Arabic Language Characteristics that Make its Automatic Processing Challenging

Ilhem Boulesnam
Textual Linguistics and Discourse Analysis (LTAD)
Laboratory, Kasdi Merbah University, Algeria
boulesnam.ilhem@univ-ouargla.dz

Rabah Boucetti
Knowledge Engineering and Computer Security (ICOSI)
Laboratory, Abbes Laghrour University, Algeria
boucetti.rabah30@gmail.com

Abstract: *Arabic Natural Language Processing (ANLP) is an area of Artificial Intelligence (AI) that enables computers to process and understand Arabic text and speech. It powers applications such as translation, sentiment analysis, and speech recognition, opening new horizons. However, despite the worldwide popularity of Arabic, ANLP is lagging. The Arabic language possesses unique linguistic features that complicate computational processing. In this context, our article seeks to explore these challenges and examine Arabic's inherent characteristics in orthography, morphology, grammar, syntax, and linguistic diversity, which contribute to the complexities of its Natural Language Processing (NLP). Indeed, the main challenges are the diversity of dialects, the morphological and syntactic richness of Arabic, diglossia, and the absence of short vowels. In addition, the scarcity of Arabic resources further complicates NLP efforts. This review can serve as a guide for practitioners in the field of Arabic NLP, whether they are computer scientists or linguists. It also calls on the Arab community scientists to take steps to meet the potential challenges and increase efforts in the field to promote Arabic NLP.*

Keywords: *Computational linguistics, natural language processing, Arabic natural language processing, Arabic characteristics, Arabic language automation.*

Received July 15, 2024; accepted April 28 2025
https://doi.org/10.34028/iajit/22/4/14

1. Introduction

Nowadays, technological advances affect every area of human daily life. Traditional data processing, which relies exclusively on humans, is no longer relevant in the face of the efficiency brought by these new technologies. Language, as an important human activity, will not escape the rule. Using these technologies, and in particular Artificial Intelligence (AI) techniques, linguistic data will be processed automatically through what is called Natural Language Processing (NLP). It consists of developing computer programs based on AI algorithms capable of automatic translation, spelling and grammar checking, language didactic, and so on [41, 57, 80]. Progress is being made despite the major challenges, and research is continuing with the aim of achieving advanced levels of automatic human language processing, in which the machine becomes completely autonomous without human assistance [53, 82, 96].

Language can be defined as a set of rules or symbols where these symbols are combined and used to enable communication between human beings by transmitting or disseminating information [47, 64, 83]. The famous linguist Chomsky [35] says that "language is the inherent capability of native speakers to understand and form grammatical sentences. A language is a set of (finite or infinite) sentences, each finite length constructed out of a limited set of elements". Linguistic diversity is an important aspect of human culture and communication. Thousands of languages co-exist in the world. Although it is difficult to determine the exact number of languages, it is estimated that between 5,000 and 7,000 languages are spoken worldwide and around 300 writing systems [5, 99]. Some languages are spoken by just a few people, while others are spoken by millions or even billions, such as Mandarin Chinese, and English. Among the five most spoken languages in the world, is the Arabic language [61]. Arabic is the official language of 22 countries, the mother tongue of over 400 million speakers, and is considered to be the 4th most widely used language on the internet [30]. Arabic is also the Semitic language of the Quran, with a rich and complex morphology that differs significantly from Latin languages such as English and French. Furthermore, Arabic can be categorized into three main varieties [53]:

1. Classical Arabic (CA) which is the form of the Arabic language used in literary texts composed by early Arab scholars. The Quran can be considered as the highest form of CA texts.
2. Modern Standard Arabic (MSA), used for formal writing and conversation. It morphologically and syntactically resembles CA, although CA's style is much richer [87].
3. Arabic Dialects (AD) which are used in everyday life communication and informal exchanges. These dialects are mainly divided into Egyptian, Levantine, Gulf, Iraqi, Maghrebi, and others. They are

geographically distributed in the 22 countries of the Arab world [53].

For all these reasons and others, the Arabic language is of considerable importance and deserves to be a subject of study.

On the other hand, linguistics is the scientific study of language in its particular properties and characteristics in general. It is a multidisciplinary field that involves the study of sounds or phonology, word morphological formation, grammatical structure of sentences, meaning and semantics, language acquisition, language processing, and the relationship between language and other aspects of human behavior [93, 103]. Linguistics also encompasses several important subfields, among others, Computational Linguistics (CL) and NLP. While CL is more concerned with the theoretical and mathematical foundations of language processing, the NLP focuses on the practical implementation of algorithms and models for processing natural language. They both contribute to our understanding and development of language processing systems by analyzing and processing human language using mathematical models and algorithms. They aim to enable computers to understand statements or words written or spoken in human languages to facilitate human machine communication and to satisfy the desire to communicate with the computer using natural language. NLP can be categorized into two parts: Natural Language Understanding (NLU), which enables NLU and analysis by machines (extracting concepts, entities, emotions, keywords, etc.,); and Natural Language Generation (NLG), which evolves the task to make the machine understand and generate language. NLG is the process of producing expressions, sentences, and paragraphs that make sense [64].

While NLP research is making considerable progress, particularly in Latin languages such as English, the results increasingly show the importance of working with languages other than English. Given the importance of Arabic as a widely spoken language, extending NLP works to Arabic can lead to more inclusive and effective applications that address a large and diverse population and explore other cultures. However, Arabic NLP suffers, until today, from enormous delays despite the efforts made in the form of workshops, conferences, resource development, etc., [23, 52, 81]. Arabic NLP is particularly challenging due to several factors. It is known as a complex language, with a rich vocabulary, intricate grammar, and multiple dialects. To go further in this field, and give Arabic automatic processing a promising future, we first need to identify all the causes hindering its development. Then, the annotation of linguistic data is essential for creating diverse resources, designing and improving linguistic models, and prioritizing research to overcome these ambiguities and challenges. Our work is part of this approach, in which we will attempt to identify the various causes and specifications that may explain this delay. We will try to give researchers and developers, whether computer scientists or linguists, the main reasons and information they need to work with Arabic NLP.

The remainder of this document is organized as follows: Section 2 addresses the related work. Section 3 gives an outline of human languages and linguistics, which form the general context of our study. Section 4 describes CL and NLP. An overview of the Arabic language is given in section 5, while the specific features that make it a challenge for NLP will be reviewed in section 6. In section 7, the current situation of the Arabic NLP is discussed. Section 8 is devoted to the conclusion and perspectives.

2. Related Work

Arabic Natural Language Processing (ANLP) presents many challenges due to the language's complex morphology, intricate grammatical structures, dialectal variations, data limitations, and punctuation inconsistencies. These complexities significantly impact NLP tasks and require specialized approaches to improve performance and accuracy. Recently, the subject has become increasingly important and widely addressed, given the importance of both the Arabic language and automatic human language processing. In the following, we aim to provide a summary of relevant related work.

Morphologically and grammatically, Arabic is a highly inflected language, where a single root can generate several derived forms, making it difficult to recognize words and extract meaning. NLP in Arabic faces significant challenges due to the complexity of the language. Issues such as orthographic ambiguity, and morphological richness make processing Arabic particularly difficult [10, 79]. In addition, the richness of Arabic derivational processes contributes to polysemy, requiring advanced algorithms to accurately interpret context. The grammatical complexities of Arabic, such as subject-verb agreement, case marking, and non-concatenated morphology, further complicate automated text processing [72]. Ambiguous diacritics, lack of capitalization, broken plurals, and semantic intricacies are other challenges that complicate ANLP tasks [46].

Dialect variations and informal language are major challenges for Arabic NLP. The difficulties of NLP in Arabic stem from the significant differences between MSA and the multitude of regional dialects, which affect the consistency of text processing [10, 73]. In addition, many social media and conversational texts are written in dialects, and the informal nature of dialectal Arabic introduces ambiguity, making NLP tasks, like sentiment analysis, topic classification, and Named Entity Recognition (NER), particularly challenging [19].

Another issue hindering the progress of NLP in Arabic is the difficulty of text pre-processing and data limitations. Text preprocessing techniques, including normalization, tokenization, and stemming, are especially challenging to implement due to Arabic's unique character set and grammatical rules [76]. The lack of high-quality annotated datasets exacerbates these challenges. As a result, the effectiveness of machine learning approaches is limited. Studies have highlighted the need for better pre-processing techniques to improve the performance of Arabic NLP systems [1, 73].

The inconsistencies in Arabic punctuation, shaped by its distinct syntactic structures, create challenges for NLP applications, particularly in detecting sentence boundaries and predicting punctuation. The legibility and consistency of machine-generated text have been improved through the development of annotated datasets on Arabic punctuation [102].

However, recent advances in deep learning, particularly transformer-based models such as Arabic Bidirectional Encoder Representations from Transformers (AraBERT) [23], have shown significant improvements in Arabic NLP tasks, outperforming traditional models such as Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). In [73], a comparative study showed that transformer-based models achieve better accuracy in a variety of language processing tasks, including sentiment analysis and Machine Translation (MT).

A growing area of research is the integration of Arabic language skills into Large Language Models (LLMs). Recent studies have highlighted the need for appropriate datasets and models to improve the performance of Arabic LLMs, with a focus on security, dialect diversity, and cultural adaptation [9, 17, 26, 74]. Conversely, while progress on LLMs in Arabic is promising, issues such as the under-representation of dialects and the need for comprehensive security measures persist, suggesting that further research is crucial to realize the full potential of LLMs in Arab contexts.

After this brief review of the literature, we can see that no paper addresses the subject in its entirety, i.e., a document that deals with all the challenges of the Arabic language for its automatic processing. The morphological and grammatical challenges have been addressed by the authors of [10, 46, 72, 79]. The aspect of AD variation and the widespread use of informal Arabic is highlighted by the authors of [10, 19, 73]. In [1, 73, 76], the difficulties of the pre-processing of Arabic texts and the lack of linguistic data are discussed. The syntactic structures of the Arabic language were studied by Yagi *et al*. of [102]. Concerning the integration of LLMs in the Arabic language, Ashraf *et al*. [26] and Aloui *et al*. [17] have stated the need for appropriate data sets and models. To provide the scientific community with a single reference document, our article will try to summarize all these and other aspects and characteristics of the Arabic language that may pose a challenge to its automatic processing. It also provides some key statistics and findings that illustrate the current state of research and resources in this area.

3. Human Languages and Linguistic Science

3.1. Human Languages

With the primary purpose of satisfying the natural need for communication between humans, several languages come into existence. Language constitutes a communication form that is specific to human beings. It can evolve naturally or intentionally constructed, but in either case, its main characteristic is its use for communication between humans. It is used to express an infinite variety of meanings (thoughts, feelings, ideas, etc.,) in a complex and dynamic system that varies from one culture and region to another, by combining specific discrete units of sound, symbols, words, grammatical structures, and even gestures. Every language is a set of knowledge and skills that allow the speakers of the language to communicate with each other, to express ideas, assumptions, emotions, desires, and all other things that need to be expressed [85]. Among the human language characteristics is that the same concept can be represented with the same utility by different words. Sometimes these words are even very different from one language to another.

Although there may be historical links and connections between different languages, the absolute implication that they evolved from a single original language is not always true. Despite this, many studies have initially sought to trace the roots of language back to a single source [106]. It is interesting to note that human language is not always used to communicate with other humans and in some cases is incorporated into other languages. Taking the example of the language used to train and command animals, although it is inspired by human language, it cannot be used by both parties for full communication.

Like all aspects of human life, language evolves, over time, in a complex way and there are many different theories about how it developed [51]. Individuals acquire one or more languages during their childhood and can be considered native speakers of these languages [107]. There are about 5,000 to 7,000 different languages worldwide. They are distinguished in their vocabularies, their sounds, and their grammatical rules. Although there are many different languages, each has the same value for communicating thoughts, even if they do so in different ways, but their main purpose is the same in all human cultures [5, 91].

One of the most important human languages in the world, the Arabic language. Arabic is the mother tongue and the first language of over 400 million people in the

22 countries that make up the Arab world from the Atlantic Ocean to the Persian Gulf. It is also the fifth most spoken language in the world, and one of the six official languages of the United Nations [14, 36]. Arabic is indeed linked to Islam, as it is the language in which the Quran, the sacred book of Islam, was revealed. Muslims the world over use it for religious rituals and prayers. All Muslims pray five times a day and perform other rituals in Arabic, even if it's not their mother tongue, which shows the importance of the Arabic language worldwide [6].

Following all that has been said before, regarding the dynamism, complexity, and even the continuous evolution of human languages, it is impossible to define it completely, because the subject is still very controversial. Among others, the origin of languages and even how humans acquire languages are not fully understood, making the study of human languages a constantly evolving field. Furthermore, the study of human language (linguistics) is an interdisciplinary field that can encompass sociology, psychology, neuroscience, and computer science, among others [31, 32, 84].

3.2. Linguistic Science

As mentioned earlier, human language is a set of knowledge and skills that allow us to communicate with each other and express our ideas. Linguistics, also known as linguistic science, is the study of such knowledge and skills in all their aspects. It consists of the language's scientific study, including its structure, grammar, syntax, semantics, phonetics, and phonology. Linguists (experts in linguistics) dive into the analysis of the form, meaning, and context of language. They analyze grammar or syntax, semantics or the speech meaning, or how language is used in context. They use scientific procedures such as data gathering, investigation, and analysis. Empirical methods also are used to analyze linguistic data and thus provide objective and systematic descriptions and explanations of linguistic phenomena [67]. Indeed, the main focus of linguists is to comprehend the nature of language in general by asking questions such as:

- What are the common characteristics of all human languages?
- What is the relationship between language and other types of human behavior?
- What are the links between the modes of linguistic communication (speech, writing, sign language)?

Language is a multi-layered phenomenon, from the sounds produced by speakers to the meanings these sounds express, encompassing various aspects and sub-subfields. These subfields help us to understand the different dimensions of language and its use [89]. Most professional linguists specialize in one or more of them. The main ones are:

3.2.1. Phonetics

Phonetics is a linguistics branch that studies the sounds of human speech. It is concerned with the physical properties of speech sounds, how these sounds are produced and perceived by humans, and how they are represented in written language. Phonetics is concerned with the anatomy and physiology of the vocal tract, and how waves move through it to produce speech sounds. It is also interested in the acoustic properties of speech sounds, the way they are pronounced, and their classifications. Phonetics is an important study era that allows us to understand how language is produced and how it differs from one language to another [58, 108]. Phonetics experts' study both the production of speech sounds by the human speech organs (articulatory phonetics) and the sound features themselves (acoustic phonetics).

3.2.2. Phonology

Phonology is the branch of linguistics that is focused on how speech sounds are organized in a language and the basic rules that govern the combinations of these sounds. Phonology studies both the physical properties of speech sounds (phonetics) and the abstract relationships between these sounds (phonemes). The aim is to analyze the sound structure of a language and understand its underlying form. Simply put, phonology is the study of the sound features of languages and how they are used to communicate meaning [58].

3.2.3. Morphology

Morphology in linguistics is the study of how words are formed and their internal structure. It consists of analyzing the smallest units of meaning in language, called morphemes. In other words, how to distinguish between roots and affixes in the composition of words, and how they can be combined to create different words. Morphology is essential for understanding how words are structured or composed to convey meaning in language [24, 69, 100].

3.2.4. Syntax

Syntax is the set of principles, rules, and processes that shape the structure of sentences in a language. It is the study of how elementary words or sound units are combined to form larger units such as clauses, phrases, and sentences, and how these units are organized to create meaning. Syntax deals with issues such as the order of the words in sentences, how they are grouped to form sentences and clauses, and the relationship between different sentence parts. Overall, syntax is concerned with the grammatical structure of language, rather than with its meaning or expression. Syntax in linguistics is important for a good understanding of the differences between languages, as well as how language is learned and processed. It is a crucial element of linguistic analysis that helps us understand many other linguistic concepts such as semantics and morphology [8, 33, 70].

3.2.5. Semantics

Semantics is a linguistics branch that focuses on the sense of words, phrases, sentences, and larger units of speech. It is interested in how language users make meaning, and how they interpret and understand meaning in the context of their interactions. Semantics is associated with syntax, while syntax is concerned with the formal rules of language, semantics deals with the meaning that these rules transmit. Several levels of semantics can be differentiated, lexical semantics, which deals with the meaning of single words; phrasal semantics, which examines the meaning of phrases composed of several words; and sentential semantics, which examines the meaning of whole sentences [27, 50].

3.2.6. Pragmatics

Pragmatics is a discipline in linguistics that looks at how the context affects the meaning of a sentence. It interests in how speakers use language to transmit meaning and how listeners interpret that meaning in light of various contextual factors. Pragmatics goes beyond the literal meaning of words and sentences to take into account elements such as the social context. Pragmatics focuses on conversational implicatures, that is, what a speaker implies and what the listener infers [34, 68].

Furthermore, linguistics also encompasses other aspects such as how languages are formed and how they function, as well as, the historical and social contexts in which they are used and evolved, and the relationship between languages, cultures, and societies [25]. Consequently, linguistics is a highly interdisciplinary field that involves the study of language in a wide range of contexts [7, 40], such as sociolinguistics, where the language is studied in its relation to society [60]; psycholinguistics, language and its relation to the mind [92, 94]; and CL, which consists of studying processing language by computers [37, 64, 71].

4. Computational Linguistics and Natural Language Processing

CL is a sub-field of linguistics that combines the theories and results of linguistics in all its research areas, such as phonetics, phonology, syntax, semantics, etc., with computational techniques, mathematical models, and algorithms to study and analyze human language. This involves the scientific study of language and its underlying structure to understand its meaning and construction so that it can be modeled using computer models and algorithms. That is to say, CL applies computational techniques to linguistic data (speech and text) to achieve a better understanding of linguistic phenomena. It seeks to understand how humans use language and how this can be modeled and represented in machine-readable form, by designing models and algorithms to automate linguistic analysis tasks, such as sentence parsing, semantic information extraction, or language model generation [78, 90].

While CL aims to use mathematical methods and algorithms to better understand linguistic phenomena, NLP focuses on developing practical applications that enable computers to understand, interpret, and produce human language. It should be noted, however, that there is often an overlap between CL and NLP, as they are closely related fields. Whereas, NLP uses techniques from CL to develop models and algorithms to process human language. In another way, CLencompasses much more than NLP, as it also covers text mining, information extraction, and more [98]. NLP is an interdisciplinary field that combines computer science, CL results, and machine learning models to process human language. It is a sub-field of AI and computer science that deals with the automatic processing of linguistic data, which enables computers to understand, analyze, and generate human language. This allows us to bridge the gap between human communication and computer understanding, enabling and facilitating the way humans interact with machines [42].

Recently, with the increased interest in human-machine communications, NLP has seen rapid and remarkable progress thanks to technological advances, computing power, and the reliability of machine learning algorithms [96]. LLMs advanced AI models designed to process and generate human-like text based on deep learning techniques, particularly transformer architectures, where multi-head attention layers are stacked in a deep neural network. These models contain hundreds of billions (or more) of parameters. They are trained on large amounts of text data, enabling them to understand linguistic patterns, context, and semantics in a wide range of domains [105]. LLMs can learn from context and adapt to different tasks through fine-tuning or prompt engineering. Currently, LLMs such as GPT-4 [28], BERT [65], and Galactica [95] can perform a variety of NLP tasks, including text generation, translation, sentiment analysis, question answering, and summarization. These enable humans to interact with computers more naturally and intuitively, making the machine capable of processing and understanding natural languages, which opens the door to a variety of applications. There exists a wide range of real-life NLP applications, in the following we list some of the most common [64]:

- **Chatbots and Virtual Assistants**

NLP is used to develop intelligent chatbots and virtual assistants, both of which are AI-based software programs that interact with humans, capable of understanding and responding to human queries and commands. This is leading to the production of systems capable of enabling robots to interact with humans in natural languages, such as Google's assistant and Amazon's Alexa.

Arabic chatbots and virtual assistants are working to evolve and overcome language challenges. Major tech companies such as Google, Amazon, and Microsoft have introduced support for Arabic in their virtual assistants (Google Assistant, Alexa, and Cortana). There are also local initiatives to develop AI-enabled Arabic conversational agents using models such as AraBERT, enhancing chatbot accuracy in industries like banking and customer service (healthcare, and e-learning) [11, 23]. However, issues such as Arabic's rich morphology, ambiguous diacritics, numerous dialects, and lack of data are hindering the progress of Arabic chatbots and virtual assistants to make them more effective and accessible [15].

- **Information Extraction and Text Mining**

Information extraction is the process of outputting and extracting structured information from unstructured or semi-structured data sources, such as text documents or web pages. This operation involves NLP techniques to analyze and understand the content of the target source and identify relevant information such as named entities, relationships between them, and linked events. In many cases, extracting entities such as names, locations, events, dates, times, and prices is a powerful way to obtain information relevant to a user's needs.

Information extraction and text mining in Arabic have seen some progress in recent years, thanks to advances in NLP and AI techniques. Pre-trained models such as AraBERT and Collaborative Arabic Multi-dialectal Enhanced Language Bidirectional Encoder Representations from Transformers (CAMeL-BERT) have improved Arabic text mining, enabling better NER topic modeling [18, 43]. Arabic text mining is commonly applied in the financial, health, and government sectors for purposes such as fraud detection, opinion mining, and automated document processing. However, difficulties remain due to the complexity of the Arabic language, including its rich morphology, dialectal variations, and the lack of high-quality annotated datasets.

- **Text Summarization**

Summarizing long documents is one of the important NLP applications. It consists in condensing and summarizing a long text document into a short, concise text. Text summarization has become more accessible and efficient thanks to AI algorithms and online tools, which offer major advantages such as a significant reduction in reading time, key information extraction, and generating summaries for various purposes.

Automatic summarization of Arabic text has recently become more efficient thanks to the efforts and improvements made in the Arabic NLP. To improve summarization accuracy, researchers have developed extractive and abstractive summarization models using the AraBERT, T5, and Seq2Seq architectures [4, 46, 59, 101]. Arabic summarization is commonly used in news aggregation, legal document processing, and academic research, with the availability of datasets and the adaptation of AI models constantly improving. However, challenges remain due to Arabic language issues.

- **Speech Recognition**

Also known as Automatic Speech Recognition (ASR) or speech-to-text, which is the ability of a machine or program to identify words spoken aloud and convert them into readable text. It is a capability that enables a program to convert human speech into written form. This technology represents a giant step towards simple and practical communication between man and machine, which allows it to acquire great importance. Speech recognition plays a crucial role in voice assistants, dictation engines, language-learning applications, etc.

Arabic speech recognition has made significant strides. Several systems now support Arabic, including Google speech-to-text, International Business Machines (IBM) Watson, and Mozilla DeepSpeech. In addition, regional initiatives such as Qatar Computing Research Institute Automatic Speech Recognition (QCRI ASR) and Arabic speech corpus are making significant contributions to its development [21, 75]. However, dialect diversity, lack of labeled speech data, phonetic ambiguity, and other problems related to the Arabic language remain.

- **Machine Translation (MT)**

MT is considered one of the most important NLP applications, and its value has increased with Internet democratization. MT is the use of automated software to translate text from one language to another without human intervention. MT has remarkably facilitated communication between people of different languages and cultures. This has made it possible to draw on the immense wealth of knowledge in the different communities. It's a technology that has evolved, and with the advent of AI, Neural Machine Translation (NMT) is considered the most accurate and advanced approach. Nowadays, automatic translation engines enable translation from and into several languages, for example, Google's translation engine enables translation into over sixty languages. Nevertheless, despite all these advances, MT still has its limitations and can come up against complex linguistic structures and cultural nuances.

Advances in NMT and deep learning have been instrumental in improving Arabic MT. Arabic-English translations are remarkably reliable in most models, like Google Translate, Microsoft Translator, and Meta's No Language Left Behind (NLLB). Specialist models such as AraT5 and MARBERT enhance performance [3, 77].

Despite the problems associated with the Arabic language, research is underway into Arabic dialectal translation and context-aware models to improve accuracy and fluency.

- **Sentiment Analysis**

Also known as Opinion Mining. Sentiment analysis is an NLP technique used to analyze textual data to determine the feeling or emotion expressed in it. Sentiment analysis is an important business intelligence tool that helps companies improve their products and services. It involves analyzing large volumes of textual data, such as emails, transcripts of customer service discussions, social media comments, and reviews, to better understand customer attitudes and improve services. In politics, sentiment analysis is also used to find out people's tendencies towards a party or an individual.

Arabic sentiment analysis is advancing through the development of deep learning models such as AraBERT, MARBERT, and CAMeL-BERT, which try to improve the accuracy of emotion and opinion detection [22]. It monitors social media, analyzes customer feedback, and tracks political sentiment. Arabic sentiment analysis model training has been improved using annotated datasets such as Arabic Sentiment Analysis System (ArSAS) and Arabic Sentiment Tweets Dataset (ASTD) [49]. However, sentiment analysis in Arabic lags because of dialectal diversity, code-switching (mixing Arabic with English or French), and ambiguity due to the absence of diacritics.

These are just a few examples of the many NLP applications. The field is still evolving, new needs arise and new applications are being developed to solve various language-related challenges. While NLP applications in English have reached a satisfactory level, the need for other languages, such as Arabic, is growing. Given the recent technological advances in the Arab world and its openness to different cultures, the need for efficient tools for the automatic processing of Arabic will be of great importance.

5. Arabic Language Overview

Arabic language is a rich and diverse language with a complex grammatical structure. In light of the three key parameters morphology, syntax, and lexical blends, the Arabic language is composed of three main classes or types: CA, MSA, and Dialect Arabic (DA):

1. CA, also called Quranic Arabic (the language of the Quran, the holy book of Islam), is used in religious writings such as the Sunnah and Hadith, as well as in many ancient Arabic manuscripts.
2. MSA is the type of formal communication understood by the majority of Arabic speakers as the official language of the Arab world. MSA is the main language of the media, education, television, newspapers, and books. It is mainly based on the syntax, morphology, and phonology of CA, and can be transformed to accommodate new words that need to be created as in the field of science or technology.
3. The AD or "colloquial Arabic," in contrast, are the true native language forms. They are used only in informal everyday Arab communication, are not taught in schools, and are not standardized. Dialects are mainly spoken and not written. However, the situation is changing as more and more Arabs have access to electronic means of communication such as social media platforms, where people prefer to express themselves in their dialect. These ADs are less related to CA and differ from MSA. They are the result of interaction between different ancient CA dialects and other neighboring and/or colonial languages, for example, the Algerian dialect is deeply influenced by Berber and French. The Arabs do not consider MSA and DA to be two distinct languages despite the two variants having clear domains of prevalence: formal written MSA and informal spoken (dialect); which leads to a particular type of coexistence between the two forms of language. This type of situation is what linguistics calls diglossia [88]. ADs vary geographically throughout the Arab world and can be divided into:

   - Maghrebi Arabic: covers the dialects of Mauritania, Morocco, Algeria, Tunisia and Libya.
   - Egyptian Arabic: covers the dialects of the Nile Valley: Egypt and Sudan.
   - Levantine Arabic: includes dialects of Lebanon, Syria, Jordan, and Palestine.
   - Gulf Arabic: includes the dialects of Saudi Arabia (although it has a wide range of sub-dialects), Yemen, Kuwait, the United Arab Emirates, Bahrain, Oman, and Qatar.
   - Iraqi Arabic: dialects of Iraq and containing elements from the Levantine and Gulf.

The Arabic language is read and written from right to left. It is written using the Arab script, which is also used to write other languages around the world that are not related to Arabic, such as Persian and Kurdish. Arabic script uses two types of symbols to write words: Letters and diacritical marks [56].

- **Letters**

Arabic letters are written from right to left in cursive style whether printed or handwritten form, with no upper or lower case letters. They consist of two parts: the letter form and the letter mark. The letter form is an essential element of the letter. In the Arabic language, there is a total of 19 letter forms (Figure 1).

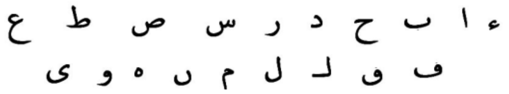


Figure 1. Arabic letter forms [56].

Letter marks, also known as consonant diacritical marks: Dots (one, two, or three to go above or to go below the letter form). The short Kaf, which is used to mark Kaf (كاف ك) letter shapes, and the Hamza (همزة ء) which can appear above or below particular letter forms. Note that the term Hamza is used for both the letter form and the letter mark. The letter mark Madda (آ) is considered a variant of Hamza (Figure 2).

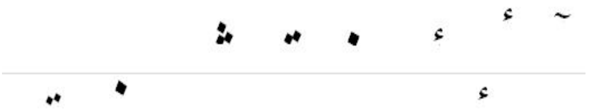
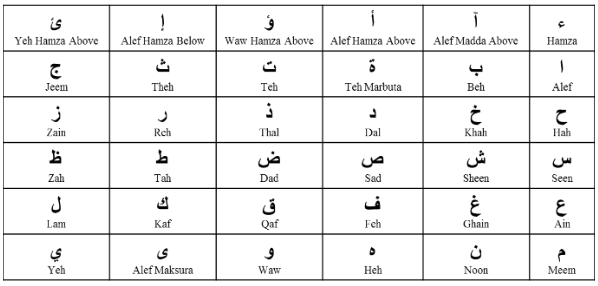


Figure 2. Arabic letter marks [56].

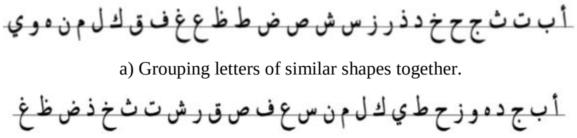
Combinations of letter forms and letter marks give the 36 letters of the Arabic alphabet used to write MSA. Note that, some letters are created using letter forms only, without any letter marks (Figure 3).



ئ Yeh Hamza Above	إ Alef Hamza Below	ؤ Waw Hamza Above	أ Alef Hamza Above	آ Alef Madda Above	ء Hamza
ج Jeem	ث Theh	ت Teh	ة Teh Marbuta	ب Beh	ا Alef
ز Zain	ر Reh	ذ Thal	د Dal	خ Khah	ح Hah
ظ Zah	ط Tah	ض Dad	ص Sad	ش Sheen	س Seen
ل Lam	ك Kaf	ق Qaf	ف Feh	غ Ghain	ع Ain
ي Yeh	ى Alef Maksura	و Waw	ه Heh	ن Noon	م Meem

Figure 3. Arabic letters [56].

Of the 36 Arabic letters used in MSA, the basic letters of the Arabic alphabet correspond to the 28 consonant sounds. They are formed using all letter forms except the Hamza letter form. In all these letters, the letter marks are fully discriminating and distinguish the different consonants from each other. There are two commonly used Arabic alphabetical sorts. The former is based on the latter shapes. It is called in Arabic (الفبائية) 'Alyfbaiyah'. It consists of grouping letters of similar shapes together (Figure 4-a)). The second called (أبجدية) 'Abjadiyah', is loosely based on the ancient Phoenician alphabetical order. This order is mainly used for enumerating small lists, as is done in Arabic dictionaries, where words are generally listed in groups sorted in Abjad order (Figure 4-b)).



أ ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي

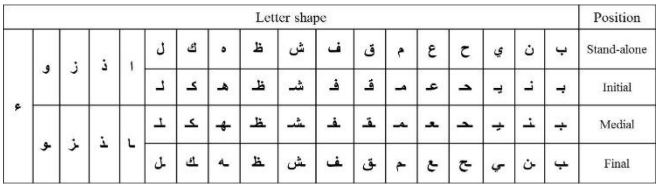
a) Grouping letters of similar shapes together.

أ ب ج د ه و ز ح ط ي ك ل م ن س ع ف ص ق ر ش ت ث خ ذ ض ظ غ

b) Abjad order.

Figure 4. Arabic's alphabetical sorting order.

Depending on their position in the word: initial, medial, final, or stand-alone, Arabic letters are not written in the same way, but take different shapes. The shapes of the initial and medial letters are generally similar, as are the final and stand-alone shapes. Most letters in a word are written in a fully connected form. A few letters are connected to the preceding letters but not to the following ones; they are either initial or standalone. Small white spaces follow disconnected letters, creating visually isolated islands of connected letters, called word parts. Figure 5 illustrates the different shapes of some Arabic letters according to their position in the word.



Letter shape																		Position
ء	و	ز	ذ	ا	ل	ك	ه	ظ	ش	ف	ق	م	ع	ح	ي	ن	ب	Stand-alone
					لـ	كـ	هـ	ظـ	شـ	فـ	قـ	مـ	عـ	حـ	يـ	نـ	بـ	Initial
	ـو	ـز	ـذ	ـا	ـلـ	ـكـ	ـهـ	ـظـ	ـشـ	ـفـ	ـقـ	ـمـ	ـعـ	ـحـ	ـيـ	ـنـ	ـبـ	Medial
					ـل	ـك	ـه	ـظ	ـش	ـف	ـق	ـم	ـع	ـح	ـي	ـن	ـب	Final

Figure 5. Arabic letters with their different shapes [56].

- **Diacritics**

Diacritical marks (تشكيل) 'Tashkeel', make up the second class of Arabic script symbols. These signs are used in Arabic script to indicate various linguistic features, such as vowels and pronunciation, to help readers overcome the ambiguity of certain words. These diacritical marks play a crucial role in clarifying the meaning and pronunciation of Arabic letters. While letters are always written, diacritics are optional: written Arabic can be fully diacritized, as in religious and educational texts for children; partially diacritized; or entirely non-diacritized. Diacritic is formed by diacritical marks above or below a consonant (letter mark) to give it a sound.

Diacritical marks are divided into three groups: Vowel, Nunation, and Shadda. Vowel diacritics represent the three short vowels of Arabic: 'Fatha' (فتحة), 'Damma' (ضمة), and 'Kasra' (كسرة), and no vowel 'Sukun' (سكون). The nunation can only appear in the final position of words in nominal nouns (nouns, adjectives, and adverbs). Nunation diacritics look like a doubled version of their corresponding short vowels. Shadda is a double consonant diacritic, usually combined with a vowel or nunation diacritic (Figure 6 shows the different diacritical marks with the latter ب).

Diacritical marks are divided into three groups: Vowel, Nunation, and Shadda. Vowel diacritics represent the three short vowels of Arabic: 'Fatha' (فتحة), 'Damma' (ضمة), and 'Kasra' (كسرة), and no vowel 'Sukun' (سكون). The nunation can only appear in the final position of words in nominal nouns (nouns, adjectives, and adverbs). Nunation diacritics look like a doubled version of their corresponding short vowels. Shadda is a double consonant diacritic, usually combined with a vowel or nunation diacritic (The Figure 6 shows the different diacritical marks with the latter ب).

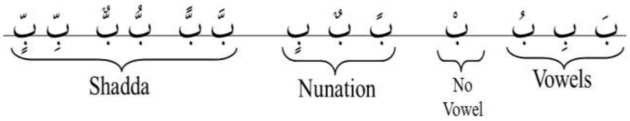


Figure 6. Arabic diacritic marks [56].

On the whole, Arabic is a widely spoken language with a rich history and culture. It is known for its complexity and richness, which makes it challenging for many non-native speakers to learn. Furthermore, its grammar is so complex that it is even difficult for Arabic speakers to master. If Arabic presents these specific challenges for learners, they will certainly make it more difficult to process automatically. These challenges will contribute negatively to the development of effective Arabic NLP applications.

6. Specific Arabic NLP Challenges

After more than fifty years of research to develop techniques enabling computers to automatically process natural languages, the field has reached a reasonable level of maturity, particularly for the English language. This is due to the widespread use of English as a world language and the extensive research carried out in English-speaking countries. Although English NLP has a more mature ecosystem and a significant advantage, other languages are trying to catch up to varying degrees. When comparing English and Arabic NLP, the latter is at least a decade behind. We can attribute this to several factors, some are linked to the Arabic language itself, others rather to the shortcomings of the works in question [29].

Arabic is an extremely twisted language with a complicated structure in terms of orthography, morphology, grammar, and syntax. It has many specific features that make its NLP challenging. In what follows, we will try to summarize the main ones.

6.1. Orthographic Features

As we've seen, the shape of an Arab letter can be modified depending on whether it's at the beginning, middle, or end of the word; linked to a preceding and following letter, or simply linked to a preceding. This already constitutes a non-standardization that needs to be taken into account when designing Arabic NLP applications. On the other hand, Arabic has no capital letters, unlike many Latin-based languages such as English and French, in which the capital letter is an orthographic marker showing that a word is a phrase's beginning or a Named Entity (NE). Without that spelling marker, proper nouns derived from adjectives are a source of ambiguity: for example, the same word "إلهام" can mean both a girl's first name and inspiration. This makes it all the more challenging to recognize Named Entities in Arabic, such as names of people, places, or things. Moreover, the absence of capital letters can make the use of acronyms unhelpful and insignificant. This can lead to confusion, particularly when translating, extracting information, or naming entity processing in general [48].

Also, Arabic includes a set of orthographic symbols, called diacritical marks (short vowels), that indicate the correct pronunciation of words. These marks can completely change the sense and meaning of a word. For instance, the non-vowel word (بر) supports several alternatives such as (بِرْ), which means obedience, kindness, and virtue; ((بَرْ meaning land or ground; and (بُرْ) that means wheat. Overcoming ambiguities requires knowledge of the context in which the word is introduced. Note that the non-vowel orthography is the standard form of the Arabic language. MSA is generally devoid of diacritics, and restoring or generating them is an additional and crucial task for Arabic NLP [44].

Arabic script, likewise, includes three other long vowels (حروف العلة): "أ" (alef), "و" (waw), and "ي" (yaa). A word that contains a vowel letter fluctuates depending on its position in the sentence and does not keep a stable state. Sometimes we find them fixed, sometimes we remove them, and sometimes diacritical marks are prescribed to prevent them from appearing, either through incapacity or heaviness. The defective (weak) verb (الفعل المعتل), a well-known case in MSA, is any verb whose root has a long vowel as one of its three radicals. These verbs will undergo vowel changes when conjugated. This specificity of long vowels in Arabic gives rise to two forms of spelling: with or without vocalization. Automation systems must remedy this situation [20].

Hamza (الهمزة ء) spelling is also subject to numerous and changing rules, making it difficult to master, even for Arabic speakers. The letter Hamza has several spelling forms (ؤ ,أ ,ء , or ئ), which is called "seat of Al-Hamza", depending on its diacritical mark and that of the letter preceding it. When Hamza arrives at the beginning of the word, it is written on or below an "Alef" (إ ,أ) according to its diacritic sign. Moreover, there is an exception to this rule when it concerns Hamzat El Kata (همزة القطع) or Hamzat El Wasl (همزة الوصل). The former is always written and pronounced, while the latter is neither written nor pronounced, and a "bare Alef" (ا) is used. Distinguishing the type is itself a challenge, the rule is to add the conjunction "and" (و) before and see if it is pronounced; therefore, written. Otherwise, when the Hamza appears in the middle or at the end of a word and is written on its seat (ؤ ,أ , or ئ), the choice of the right one follows the vowels hierarchy in the Arabic language: the Kasra, the Dama, and then the Fatha which has the lowest priority. Determining Al-Hamza orthographic variants is of notorious complexity and therefore requires special handling in the elaboration of the computational system [104].

All these orthographic features among others show clearly the inconsistency of Arabic orthographic rules, making its automatic processing more difficult. The solution is generally to rely much more on contextual analysis, which is no easy task either.

6.2. Morphological Features

One of the well-known properties of Arabic is that it is very rich and morphologically complex. Its vocabulary can be flexibly extended using roots and morphological samples. Arabic words are generally formed from roots. The root refers to a basic three or four-consonant word that represents a central meaning or concept. Most of them consist only of three consonants and a large number of words can be derived from each root by adding vowels and/or other consonants. The root system is a unique concept specific to the Arabic language, allowing to expansion of the vocabulary. Compared to English, Arabic words are highly inflected. While English relies heavily on affixation, Arabic is based on a morphology that is both concatenative (affixes and stems) and templatic (root and patterns). The most complicated situation is that of words that can represent an entire sentence. These are generally the longest words in the Arabic dictionary. They are made up of a root and hidden and/or connected pronouns. For instance, the word أَنُلْزِمُكُمُوهَا)) can be translated into English as the following sentence: Do we oblige you to do so? The word consists of: the root (لزم), the subject doing the action (نـ), then the first object (كم), and the second object (ها); and the interrogative hamza (أ) is added at the beginning to become a complete sentence in one word: (أنلزمكموها ؟). Distinguishing each part according to its role in the word is a challenge [39].

Another aspect of the morphology of Arabic words is that they generally depend on the sentence, and it is difficult to draw a line between the word and its sentence. This leads to morpho-syntactic structures according to inflection, declension, and clitics. The syntactic relationship of a word to other ones in the sentence is indicated by its inflectional endings, not by alternative words. For example, in the sentence, (وبخ الأب إبنه) which means: The father scolded his son. The suffixed pronoun (ـه) in the word (إبنه), refers to the word (الأب). The morpho-syntactic structure makes it possible to incorporate a large number of additions to a word, hence the richness of the language's vocabulary, and it would be difficult to pinpoint all the possible cases [16].

In MSA, a word can be made up by joining two words together (can be nouns, verbs, or particles), which is called annexation. In general, the meaning of the word compound is clear, insofar as its meaning corresponds to that of the parts assembled. As in the word (برمائية), which is made up of the two words: (بر) "land" and (مائية) "water", to refer to amphibians. However, the absence of a rule governing annexation to form compound words adds another challenge to the automatic processing of the Arabic language [97].

The Arabic language is highly derivational, and inflection leads to a high degree of fluctuation in morphology, which complicates modeling and, consequently, computational processing will be more challenging.

6.3. Syntactic Features

Syntactically, Arabic is generally considered to be a free-word language, with different word order patterns in a sentence. This syntactic flexibility makes it possible to express oneself in a variety of ways and to adapt sentences to different contexts and communication objectives. In Arabic, a sentence can have four types of word order; all are considered correct and convey the same meaning: Subject-Verb-Object (SVO), Verb-Subject-Object (VSO), Verb-Object-Subject (VOS), and Object-Verb-Subject (OVS). This flexibility in word order is a feature of Arabic syntax that further contributes to the richness and even to the complexity of the language. Therefore, generating or understanding sentences for different Arabic NLP applications will be a challenge [62].

Anaphora resolution is another syntactic issue that Arabic NLP applications have to address. Anaphora consists of referring nouns in the sentence by particular entities or pronouns. Almost all NLP applications require a successful mechanism for identifying and resolving anaphora. Nevertheless, this task is classically recognized as a very difficult problem in language processing, especially in Arabic. It is a complex task that requires a great deal of time and effort for NLP systems to understand and resolve references to earlier or later words in a discourse. Without finding the appropriate antecedent of the anaphora, the meaning of the sentence would not be fully and correctly understood. One of the most common types of anaphora in Arabic is the pronominal anaphora. It could be a third personal pronoun, which is known in Arabic as (ضمير الغائب). It has an empty semantic structure and has no meaning independently of its antecedent or main subject. The Arabic pronoun does not make the linguistic distinction between the human pronoun (he/she for example in English) and the non-human pronoun (it), which may cause ambiguity and external knowledge is required to correctly identify the antecedent. Another very common type of anaphora in Arabic is the hidden or zero anaphora (الضمير المستتر). It occurs when there is no entity acting as a subject. Zero anaphora can be determined easily by the human mind, but it can represent a great challenge for automated systems [38].

Agreement is also a syntactic feature that must be taken into account in NLP systems. In Arabic, an adjective follows generally the noun it describes in terms of number, gender, case, etc. Nevertheless, it also depends on the word order and can be total or partial, which leads to exceptions. For example, in the SVO order, the verb agrees with the subject in terms of number and gender. On the other hand, verbs in VSO order agree with the subject only in gender. In the auxiliary sentence, if the auxiliary comes before the subject, it agrees in gender only, while the verb agrees in both gender and number with the subject. However, if the subject precedes the auxiliary, the auxiliary and the verb both agree in gender and number. On the other hand, when it comes to agreement between numbers and countable nouns, there is a complex set of rules for determining the literal number that agrees with the counted noun. The literal generation of numbers depends on the expression of the noun counted in the sentence. Agreement in Arabic is either total or partial and sensitive to word order in the sentence; several cases are envisaged, adding further challenges to Arabic NLP [54].

6.4. Diglossia and Dialectal Variation

Diglossia refers to a linguistic situation in which two distinct varieties of a language are used in a community, high and low variety. While the high variety is used in formal contexts, the low one is used in everyday informal conversations. In the case of Arabic, diglossia is a well-known phenomenon. Further, the term diglossia was originally used to describe the linguistic situation in Arabic-speaking countries. In the Arab world, we find MSA, a high variety used in formal contexts such as literature, newspapers, and education; and various spoken colloquial dialects, low varieties used in everyday communication. In addition, Arabic speakers have a wide range of dialects that vary considerably from one another. Each dialect has its own vocabulary and pronunciation. We can add to all this variety, a non-standard romanization called 'Arabizi', in which the Arabic text is written by mixing Latin characters, numbers, and some punctuation marks, mainly used by Arab Internet users on social networks, Short Messaging System (SMS), and discussion forums. All these varieties will constitute a challenge and will add further difficulties to NLP applications which have to deal with both MSA and its different dialects [45].

6.5. Challenges for Large Language Models Application

LLMs face significant challenges when applied to Arabic due to the language's complex morphology, dialectal diversity, and specific writing issues. As mentioned earlier, the richness of the Arabic inflectional and derivational system makes word segmentation and contextual interpretation difficult. This involves developing tokenization methods specific to Arabic, attention mechanisms, or pre-training objectives that take into account Arabic's properties [86]. The presence of several dialects, many of which have insufficient annotated data, limits the model's performance beyond MSA. Furthermore, the scarcity of high-quality labeled datasets, even for MSA, further limits the training and fine-tuning of Arabic NLP models. Arabic's writing and orthographic characteristics, such as the absence of capital letters, the ambiguity of diacritics, and the shape of letters according to context, pose additional difficulties for tokenization [12]. The right-to-left writing system of Arabic also leads to formatting and processing problems in NLP models, which are primarily designed for languages written from left to right. In addition, certain dialects and communities may be misrepresented due to biases in the training data. Most Arabic-specific models lag behind their English counterparts due to smaller datasets and high training costs, leading to computational limitations that further hamper the development of the Arabic LLMs [26].

It's worth noting that the challenges mentioned above are not an exhaustive list, and there may be additional challenges specific to certain Arabic NLP tasks or applications.

7. Current Situation and Discussion

Despite all the challenges mentioned above, Arabic ANLP has experienced significant growth and development in recent years. Work and concerted efforts to promote the field are in full swing, and extensive linguistic resources and models adapted to the unique characteristics of the Arabic language have been created. Key statistics and information illustrating the current state of research and resources, including:

1) Conferences and shared tasks

- The 2nd ArabicNLP conference [55]: held in August 2024, this conference presented eight shared tasks, attracting a total of 79 papers, 8 review papers, and 71 system descriptions. Two shared tasks were particularly noteworthy: Arabic Financial NLP (AraFinNLP), which received 9 papers, dealing with applications in the financial field; and News Media Narratives of the Israel War on Gaza (FIGNEWS), which received 17 papers, highlighting the interest in dealing with news media content.
- The 1st Arabic Natural Language Understanding Shared Task (ArabicNLU) [63]: this task focused on Word Sense Disambiguation (WSD) and Location Mention Disambiguation (LMD). Of the 38 teams registered, 3 participated in the final evaluation.

2) Language resources

- 101 billion Arabic words dataset [17]: released in April 2024, this is the largest Arabic corpus to date, compiled from common crawl Web Extracted Text (WET) files. It has been rigorously cleaned and de-duplicated to ensure data quality and provides a substantial resource for training authentic Arabic language models.
- ArabicaQA dataset [2]: introduced in March 2024, it contains 89,095 answerable questions and 3,701 non-answerable questions, as well as additional labels for open-domain questions. The ArabicaQA dataset fills the gaps in resources for machine reading comprehension of Arabic and answering questions in the open domain.

3) Model development

Development of models designed for the Arabic language, like ArabianGPT and Arabic Compact Language Model (ACLM). ArabianGPT [66] Launched in February 2024, this is a range of models based on transformers specially designed for Arabic. The models, with parameters ranging from 0.1 to 0.3 billion, use the AraNizer tokenizer to handle the complex morphology of Arabic. Fine-tuned versions have significantly improved, with sentiment analysis accuracy reaching 95%. ACLM [13] is a small, efficient language model adapted for MSA. It is built using the 135-million-parameter AraGPT2 base model. Thanks to rigorous pre-training on carefully selected datasets, ACLM has achieved impressive linguistic capabilities. It offers a compact and efficient solution that bridges the gap between the high resource requirements of large models and practical needs, representing a significant advance in Arabic NLP.

All these works and others show that even though the challenges and specifications of the Arabic language have an impact on the immaturity of NLP applications in Arabic, progress can be made. We strongly believe that the main reason for this delay is the lack of research and development. Although there has been some advancement in recent years, Arabic NLP still lags behind other languages. The unavailability of tools and resources indicates that there is still work to be done for further progress. Limited research and development efforts have had the effect of restricting the number of tools, models, and datasets available for NLP tasks in Arabic such as annotated corpora, and lexicons. This lack of resources hinders the development of robust NLP applications for Arabic and limits the accuracy and performance of existing ones. To meet these challenges, we need to develop high-quality Arabic datasets, improve tokenization techniques, reduce biases, and exploit multilingual models such as mBERT and GPT-4 to improve NLP performance in Arabic. The absence of a willingness on the part of the Arab community to promote research on Arabic NLP, by launching serious projects on the subject with the involvement of qualified human resources, remains the weak link in its development. Computer scientists and linguists alike are called upon to step up their efforts, working in close collaboration to tackle the challenges mentioned above. As far as we know, there is no single, standard way of developing an IT system, whatever the field. Each domain is a separate project, with its own rules and exceptions that need to be dealt with. If English has reached this level of maturity, it is thanks to the research and efforts made in the field by the English-speaking community.

8. Conclusions

Arabic as a language is both challenging and interesting given its widespread use throughout the world. Despite this, its NLP still suffers from shortcomings compared to other languages. In this article, we have tried to look at the main causes and linguistic characteristics of Arabic that hinder its NLP development. We have tried to go deeper into the basics of word and sentence structure, as well as the relationships between sentence elements. Arabic is a phonetic language, in the sense that there is a direct correspondence between the letters and the sounds with which they are associated. An Arabic word requires changes in the letter shape depending on its position in the word, and there is no notion of capital letters. It does not dedicate letters as vowels but uses diacritical marks instead. In general, MSA texts are not diacritized since short vowels are optional. This can lead to ambiguity and difficulty in analyzing Arabic words. Morphologically, the structure of the words is both rich and complex, so that they can represent a complete phrase or sentence. Identifying and distinguishing each part according to its role in the word is a challenge. In syntactic terms, the Arabic sentence is complex, with a free word order in which the constituents of the sentence can be interchanged without affecting the meaning. Resolving anaphora is also a challenge, as contextual knowledge is required, which increases syntactic and semantic ambiguity and necessitates a more complex analysis. To illustrate these and other challenges, descriptive examples in MSA are given, in the hope that readers will appreciate the complexity associated with Arabic NLP.

Consequently, Arabic differs from other languages because of its complex and ambiguous structure. Its NLP presents unique and specific challenges. These challenges underline the need for specialized tools and techniques to deal with the unique features of the Arabic language in the NLP domain. Researchers and developers are called upon to work actively to find possible solutions.

References

[1] Aarab A., Oussous A., and Saddoune M., "Optimizing Arabic Information Retrieval: A Comprehensive Evaluation of Preprocessing Techniques," *in Proceedings of the IEEE 12th International Symposium on Signal, Image, Video and Communications*, Marrakech, pp. 1-4, 2024, DOI:10.1109/ISIVC61350.2024.10577827

[2] Abdallah A., Kasem M., Abdalla M., Mahmoud M., Elkasaby M., Elbendary Y., and Jatowt A., "ArabicaQA: Comprehensive Dataset for Arabic Question Answering," *in Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Washington (DC), pp. 2049-2059, 2024. https://doi.org/10.1145/3626772.3657889

[3] Abdul-Mageed M., Elmadany A., and Nagoudi E., "ARBERT and MARBERT: Deep Bidirectional Transformers for Arabic," *in Proceedings of the 59th Annual Meeting of the Association for*

Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online, pp. 7088-7105, 2020. DOI: 10.18653/v1/2021.acl-long.551

[4] Abu Nada A., Alajrami E., Al-Saqqa A., and Abu-Naser S., "Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach," *International Journal of Academic Information Systems Research*, vol. 4, no. 8, pp. 6-9, 2020. http://ijeais.org/wp-content/uploads/2020/8/IJAISR200802.pdf

[5] Africa A., Lamdagan R., and Lacanilao J., "Audio-based Assessment in Determining Language," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 7, pp. 2984-2988, 2020. https://doi.org/10.30534/ijeter/2020/16872020

[6] Ahmed A., Ali N., Alzubaidi M., Zaghouani W., Abd-alrazaq A., and Househ M., "Freely Available Arabic Corpora: A Scoping Review," *Computer Methods and Programs in Biomedicine Update*, vol. 2, pp. 100049, 2022. https://doi.org/10.1016/j.cmpbup.2022.100049

[7] Akbulut F., "A Study on Interdisciplinary Nature of Translation Studies," *Journal of Language Research*, vol. 6, no. 1, pp. 45-56, 2022. https://doi.org/10.51726/jlr.1193899

[8] Akhmanova O. and Mikaeljan G., *The Theory of Syntax in Modern Linguistics*, Mouton, 1969. https://api.pageplace.de/preview/DT0400.9783112414668_A44908655/preview-9783112414668_A44908655.pdf

[9] Al Ghanim M., Almohaimeed S., Zheng M., Solihin Y., and Lou Q., "Jailbreaking LLMs with Arabic Transliteration and Arabizi," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Florida, pp. 18584-18600, 2024. https://aclanthology.org/2024.emnlp-main.1034/

[10] Al Moaiad Y., Alobed M., Alsakhnini M., and Momani A., "Challenges in Natural Arabic Language Processing," *Edelweiss Applied Science and Technology*, vol. 8, no. 6, pp. 4700-4705, 2024. https://doi.org/10.55214/25768484.v8i6.3018

[11] Al-Ghadhban D. and Al-Twairesh N., "Nabiha: An Arabic Dialect Chatbot," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 3, pp. 452-459, 2020. https://www.researchgate.net/publication/340403610_Nabiha_An_Arabic_Dialect_Chatbot

[12] Ali A., Siddiqui M., Algunaibet R., and Ali H., "A Large and Diverse Arabic Corpus for Language Modeling," *Procedia Computer Science*, vol. 225, pp. 12-21, 2022. https://doi.org/10.1016/j.procs.2023.09.086

[13] Alkaoud M., Alsaqoub M., Aljodhi I., Alqadibi A., and Altammami O., "ACLM: Developing a Compact Arabic Language Model," *The International Arab Journal of Information Technology*, vol. 22, no. 3, pp. 535-546, 2025. https://doi.org/10.34028/iajit/22/3/9

[14] Almars A., "Attention-based Bi-LSTM Model for Arabic Depression Classification," *Computers, Materials and Continua*, vol. 71, no. 2, pp. 3091-3106, 2022. https://doi.org/10.32604/cmc.2022.022609

[15] Almurayh A., "The Challenges of Using Arabic Chatbot in Saudi Universities," *IAENG International Journal of Computer Science*, vol. 48, no. 1, pp. 1-12, 2021. https://www.iaeng.org/IJCS/issues_v48/issue_1/IJCS_48_1_21.pdf

[16] Alothman A. and Alsalman A., "Arabic Morphological Analysis Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 214-222, 2020. file:///C:/Users/user/Downloads/Arabic_Morphological_Analysis_Techniques.pdf

[17] Aloui M., Chouikhi H., Chaabane G., Kchaou H., and Dhaouadi C., "101 Billion Arabic Words Dataset," *arXiv Preprint*, vol. arXiv:2405.01590v1, pp. 1-15, 2024. https://doi.org/10.48550/arXiv.2405.01590

[18] Alsaleh A., Althabiti S., Alshammari I., Alnefaie S., et al., "LK2022 at Qur'an QA 2022: Simple Transformers Model for Finding Answers to Questions from Qur'an," *in Proceedings of the OSACT Workshop, ELRA European Language Resources Association*, Marseille, pp. 120-125, 2022. https://aclanthology.org/2022.osact-1.14.pdf

[19] AL-Sarayreh S., Mohamed A., and Shaalan K., "Challenges and Solutions for Arabic Natural Language Processing in Social Media," *in Proceedings of the International Conference on Business Intelligence and Information Technology, Smart Innovation, Systems and Technologies*, Harbin, pp. 293-302, 2023. https://doi.org/10.1007/978-981-99-3416-4_24

[20] Alshaari M., Modern Standard Arabic Speech Recognition: Using Formants Measurements to Extract Vowels from Arabic Words' Consonant-Vowel-Consonant-Vowel Structure, Doctoral Theses, Florida Institute of Technology, 2020. https://repository.fit.edu/etd/858

[21] Alsharhan E., Ramsay A., and Ahmed H., "Evaluating the Effect of Using Different Transcription Schemes in Building a Speech Recognition System for Arabic," *International Journal of Speech Technology*, vol. 25, no. 1, pp. 43-56, 2022. https://link.springer.com/article/10.1007/s10772-020-09720-z

[22] Alturayeif N., Luqman H., Alyafeai Z., and

Yamani A., "StancEval 2024: The First Arabic Stance Detection Shared Task," *in Proceedings of the 2nd Arabic Natural Language Processing Conference*, Bangkok, pp. 774-782, 2024. DOI: 10.18653/v1/2024.arabicnlp-1.88

[23] Antoun W., Baly F., and Hajj H., "AraBERT: Transformer-based Model for Arabic Language Understanding," *in Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, Marseille, pp. 9-15, 2020. https://aclanthology.org/2020.osact-1.2/

[24] Aronoff M. and Fudeman K., *What is Morphology?*, John Wiley and Sons, 2022. https://www.wiley.com/en-ie/What+is+Morphology%3F%2C+3rd+Edition-p-9781119715221

[25] Aronoff M. and Rees-Miller J., *The Handbook of Linguistics*, John Wiley and Sons, 2017. DOI:10.1002/9781119072256

[26] Ashraf Y., Wang Y., Gu B., Nakov P., and Baldwin T., "Arabic Dataset for LLM safeguard Evaluation," *in Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics, Human Language Technologies*, New Mexico, pp. 5529-5546, 2025. https://aclanthology.org/2025.naacl-long.285/

[27] Ataboyev I. and Turgunova F., "The Concept of Semantic Field in Linguistics," *ACADEMICIA: An International Multidisciplinary Research Journal*, vol. 12, no. 3, pp. 319-324, 2022. DOI :10.5958/2249-7137.2022.00223.3

[28] Baktash J. and Dawodi M., "GPT-4: A Review on Advancements and Opportunities in Natural Language Processing," *Journal of Electrical Electronics Engineering*, vol. 2, no. 4, pp. 548-549, 2023. DOI: 10.33140/JEEE.02.04.19

[29] Bashir M., Azmi A., Nawaz H., Zaghouani W., Diab M., Al-Fuqaha A., and Qadir J., "Arabic Natural Language Processing for Qur'anic Research: A Systematic Review," *Artificial Intelligence Review*, vol. 56, no. 7, pp. 6801-6854, 2023. https://doi.org/10.1007/s10462-022-10313-2

[30] Boudad N., Faizi R., Thami R., and Chiheb R., "Sentiment Analysis in Arabic: A Review of the Literature," *Ain Shams Engineering Journal*, vol. 9, no. 4, pp. 2479-2490, 2018. https://doi.org/10.1016/j.asej.2017.04.007

[31] Boyd R. and Schwartz H., "Natural Language Analysis and the Psychology of Verbal Behavior: The Past, Present, and Future States of the Field," *Journal of Language and Social Psychology*, vol. 40, no. 1, pp. 21-41, 2021. https://doi.org/10.1177/0261927X20967028

[32] Bragg D., Koller O., Bellard M., and Berke L., et al., "Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective," *in Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, Pittsburgh, pp. 16-31, 2019. https://doi.org/10.1145/3308561.3353774

[33] Brown K. and Miller J., *Syntax: A Linguistic Introduction to Sentence Structure*, Routledge, 2020. https://doi.org/10.4324/9781003070702

[34] Cheng D., "Corpus Linguistics for Pragmatics: A Guide for Research, Written by Ruhlemann, Christoph," *Contrastive Pragmatics*, pp. 1-5, 2022. DOI:10.1163/26660393-bja10064

[35] Chomsky N., *Aspects of the Theory of Syntax*, The MIT Press, 2014. https://www.scribd.com/document/543469281/Aspects-of-the-Theory-of-Syntax-by-Noam-ChomskyAJ

[36] Chouikhi H., Chniter H., and Jarray F., "Arabic Sentiment Analysis Using Bert Model," *in Proceedings of the 13th International Conference, Advances in Computational Collective Intelligence*, Rhodes, pp. 621-632, 2021. https://doi.org/10.1007/978-3-030-88113-9_50

[37] Church K. and Liberman M., "The Future of Computational Linguistics: on beyond Alchemy," *Frontiers in Artificial Intelligence*, vol. 4, pp. 1-18, 2021. https://doi.org/10.3389/frai.2021.625341

[38] Dahou A., Abdelmoazz M., and Cheragui M., "A3C: Arabic Anaphora Annotated Corpus," *in Proceedings of the 4th International Conference on Natural Language and Speech Processing*, Trento, pp. 147-155, 2021. https://aclanthology.org/2021.icnlsp-1.17/

[39] Darwish K., Habash N., Abbas M., and Al-Khalifa H., et al., "A Panoramic Survey of Natural Language Processing in the Arab World," *Communications of the ACM*, vol. 64, no. 4, pp. 72-81, 2021. https://doi.org/10.1145/3447735

[40] Dohma U., "Cognitive Linguistics with its Theoretical Aspects," *Uludag Universitesi Fen-Edebiyat Fakultesi Sosyal Bilimler Dergisi*, vol. 23, no. 43, pp. 1235-1259, 2022. https://doi.org/10.21550/SOSBILDER.1037676

[41] Dubey S., Shukla O., and Tiwari S., "Analysis of Application of Natural Language Processing in Artificial Intelligence," *International Journal of Mechanical Engineering*, vol. 7, no. 5, pp. 419-421, 2022. https://kalaharijournals.com/resources/Special_Issue_April_May_55.pdf

[42] Eisenstein J., *Introduction to Natural Language Processing*, The MIT Press, 2019. https://mitpress.mit.edu/9780262042840/introduction-to-natural-language-processing/

[43] El-Alami F., El Alaoui S., and Nahnahi N., "Contextual Semantic Embeddings based on Fine-Tuned AraBERT Model for Arabic Text Multi-

Class Categorization," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 10, pp. 8422-8428, 2022. https://doi.org/10.1016/j.jksuci.2021.02.005

[44] Elgamal S., Obeid O., Kabbani T., Inoue G., and Habash N., "Arabic Diacritics in the Wild: Exploiting Opportunities for Improved Diacritization," *in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, pp. 14815-14829, 2024. https://aclanthology.org/2024.acl-long.792/

[45] Elnagar A., Yagi S., Nassif A., Shahin I., and Salloum S., "Systematic Literature Review of Dialectal Arabic: Identification and Detection," *IEEE Access*, vol. 9, pp. 31010-31042, 2021. DOI: 10.1109/ACCESS.2021.3059504

[46] Elsaid A., Mohammed A., Ibrahim L., and Sakre M., "A Comprehensive Review of Arabic Text Summarization," *IEEE Access*, vol. 10, pp. 38012-38030, 2022. DOI:10.1109/ACCESS.2022.3163292

[47] Fabbro F., Fabbro A., and Crescentini C., "The Nature and Function of Languages," *Languages*, vol. 7, no. 4, pp. 1-10, 2022. https://doi.org/10.3390/languages7040303

[48] Faizullah S., Ayub M., Hussain S., and Khan M., "A Survey of OCR in Arabic Language: Applications, Techniques, and Challenges," *Applied Sciences*, vol. 13, no. 7, pp. 1-27, 2023. https://doi.org/10.3390/app13074584

[49] Farha I. and Magdy W., "A Comparative Study of Effective Approaches for Arabic Sentiment Analysis," *Information Processing and Management*, vol. 58, no. 2, pp. 102438, 2021. https://doi.org/10.1016/j.ipm.2020.102438

[50] Frawley W., *Linguistic Semantics*, Routledge Lawrence and Francis Group, 1992. https://api.pageplace.de/preview/DT0400.9781135441708_A23802583/preview-9781135441708_A23802583.pdf

[51] Freeman D., "Arguing for a Knowledge-Base in Language Teacher Education, then (1998) and Now (2018)," *Language Teaching Research*, vol. 24, no. 1, pp. 5-16, 2020. https://doi.org/10.1177/1362168818777534

[52] Ghaddar A., Wu Y., Bagga S., and Rashid A., "Revisiting Pre-Trained Language Models and Their Evaluation for Arabic Natural Language Processing," *in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, pp. 3135-3151, 2022. DOI: 10.18653/v1/2022.emnlp-main.205

[53] Guellil I., Saadane H., Azouaou F., Gueni B., and Nouvel D., "Arabic Natural Language Processing: An Overview," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 5, pp. 497-507, 2021. https://doi.org/10.1016/j.jksuci.2019.02.006

[54] Habash N., Bouamor H., and Chung C., "Automatic Gender Identification and Reinflection in Arabic," *in Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*, Florence, pp. 155-165, 2019. https://doi.org/10.18653/v1/W19-3822

[55] Habash N., Bouamor H., Eskander R., and Tomeh N., et al., "Proceedings of the Second Arabic Natural Language Processing Conference," *in Proceedings of the 2nd Arabic Natural Language Processing Conference*, Bangkok, pp. 1-17, 2024. https://aclanthology.org/2024.arabicnlp-1.0/

[56] Habash N., *Introduction to Arabic Natural Language Processing*, Springer Nature, 2010. https://doi.org/10.1007/978-3-031-02139-8

[57] Huang X., Zou D., Cheng G., Chen X., and Xie H., "Trends, Research Issues and Applications of Artificial Intelligence in Language Education," *Educational Technology and Society*, vol. 26, no. 1, pp. 112-131, 2023. https://doi.org/10.30191/ETS.202301_26(1).0009

[58] Islomov D., "Phonetics and Phonology," *Middle European Scientific Bulletin*, vol. 11, no. 1, pp. 575-579, 2021. https://core.ac.uk/download/pdf/480517092.pdf

[59] Ismail Q., Alissa K., and Duwairi R., "Arabic News Summarization based on T5 Transformer Approach," *in Proceedings of the 14th International Conference on Information and Communication Systems*, Irbid, pp. 1-7, 2023. DOI:10.1109/ICICS60529.2023.10330509

[60] Johnson E. and White K., "Developmental Sociolinguistics: Children's Acquisition of Language Variation," *WIREs Cognitive Science*, vol. 11, no. 1, pp. e1515, 2020. https://doi.org/10.1002/wcs.1515

[61] Julian G., "What are the Most Spoken Languages in the World," Online, pp. 1-15, 2020. http://tony-silva.com/eslefl/miscstudent/downloadpagearticles/mostspokenlangs-fluentin3months.pdf

[62] Kaddoura S., Ahmed R., and Jude Hemanth D., "A Comprehensive Review on Arabic Word Sense Disambiguation for Natural Language Processing Applications," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 4, pp. e1447, 2022. https://doi.org/10.1002/widm.1447

[63] Khalilia M., Malaysha S., Suwaileh R., Jarrar M., Aljabari A., Elsayed T., and Zitouni I., "ArabicNLU 2024: The First Arabic Natural Language Understanding Shared Task," *in Proceedings of the 2nd Arabic Natural Language Processing Conference*, Bangkok, pp. 361-371, 2024. https://aclanthology.org/2024.arabicnlp-1.30/

[64] Khurana D., Koli A., Khatter K., and Singh S., "Natural Language Processing: State of the Art, Current Trends and Challenges," *Multimedia*

Tools and Applications, vol. 82, no. 3, pp. 3713-3744, 2023. https://doi.org/10.1007/s11042-022-13428-4

[65] Koroteev M., "BERT: A Review of Applications in Natural Language Processing and Understanding," *arXiv Preprint*, vol. arXiv:2103.11943v1, pp. 1-18, 2021. https://doi.org/10.48550/arXiv.2103.11943

[66] Koubaa A., Ammar A., Ghouti L., Najar O., and Sibaee S., "ArabianGPT: Native Arabic GPT-based Large Language Model," *arXiv Preprint*, vol. arXiv:2402.15313v2, pp. 1-21, 2024. https://doi.org/10.48550/arXiv.2402.15313

[67] Kremmel B. and Harding L., "Towards a Comprehensive, Empirical Model of Language Assessment Literacy Across Stakeholder Groups: Developing the Language Assessment Literacy Survey," *Language Assessment Quarterly*, vol. 17, no. 1, pp. 100-120, 2020. https://doi.org/10.1080/15434303.2019.1674855

[68] Leech G., *Principles of Pragmatics*, Routledge, 2016. https://doi.org/10.4324/9781315835976

[69] Levesque K., Breadmore H., and Deacon S., "How Morphology Impacts Reading and Spelling: Advancing the Role of Morphology in Models of Literacy Development," *Journal of Research in Reading*, vol. 44, no. 1, pp. 10-26, 2021. https://doi.org/10.1111/1467-9817.12313

[70] Matchin W. and Hickok G., "The Cortical Organization of Syntax," *Cerebral Cortex*, vol. 30, no. 3, pp. 1481-1498, 2020. DOI:10.1093/cercor/bhz180

[71] Matthiessen C., Wang B., Ma Y., and Mwinlaaru I., *Systemic Functional Insights on Language and Linguistics*, Springer Singapore, 2022. https://doi.org/10.1007/978-981-16-8713-6_5

[72] Mohamed Ali H. and Mostafa M., "Challenges Related to Grammatical and Morphological Processing of Arabic Texts by Means of Artificial Intelligence," *Journal of Electrical Systems*, vol. 20, no. 6s, pp. 1366-1380, 2024. https://doi.org/10.52783/jes.2918

[73] Mohamed M. and Alosman K., "A Comparative Study of Deep Learning Approaches for Arabic Language Processing," *Jordan Journal of Electrical Engineering*, vol. 11, no. 1, pp. 18-34, 2024. https://doi.org/10.5455/jjee.204-1711016538

[74] Mousi B., Durrani N., Ahmad F., and Hasan M., et al., "AraDiCE: Benchmarks for Dialectal and Cultural Capabilities in LLMs," *in Proceedings of the 31st International Conference on Computational Linguistics*, Abu Dhabi, pp. 4186-4218, 2025. https://aclanthology.org/2025.coling-main.283/

[75] Mubarak H., Hussein A., Chowdhury S., and Ali A., "QASR: QCRI Aljazeera Speech Resource-A Large Scale Annotated Arabic Speech Corpus," *in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Online, pp. 2274-2285, 2021. https://aclanthology.org/2021.acl-long.177/

[76] Nafea A., Muayad M., Majeed R., Ali A., Bashaddadh O., Khalaf M., Sami A., and Steiti A., "A Brief Review on Preprocessing Text in Arabic Language Dataset: Techniques and Challenges," *Babylonian Journal of Artificial Intelligence*, vol. 2024, pp. 46-53, 2024. https://doi.org/10.58496/BJAI/2024/007

[77] Nagoudi E., Elmadany A., and Abdul-Mageed M., "TURJUMAN: A Public Toolkit for Neural Arabic Machine Translation," *in Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, Marseille, pp. 1-11, 2022. https://aclanthology.org/2022.osact-1.1/

[78] Nelson L., Burk D., Knudsen M., and McCall L., "The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods," *Sociological Methods and Research*, vol. 50, no. 1, pp. 202-237, 2021. https://doi.org/10.1177/0049124118769114

[79] Nouhaila B., Habib A., Abdellah A., and Abdelhamid I., "Assessing the Impact of Static, Contextual and Character Embeddings for Arabic Machine Translation," *Journal of Information and Knowledge Management*, vol. 23, no. 2, pp. 2450009, 2024. https://doi.org/10.1142/S0219649224500096

[80] Nozza D., Passaro L., and Polignano M., "Preface to the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI)," *in Proceedings of the 6th Workshop on Natural Language for Artificial Intelligence, Co-Located with the 21st International Conference of the Italian Association for Artificial Intelligence*, Udine, pp. 1-5, 2022. https://ceur-ws.org/Vol-3287/

[81] Obeid O., Zalmout N., Khalifa S., Taji D., Oudah M., Alhafni B., Inoue G., Eryani F., Erdmann A., and Habash N., "CAMeL Tools: An Open-Source Python Toolkit for Arabic Natural Language Processing," *in Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, pp. 7022-7032, 2020. https://aclanthology.org/2020.lrec-1.868/

[82] Otter D., Medina J., and Kalita J., "A Survey of the Usages of Deep Learning for Natural Language Processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 604-624, 2020. DOI:10.1109/TNNLS.2020.2979670

[83] Oviogun P. and Veerdee P., "Definition of Language and Linguistics: Basic Competence," *Macrolinguistics and Microlinguistics*, vol. 1, no.

1, pp. 1-12, 2020. https://doi.org/10.21744/mami.v1n1.1

[84] Peniro R. and Cyntas J., "Applied Linguistics Theory and Application," *Linguistics and Culture Review*, vol. 3, no. 1, pp. 1-13, 2019. DOI:10.21744/lingcure.v3n1.7

[85] Pereltsvaig A., *Languages of the World*, Cambridge University Press, 2020. https://books.google.jo/books/about/Languages_of_the_World.html?id=ucjlEAAAQBAJ&redir_esc=y

[86] Qarah F. and Alsanoosy T., "A Comprehensive Analysis of Various Tokenizers for Arabic Large Language Models," *Applied Sciences*, vol. 14, no. 13, pp. 1-17, 2024. https://doi.org/10.3390/app14135696

[87] Sharaf A. and Atwell E., "QurAna: Corpus of the Quran Annotated with Pronominal Anaphora," *in Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, pp. 130-137, 2012. http://www.lrec-conf.org/proceedings/lrec2012/pdf/123_Paper.pdf

[88] Shendy R., "The Limitations of Reading to Young Children in Literary Arabic: The Unspoken Struggle with Arabic Diglossia," *Theory and Practice in Language Studies*, vol. 9, no. 2, pp. 123-130, 2019. http://dx.doi.org/10.17507/tpls.0902.01

[89] Shparberg A., "Linguistics Database," *The Charleston Advisor*, vol. 23, no. 4, pp. 30-32, 2022. https://doi.org/10.5260/chara.23.4.30

[90] Smith N., *Linguistic Structure Prediction*, Springer Nature, 2022. https://doi.org/10.1007/978-3-031-02143-5

[91] Staib M., Teh T., Torresquintero A., Mohan D., Foglianti L., Lenain R., and Gao J., "Phonological Features for 0-Shot Multilingual Speech Synthesis," *arXiv Preprint*, vol. arXiv:2008.04107, pp. 2942-2946, 2020. https://doi.org/10.48550/arXiv.2008.04107

[92] Sterling J., Jost J., and Bonneau R., "Political Psycholinguistics: A Comprehensive Analysis of the Language Habits of Liberal and Conservative Social Media Users," *Journal of Personality and Social Psychology*, vol. 118, no. 4, pp. 805-834, 2020. DOI:10.1037/pspp0000275

[93] Tasheva N., "Exploring the Rich Tapestry of Linguistics: A Comprehensive Overview," *Science and Innovation in the Education System*, vol. 2, no. 11, pp. 51-57, 2023. https://doi.org/10.5281/zenodo.10006452

[94] Tatlılıoglu K. and Senchylo-Tatlilioglu N., "Language Development at Early Childhood: An Overview in The Context of Psycholinguistics," *in Proceedings of the 16th Scientific and Practical Conference on Psycholinguistics in a Modern World*, Pereiaslav, pp. 283-288, 2021. https://doi.org/10.31470/2706-7904-2021-16-283-288

[95] Taylor R., Kardas M., Cucurull G., and Scialom T., et al., "Galactica: A Large Language Model for Science," *arXiv Preprint*, vol. arXiv:2211.09085v1, pp. 1-58, 2022. https://doi.org/10.48550/arXiv.2211.09085

[96] Torfi A., Shirvani R., Keneshloo Y., Tavaf N., and Fox E., "Natural Language Processing Advancements by Deep Learning: A Survey," *arXiv Preprint*, vol. arXiv:2003.01200v4, pp. 1-23, 2020. https://doi.org/10.48550/arXiv.2003.01200

[97] Torjmen R. and Haddar K., "Tunisian Dialect Agglutination Processing with Finite Transducers," *Computacion y Sistemas*, vol. 26, no. 3, pp. 1215-1223, 2022. https://doi.org/10.13053/cys-26-3-4344

[98] Tsujii J., "Computational Linguistics and Natural Language Processing," *in Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*, Tokyo, pp. 52-67, 2011. https://dl.acm.org/doi/10.5555/1964799.1964806

[99] Vinson R., *Language, Culture and Society*, Online, 2022. https://www.bibliotex.com/explore;searchText=Randy%20Vinson;mainSearch=1;themeName=Default-Theme

[100] Vocroix L., "Morphology in Micro Linguistics and Macro Linguistics," *Macrolinguistics and Microlinguistics*, vol. 2, no. 1, pp. 1-20, 2021. https://doi.org/10.21744/mami.v2n1.11

[101] Wazery Y., Saleh M., Alharbi A., and Ali A., "Abstractive Arabic Text Summarization based on Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, pp. 1-14, 2022. https://onlinelibrary.wiley.com/doi/10.1155/2022/1566890

[102] Yagi S., Elnagar A., and Yaghi E., "Arabic Punctuation Dataset," *Data in Brief*, vol. 53, pp. 110118, 2024. https://doi.org/10.1016/j.dib.2024.110118

[103] Yule G., *The Study of Language*, Cambridge University Press, 2017. https://archive.org/details/georgeyulethestudyoflanguage2017cambridgeuniversitypress/page/n27/mode/2up

[104] Zalmout N., Erdmann A., and Habash N., "Noise-Robust Morphological Disambiguation for Dialectal Arabic," *in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, pp. 953-964, 2018. https://aclanthology.org/N18-1087/

[105] Zhao W., Zhou K., Li J., and Tang T., et al., "A Survey of Large Language Models," *arXiv Preprint*, vol. arXiv:2303.18223v16, pp. 1-144,

2023. https://doi.org/10.48550/arXiv.2303.18223

[106] Zlatev J., Zywiczynski P., and Wacewicz S., "Pantomime as the Original Human-Specific Communicative System," *Journal of Language Evolution*, vol. 5, no. 2, pp. 156-174, 2020. DOI:10.1093/jole/lzaa006

[107] Zokirov M. and Dadabayeva S., "About the Role of Languages Contacts in the Development of Languages," *Theoretical and Applied Science*, vol. 84, no. 4, pp. 687-691, 2020. DOI:10.15863/TAS.2020.04.84.118

[108] Zokirov M. and Zokirova S., "On Researching Phonetic Level of the Languages," *GIS Business*, vol. 15, no. 6, pp. 148-154, 2020. https://gisbusiness.org/index.php/gis/article/view/20223

Ilhem Boulesnam obtained her Master's degree in 2011 and her Ph.D. in 2018 from the Department of Linguistics at Ben Youssef Benkheda University-Algiers 2, Algeria. Specialized in linguistics, she is currently a lecturer at Kasdi Merbah University-Ouargla, Algeria. Her areas of research are Applied Linguistics, Computational Linguistics, Language Programming, Didactics and Translation.

Rabah Boucetti received his Ph.D. in 2022, in Artificial Intelligence from the Math and Computer Science Department of Abbes Laghrour University-Khenchela, Algeria. His research areas include Machine Learning and Data Analysis, Internet of Things, IoT Services Composition, IoT Services discovery, Natural Language Processing, and Optimization.