamins.
- Venter, E. (2017). Bridging the communication gap between Generation Y and the Baby - - Boomer generation. *International Journal of Adolescence and Youth*, *22*(4), 497–507. https://doi.org/10.1080/02673843.2016.1267022
- Dash, N. S. (n.d.). *Corpus Linguistics: An Introduction*. Pearson Education India.
- Davies, M. (2008). *Corpus of Contemporary American English (COCA)*. English-Corpora.org. https://www.english-corpora.org/coca/
- Desagulier, G. (2022, July 4). *Noam Chomsky's colorless green idea: "corpus linguistics doesn't mean anything."* Around the Word. https://corpling.hypotheses.org/252
- Endarto, I. T. (2017). *Systemic Functional Linguistics: A Brief Introduction*.
- *Halliday: Meaning, Functions, Theory, Examples | StudySmarter*. (n.d.). StudySmarter UK. https://www.studysmarter.co.uk/explanations/english/language-acquisition/halliday/
- Hamzaoui , C. (2020). *an Introduction to corpus linguistics.* Belhadj Bouchaib University – Ain- Temouchent.

Importance and Application of Corpus Linguistics. (2023, June 16). BSELN.COM. https://www.bseln.com/2023/06/importance-and-application-of-corpus.html

- Kennedy, G. (2014). *An Introduction to Corpus Linguistics*. Routledge.
- Lavid, J. (2008). To the memory of John Sinclair, Professor of Modern English Language. *Www.academia.edu*, *15*(2). https://www.academia.edu/78833528/To_the_memory_of_John_Sinclair_Professor_of_Modern_English_Language

LIST OF REFERENCES

- Lu, X. (2014). *Computational Methods for Corpus Annotation and Analysis*. Springer.
- Lukl, J. (2019). A TRIBUTE TO MICHAEL HALLIDAY. *Theory and Practice in English Studies, 8*(1), 9.
- McEnery, T., & Hardie, A. (2012). *Corpus Linguistics*. Cambridge University Press.
- Mcenery, T., & Wilson, A. (2001). *Corpus linguistics : an introduction*. Edinburgh Textbooks in Empirical Linguistics.
- O'Keeffe, A., & McCarthy, M. (2012). *The Routledge handbook of corpus linguistics*. Routledge.
- OUCS. (2014). *British National Corpus*. Ox.ac.uk. http://www.natcorp.ox.ac.uk/
- Sinclair, J. (1987). *Collins COBUILD English Language Dictionary*. HarperCollins.
- *UZH - International Corpus of English (ICE)*. (n.d.). Www.ice-Corpora.uzh.ch. https://www.ice-corpora.uzh.ch/en.html
- Wonner, B. (2007). *The Development of Corpus Linguistics to Its Present-day Concept*. GRIN Verlag

ملخص

هذه الدراسة تستهدف استكشاف الاختلافات اللغوية بين الأجيال في تيارت، مع التركيز على تنوع الكلام لدى الجيل الأكبر، بهدف الحفاظ على هذا التنوع من خلال نهج قائم على علم اللغة النصي. على الرغم من التحديات في تحويل البيانات إلى شكل رقمي بسبب القيود التكنولوجية، تشدد الجهود على ضرورة توفير أدوات متقدمة تدعم اللغات الأقل شيوعًا والتي تتمتع بأبحاث أقل معرفة مثل اللغة التيارتية. تتضمن التوصيات تعزيز الموارد التكنولوجية وتوافر البيانات للحفاظ على وتحليل ملامح اللهجات بشكل فعال. في النهاية، تسلط هذه البحوث الضوء على أهمية الحفاظ المستمر على التنوع اللغوي من أجل التراث الثقافي من خلال أساليب علم اللغة النصي

Abstract

This study targets the exploration of linguistic disparities between generations in Tiaret, focusing on the older generation's speech variety , with the ultimate aim of preserving it through a corpus based approach . Despite challenges in digitising data due to technological limitations, efforts emphasise the need for advanced tools supporting minority and less researched languages like the Tiaretian language . Recommendations include enhancing both technological resources and data availability to preserve and analyse dialectal features effectively. Ultimately, this research underscores the ongoing importance of preserving linguistic diversity for cultural heritage through innovative corpus linguistic methods.

Résume

Cette étude vise à explorer les disparités linguistiques entre les générations à Tiaret, en mettant l'accent sur la variété de discours de la génération plus âgée, dans le but ultime de la préserver à travers une approche basée sur un corpus. Malgré les défis liés à la numérisation des données en raison de limitations technologiques, les efforts soulignent la nécessité d'outils avancés soutenant les langues minoritaires et moins étudiées comme la langue tiarétienne. Les recommandations incluent l'amélioration des ressources technologiques et la disponibilité des données pour préserver et analyser efficacement les caractéristiques dialectales. En fin de compte, cette recherche souligne l'importance continue de préserver la diversité linguistique pour le patrimoine culturel à travers des méthodes innovantes de linguistique de corpus.