# The DAGC-ATS Database for Arabic Grammar Correction for Arabic Summaries

Nada Essa[1], Mostafa. M. EL-GAYAR[2], Eman M. El-Daydamony[3]
Department of Information Technology-Faculty of Computers and Information,
Mansoura University, Mansoura 35516, Egypt[1, 2, 3]
Department of Computer Science, Arab East Colleges, Riyadh 11583, Saudi Arabia[2]

***Abstract*—Arabic Grammar Correction is a comprehensive open-domain. Modern methods of correcting Arabic language errors rely on a database specific to a particular field and containing specific words and phrases, which leads to the emergence of the problem of out-of-context words. Due to the growth in recent work for Arabic text summarization and Arabic grammar correction, out-of-context words and the complex nature of Arabic grammar, an open-domain Arabic database is a required resource for Arabic language processing techniques. In this study, A new open-domain Database for Arabic Grammar Correction (DAGC-ATS) is presented to solve the out-of-context words problem, limited domain existing databases for training. The proposed database is based on the description of Arabic grammar using part-of-speech tags and relations between words by a dependency parser. The DAGC-ATS database is based on grammar error detection and correction at the simple sentence level. The database contains entries that describe Arabic grammar rules. The DAGC-ATS database contains two files, one for correcting Arabic simple sentences and the other for correcting incorrect Arabic basic sentences in grammar. It is designed for use only in the training stage. Every entry in the database describes one different grammatical problem, such as gender, number, singular, dual, or plural faults. It contains 9309888 entries. Using the QALB dataset, the system's precision, recall, and F-measure scores were 96.9, 94.80, and 95.83. Additionally, the same system was tested using the EASC database with 785 summaries, and the results for precision, recall, and F-measure were 99.73%, 95.90%, and 97.77%.**

***Keywords*—Arabic grammar correction; Arabic natural processing; open domain database**

## I. Introduction

The Arabic Natural Language Processing (NLP) community has recently given considerable attention to correcting grammatical errors in text, although most of this research has focused on English [1]. The Arabic language is particularly important and complex. The Arabic language is structured and inflected, making error diagnosis and challenging feedback tasks. Arabic Grammar Error Correction (AGEC) has attracted increased interest due to notable studies and encouraging shared challenges [2]. The difficulties faced by AGEC include the prevalent morphology, the scarcity of corpora, and the complexity of Arabic grammar. Spelling error detection and correction are essential for various applications and natural language processing tasks, such as optical character recognition, search query processing, parsing, machine translation, and intelligent tutoring systems. Grammar error correction has two main stages: determination of grammar errors and correction of grammar errors. Most databases for AGEC are for specific data and domains [3]. A spelling error correction system contains two primary modules. The first is detecting errors within the second, which is text, and correcting those errors. The simplest approach to detection is to look up the primitive form of the input word in a lexicon. If the word is not found in the lexicon, it is considered an ill-formed word that flags a spelling error. The error correction module generates a list of candidate words and ranks them in a way that could be considered as corrections for the erroneous word. Grammatical error correction is a challenging task, and existing methods have limitations that involve post-editing [4]. Recent work in the Arabic language has increased, especially for Arabic text summarization [5],[6],[7] and Arabic grammar correction [8],[9],[10]. Most databases for Arabic grammar correction consist of specific Arabic entries for training and testing for certain domains [11]. In other words, the data used in one database is different from that in another database. So, the need to produce a new open database to check errors in Arabic writing has become necessary, which can be used in any specific domain or data. This study presents a new database that describes the meaning and structure of Arabic sentences and is designed to correct Arabic grammar for simple Arabic sentences. In other words, the summarized text contains simple sentences.

## II. Related Works

Because of Arabic's rich vocabulary, syntactic flexibility, and morphological complexity, Arabic presents a special challenge for tasks involving natural language processing. According to the lack of standard databases for AGER and the complexity of Arabic, Recent work for AGER is based on training and testing data for a specific domain. Ahlam et al used ChatGPT to create an Arabic corpus named "Tibyan" for grammatical error correction. ChatGPT is a data augmentation tool that compares two Arabic sentences with grammatical errors to a sentence that is error-free and taken from Arabic books. These sentences are known as guide sentences. The Tibyan corpus was created in a number of steps, including gathering and pre-processing two Arabic texts from diverse sources, including open-access corpora and books. After that, ChatGPT is used to create a parallel corpus using the previously gathered text as a model for building sentences with various kinds of errors. To ensure the automatically generated sentences were accurate and error-free, they hired linguistic specialists to examine and validate them. The corpus is improved and validated. The accuracy of the corpus is increased through iterative validation and refinement based on input from linguistic experts. Lastly, they examined several types of errors in the Tibyan corpus using the Arabic Error [22] Type Annotation tool (ARETA). The seven categories of errors found in our corpus were orthography, morphology, syntax, semantics, punctuation, merge, and split. There are 600,000 tokens in the Tibyan corpus [12]. Chouaib Moukrim et al introduced a syntactic error correction system that is based on automatically producing proper Arabic sentences. To produce all possible syntactically correct sentences, they first extracted the words from the sentence under consideration. This is done using an ontology that logically describes the rules of Arabic grammar. There is no good corpus that contains annotations on several levels of Arabic grammar, which firstly leads to an annotation of a new reference corpus containing 360 sentences. Of the 360 Arabic sentences [13]. Zainab Althafir et al presented an Arabic Bidirectional Encoder Representations from Transformers (AraBERT) and a rule-based system. Seventy million sentences gathered from two distinct sources were used to train the model. The first consists of 5 million articles with 1.5 billion words gathered from 10 news sources. The other is the 3.1 million articles from 31 news sources, which make up the Open-Source International Arabic News (OSIAN) corpus. Also, the model used the Qatar Arabic Language Bank (QALB) corpus to create an Arabic grammatical auto-corrector for non-native speakers [14]. The corpus consists of 622 sentences with spelling and grammar mistakes made by non-native speakers [15]. Sarah AlOyaynaa et al suggested a novel study for AGER using pre-trained language models based on transformers. They suggested refined language models that were based on previously Arabic GED using two main approaches: level and sentence level, using language models AraBERT and M-BERT. Several Arabic datasets are public. accessible were used for fine-tuning. To enrich the GEC domain, they first used SCUT, a publicly accessible dataset with two columns named source (i.e., incorrect erroneous sentences) and targets (i.e., correct sentences) [16]. Aiman Solyman et al created the dataset to overcome the dearth of Arabic-language resources for the GEC task. A large-scale parallel synthetic corpus based on the confusion function was created in an unsupervised manner. The Al-Watan corpus, a direct access collection of 10,000,000 (ten million) words composed by professional journalists in Modern Standard Arabic (MSA), served as the data source. 18,061,610 million words, or a portion of this data, were extracted, and the confusion function. They reformatted the dataset to make it appropriate for the intended task because there isn't a lot of data pertaining to it. There were two stages to the reformatting process: first, a new column called "label" was added, followed by the target being categorized as correct (i.e., 1) and the source as erroneous (that is, 0). And a subset of the primary dataset was used for that. It was then used during the fine-tuning process and saved in a new file. The final dataset was then divided into two datasets: the test dataset, which is used to evaluate the model's performance, and the training dataset, which is used for the fine-tuning step. For train-validation datasets, the splitting was applied at random at a percentage of 90–10 [17]. Aiman Solyman et al suggested a GEC model for low-resource languages like Arabic that is based on the seq2seq Transformer. They suggested a technique for creating synthetic parallel data to get around the bottleneck caused by the absence of a corpus. Additionally, inspired by the success of capsule networks in computer vision, they dynamically aggregated data across layers in Arabic GEC using the Expectation-Maximization routing algorithm. The datasets they use to train the models are real and synthetic. Initially, the synthetic set was created using the 15,001,707-sentence monolingual news Arabic corpus OSIAN. Nevertheless, the technique outlined in Section 4.2 was used to generate synthetic data, which included 1,500,173 sentence pairs for the development set and 11,833,758 sentence pairs for the training set. These were employed to retrain the L2RAGEC and R2L models. Additionally, the QALB-2014 corpus, which included 1,017 examples for development and 19,411 examples for training, was used for fine-tuning. QALB-2014 data source gathered from the Arabic Learners Written Corpus, the Aljazeera news platform. Additionally, Google Translate is used to translate English Wikipedia articles into Arabic. Native speakers have corrected and annotated the source data. By dividing each unknown word into several sub-words, the Byte Pair Encoding (BPE) technique was used to overcome the challenge of rare and unknown words [18]. SOMAIA MAHMOUD presented an economical and effective method for automatically scoring Arabic essays that emphasize grammatical assessment. Essays can also be graded using the suggested method according to additional standards like prompt adherence, organization, and resemblance to the model response. The method uses various parameter-efficient techniques and makes use of the pre-trained AraBART model. For each criterion, they used two datasets. The QALB dataset is the first publicly accessible dataset for correcting grammatical errors in Arabic. It is divided into two sections: QALB-2014, which included 21.3K comments from native Arabic speakers on the Aljazeera news website, and QALB-2015, which included 622 essays by Arabic language learners. For a variety of grammatical errors, including spelling, punctuation, word choice, morphology, syntax, and dialectal usage, every text in the dataset has been manually corrected. Training, development, and test sets make up the dataset. There are two test sets in QALB-2015: one for L1 texts composed by native speakers and another for L2 texts composed by Arabic L2 learners [19]. Nizar Habash et al presented the ZAEBUC dataset as the second. It includes 214 essays authored by first-year students at Zayed University in the United Arab Emirates [20]. From the previous, it is concluded that existing dictionaries or databases are for a specific domain and a limited number of words. There is no AGER database that can be used for any dataset or domain. A comparison between the most popular databases used in AGEC is shown in Table I.

TABLE I. A Comparison Between Mostly Used Arabic Databases for AGEC

| Database Name | Domain | Statistics |
|---|---|---|
| QALB-2014 [20] | The Aljazeera news | 1,017 examples for development and 19,411 examples for training |
| QALB-2015 [21] | Essays by Arabic language learners | 622 essays |

| | | |
|---|---|---|
| Tibyan corpus [12] | Sentences from various sources | 600,000 tokens |
| OSIAN [18] | Monolingual news Arabic corpus | 15,001,707-sentence |
| Watan corpus [17] | It is composed of professional journalists in Modern Standard Arabic (MSA) | 10,000,000 (ten million) words written |
| The ZAEBUC dataset [20] | It is authored by first-year students at Zayed University in the United Arab Emirates | 214 essays |

## III. Proposed Framework

Because the database relies on a simple sentence structure, the description of each entry in the database includes a set of characteristics and the relationship of words to each other. So, the different word types, the basic, complex, and advanced sentence structures, and grammatical rules of the Arabic language used in this database are discussed in the following sentences.

### A. Pronunciation of Arabic Letters

The basic elements of the Arabic language are word, clause, and sentence. There are four grammatical states in Arabic [23]. These cases are nominative, accusative, genitive, and jussive. Any word that can express its meaning without the aid of another word is a noun. People, places, objects, concepts, and more are all named with nouns. There are two genders in Arabic: feminine and masculine. Unless they contain a feminine ending, most nouns are regarded as male. To convert male nouns and adjectives into feminine, the most common feminine ending is the taa marbuta letter. Geographical names, bodily parts that occur in pairs, and various other nouns are examples of terms that are conventionally regarded as feminine. Some nouns, on the other hand, can be classified as either masculine or feminine. The nouns in Arabic are divided into three main groups according to their numbers. The following are these groups: singular, dual, and plural. In a sentence, verbs are used to express the activity. There must be a verb in every phrase \ cite [24]. Verbs in Arabic grammar must be conjugated according to the plurality and gender of the individual carrying out the activity. Arabic verbs have three different tenses: imperfect, perfect, and imperative. Adjectives are used to characterize nouns in Arabic. Here is a quick summary of the four basic guidelines for utilizing adjectives. Prepositions are a group of words used in Arabic to show the link between other words [25]. Their purpose is to make clear where an action or item is in relation to another. It is important to remember that no matter where a preposition appears in a phrase, its pronunciation remains the same. The reason for this is that they are regarded as "Mabni" terms. In addition, every phrase that comes after a preposition ends in a Kasrah. A pronoun is a term that can take the place of a noun in Arabic. Subject pronouns, object pronouns, and genitive pronouns are three dissimilar categories of overt or conspicuous pronouns in Arabic [26]. Subject pronouns, which can be attached or used alone, function as the verb's doer or subject. Conversely, object pronouns are always connected and function as the verb's object. Finally, genitive pronouns function as either the object of the associated noun (possessive) or the object of the preposition. Words like this and those are examples of demonstrative pronouns in Arabic. Unlike many other languages, they are classified as demonstrative pronouns proximal or distal, and they also vary in terms of femininity or masculinity, as well as whether they are singular, dual, or plural. Relative pronouns are examples of how the pronouns "that, which, who, whom" are portrayed. To link nouns, pronouns, and verbs with other nouns or verbs, these pronouns are used as conjunctions\subsection {Sentence Structure in Arabic}.

There are two primary categories of sentences in Arabic: the nominal and verbal sentences. or the nominal phrase, starts with a noun, such as "the book is new". There are two primary categories of sentences in Arabic: the nominal phrase and the verbal phrase. The nominal sentence starts with a noun, such as "the book is new". The subject is explicitly stated, and the predicate provides information about the subject [27]. The predicate can be an adjective, noun, or prepositional phrase. The verbal phrase starts with a verb, such as "The man went out". In Arabic grammar, the primary distinction between verbal and nominal sentences is that the subject of a verbal sentence is the action the verb is doing, whereas the subject of a nominal phrase is the statement's major topic. Compared to English, Arabic has a more flexible word order. Nevertheless, the most widely used word orders are Verb-Subject-Object (VSO) and Subject-Verb-Object (SVO). The VSO order is the most traditional and common word order in Arabic. The verb is placed at the beginning of the sentence, followed by the subject and object. This structure often emphasizes the action that is being performed. The SVO order is also used in Arabic, particularly in less formal contexts or when the subject needs to be emphasized. Although SVO may be found in both spoken and written Arabic, the VSO order is more formal and is frequently used in writing Arabic [28]. An example of these orders for the sentence the boy wrote the book is shown in Fig. 1. Part (a) represents the VSO order. SVO is represented by part (b).

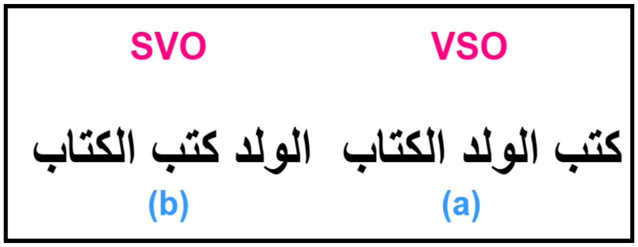


Fig. 1. A representation of SVO and VSO for Arabic sentences: a) description of VSO order in Arabic, b) description of SVO order in Arabic

### B. Complex Arabic Sentence Structure

There are intricate sentence patterns, such as conditional phrases, relative clauses, and many kinds of subordinate clauses. In Arabic, conditional sentences are usually built with particles such as for actual situations and (if) for hypothetical conditions. It is essential to comprehend the subtle differences between these particles. Conjuncts such as (and), (or), and (but) are used to join clauses in complicated Arabic sentences. Subordinate clauses, such as who or which, frequently come after a relative pronoun. Arabic introduces subordinate clauses using a variety of particles, each of which has a distinct purpose [29].

### C. *Advanced Arabic Sentences*

Because Arabic has a two-word order: SVO and SOV, advanced, complicated Arabic sentences can be converted to simple, summarized forms [30]. There are exceptional cases of Arabic sentences where nominal and verbal sentences overlap. These cases have been treated by converting two words to one word or replacing the word order without affecting the meaning, as shown in Table II.

TABLE II. THE ORIGINAL AND CONVERTED ARABIC SENTENCES

| Arabic sentence | English sentence | Converted Arabic sentence |
|---|---|---|
| الطالب الذي يدرس هنا ذكي | The student who studies here is smart | الطالب الدارس هنا |
| أريد أن أذهب | I want to go | أريد الذهاب |
| الكتاب موضوعه جميل | The topic of the book is good | موضوع الكتاب جميل |

### D. *Arabic Grammar Errors*

This section describes various errors in the Arabic language. Arabic errors range from number and gender agreement to the wrong case and mood in many cases. There are seven main types of Arabic errors [31]. First, AGE indicates Adjectives describing a noun or pronoun agree in case (i.e., both are nominative, accusative, or genitive), number (both are single, dual, or plural), definiteness (i.e., both are definite or indefinite), and gender (i.e., both are masculine or feminine). Second, the singular predicate describing the subject agrees in case, number, and gender but varies in definiteness (the predicate must be indefinite). Third, verbs in Arabic must match the subject in gender (i.e., masculine or feminine) in verbal sentences (where the verb comes before the subject). Fourth, the verb at the beginning of the sentence must be singular. Fifth, the object of verbal sentences, the predicate of en and her sisters, the subject of kan and her sisters, and verbs after present tense verb accusative tools will be in the accusative case. Sixth, the subject of the verbal sentence, the subject of kan and her sisters, and the predicate of en and her sisters must be in the nominative case. Seventh, the noun after a preposition and in the idafa construction must be in the genitive case. Eighth, verbs after the present-tense verb jussive tools will be in the jussive case [32]. Table III shows examples of each AGE sorted by error number.

TABLE III. EXAMPLES OF COMMON ARABIC ERRORS\LABEL

| Error number | Incorrect Arabic sentence | Correct Arabic sentence |
|---|---|---|
| First error | التلميذ كريمان | التلميذ كريم |
| Second error | الشجرة ثابت | الشجرة ثابتة |
| Third error | الحديقة الجميلة | الحديقة جميلة |
| Fourth error | قالت الرجل | قال الرجل |
| Fifth error | قالوا الرجلان | قال الرجلان |
| Sixth error | ان التلميذان مجتهدين | ان التلميذين مجتهدان |

## IV. DATA COLLECTION

This section describes how the database is created and shows the four substages used for the collection of the DAGC-ATS database, as shown in Fig. 2. These stages are preprocessing, feature extraction, Arabic Grammar Error Generation (AGEG), and generating DAGC-ATS database entries and statistics. First, the examples for different structures of Basic Arabic Sentences (BAS) are passed to the preprocessing stage for the determination of Arabic sentences, clauses, and words. Second, the feature extraction stage is applied by three substages: the generation of features describing Arabic words, and the selection of dedicated features that can describe BAS. Previous stages are illustrated in the work [33]. Third, AGER, using the Fundamental Principle of Counting (FPC), is applied to generate errors for each sentence structure. The final entries for the DAGC-ATS database are created into two files, and their statistics are reported. Data collection is implemented using the Stanford parser and the NLTK libraries. Stanza is a Python-Based Toolkit for Natural Language Processing in a Variety of Human Languages [34]. Natural Language Toolkit (NLTK) is a natural language processing tool for Python [35]. All stages used for data collection are shown in Fig. 2.

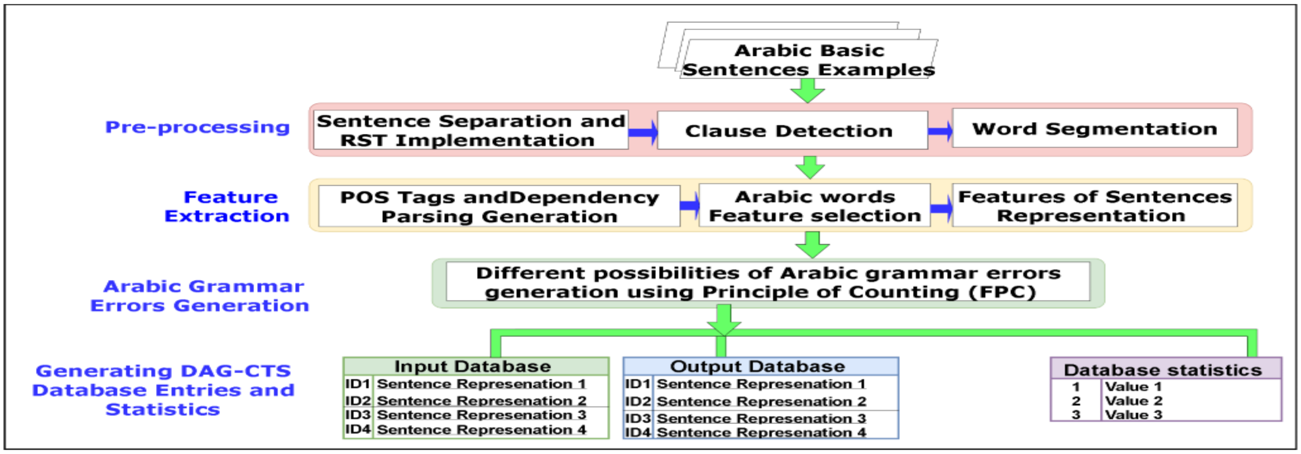

Arabic Basic Sentences Examples
Pre-processing
Sentence Separation and RST Implementation
Clause Detection
Word Segmentation
Feature Extraction
POS Tags andDependency Parsing Generation
Arabic words Feature selection
Features of Sentences Representation
Arabic Grammar Errors Generation
Different possibilities of Arabic grammar errors generation using Principle of Counting (FPC)
Generating DAG-CTS Database Entries and Statistics
Input Database
ID1 Sentence Represenation 1
ID2 Sentence Represenation 2
ID3 Sentence Represenation 3
ID4 Sentence Represenation 4
Output Database
ID1 Sentence Represenation 1
ID2 Sentence Represenation 2
ID3 Sentence Represenation 3
ID4 Sentence Represenation 4
Database statistics
1 Value 1
2 Value 2
3 Value 3

Fig. 2. The stages of DAGC-ATS database collection.

### A. *Preprocessing*

The preprocessing consists of two substages. First, a substage is sentence segmentation and Rhetorical Structure Theory (RST) implementation to ignore unnecessary sentences [36]. The second and third substages are Arabic clause detection and word segmentation. Clause detection to segment multiple Arabic sentences into basic sentences. Word segmentation to convert clauses to words. The output of this stage is a set of sentences, clauses, and words of the text input.

### B. Features Extraction

The second stage includes the extraction of features that describe the structure of BAS. This stage contains three sub-stages: Part-Of-Speech (POS) tagging and dependency parsing creation, Arabic words feature selection, and final features of sentences.

### C. POS Tagging and Dependency Parsing Creation

This stage includes extracting the features that represent the words. Every word in the Arabic text has been replaced with its characteristics, such as POS tags, case, number, gender, definiteness, and dependency relation between words. Each word in a document has been given a grammatical category, such as nouns, verbs, adjectives, and adverbs, which is known as POS tagging [37], [38] (Appendix A). This substage is based on Universal Dependencies for Arabic and Prague Arabic Dependency Treebank (PADT) [39], [40]. The Arabic features used in the DAGC-ATS database, and their counterparts in the PADT, are shown in Appendix B.

### D. Arabic Word Feature Selection

This stage relates to the conversion of complex, advanced Arabic sentences to simple sentences based on the structure of the BAS. In the feature selection of the Arabic word substage, the words that represent the BAS are selected. The final substage is considered with the representation of BAS as a set of features. The basic idea for the creation of the DAGC-ATS database is to represent the structure of BAS using POS tags, the features of words, and the dependency relations between words. The advanced nominal and verbal sentences consist of the BAS and other tools. BAS is the cornerstone of complex sentence formation. Nominal and verbal sentences have BAS structures. The nominal sentence consists of two main components, which are the subject and the predicate of the sentence. Some tools precede the nominal sentence and change its meaning and grammatical signs, including kan or en and their sisters, or negation and interrogative tools. There are sentence complements in some nominal sentences, such as semi-sentence and idafa construction. These different structures for nominal sentences are shown in Fig. 3. The overlapping rectangles illustrate the structural arrangement of the nominal sentence. The word categories present from the beginning of the sentence to the end of the sentence in order are shown in Fig. 3 from part (a) to part (d). The largest rectangle represents the beginning of the nominal sentence. It contains the additional tools that precede BAS, and they are en, kan, and their sisters, interrogative and negative tools, as shown in Part (d) of Fig. 3. Parts (c) and (d) are the components of the BAS. The smallest rectangle represents the end of the sentence, which represents the nominal complement of the sentence, and it comes after BAS for the nominal sentence, as shown in part(a). The rectangles in the middle represent the rest of the sentence.

Examples of nominal BAS, nominal sentence complement, and other tools are shown in Fig. 4. An example of an extremely basic nominal Arabic sentence is shown in Fig. 4 n parts c and d. An example of a nominal sentence with the subject is a singular noun and the predicate is a mandatory sentence complement, as shown in Fig. 4 in part a. A nominal sentence beginning with an "en tool" is shown in part b of Fig. 4.

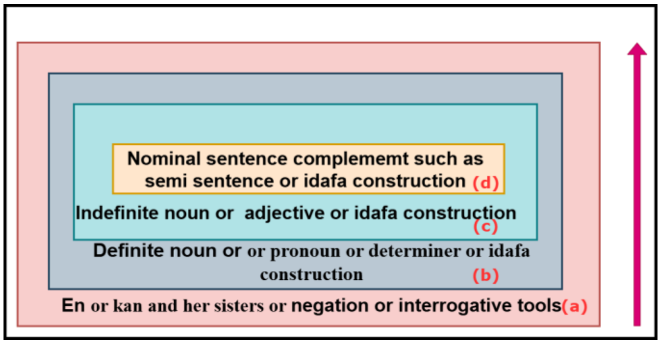

Nominal sentence complememt such as
semi sentence or idafa construction (d)
Indefinite noun or adjective or idafa construction (c)
Definite noun or or pronoun or determiner or idafa construction (b)
En or kan and her sisters or negation or interrogative tools(a)

Fig. 3. The different structures for nominal Arabic simple sentences: a) the beginning of the nominal sentence, b) description of different structures of the subject in a nominal sentence, c) description of different structures of the predicate in a nominal sentence, d) the complement of the nominal sentence.

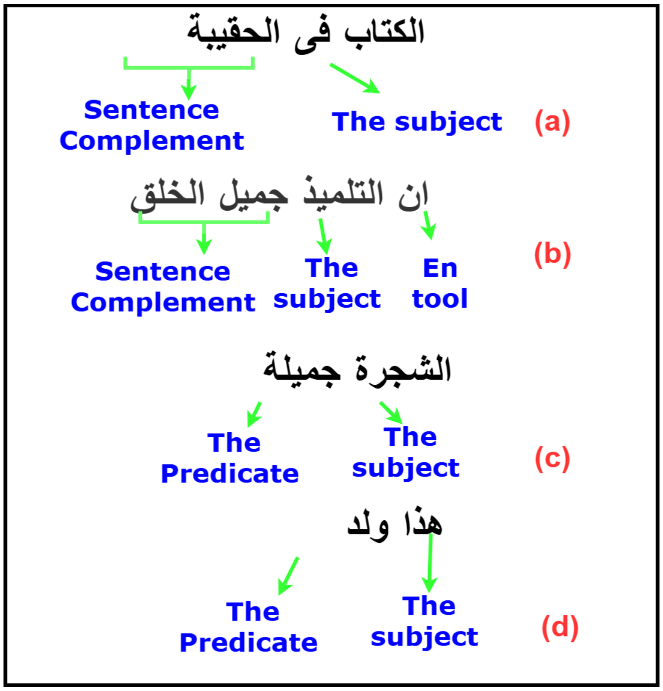


Fig. 4. Examples of Arabic nominal sentences: a) The subject is a singular noun and the predicate is a semi-sentence. b) Description of an en tool and a nominal sentence with idafa construction as a complement. c) The subject and the predicate are singular. d) The subject is a pronoun, and the predicate is a singular noun.

In verbal sentences, there are two main parts of the verbal sentence: the verb and the subject. Both direct and indirect. Verbs can be perfective or imperative. The subject can be a singular noun, an adjective, a pronoun, or a determiner. Certain tools, such as jussive and accusative tools for verbs in the present tense and interrogative tools, come before the verbal phrase and alter its meaning and grammatical indicators. Certain verbal sentences, including semi-sentences and idafa constructions, include sentence complements. The sentence is called a semi-sentence if it begins with a preposition or adverb.

The structural arrangement of the verbal sentence is represented by the overlapping rectangles. Fig. 5 shows the word categories that are present in the sentence from part (a) to part (d), in order, from the beginning to the end. The start of the verbal sentence is shown by the largest rectangle. It includes the negation or interrogative tools, or accusative or jussive tools, which come before BAS. The verbal complement of the sentences is represented by the smallest rectangle, which is the end of the sentence, and they come after BAS for the verbal sentence as shown in part (c) of Fig. 5. The BAS of the verbal sentence is represented by the rectangles in the center as shown in parts (a, b, c) of Fig. 5.

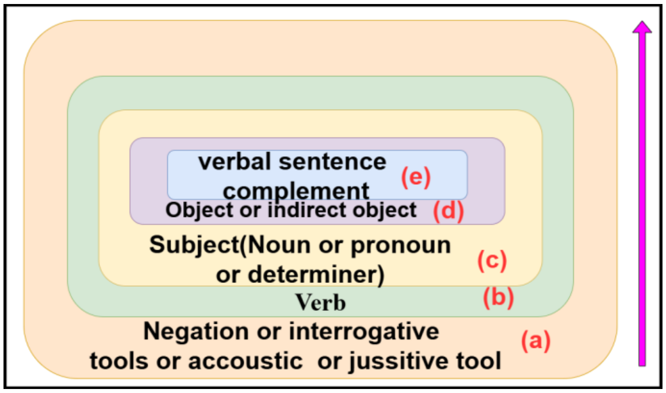

verbal sentence complement
(e)
Object or indirect object
(d)
Subject(Noun or pronoun or determiner)
(c)
Verb
(b)
Negation or interrogative tools or accoustic or jussive tool
(a)

Fig. 5. The structures of Arabic verbal sentences: a) Description of the beginning of the verbal sentence.b) The basic component of the verbal sentence is a verb. c) Description of different structures of the subject in a verbal sentence. d) The object and the indirect object are components of a verbal sentence. e) Description of the complement of the verbal sentence.

## E. Features of Sentences Representation

Every word in the BAS is represented by its POS tag, the Case, the Gender, the number, and the syntactic dependency relations. The noun and adjective are further described by definiteness. Table IV shows the word types used in the DAG-CTS database and their representations.

TABLE IV. ARABIC WORDS AND THEIR REPRESENTATIONS IN THE DAGC-ATS DATABASE

| Word type | Word features | Word features values |
|---|---|---|
| Adjective | POS tag Case Definiteness Number Dependency Relation Gender | adj Acc Ind Sing xcomp NONE |
| Noun | POS tag Case Definiteness Number Dependency Relation Number | NOUN Nom Cons Sing nsubj:pass |
| Determiner | DET Case Gender Number Pronoun Type Dependency Relation | Acc Masc Sing Dem det |
| Verb | POS tag Perf Gender Number NONE Dependency Relation | VERB Perf Masc Sing NONE root |
| Pronoun | PRON Pronoun type Nom Masc Sing nmod | Noun object & POS tag Case obj & NOUN Acc obj |
| Proper noun and unknown noun | POS tag Pronoun type Case Gender Number Dependency Relation | X root NONE |
| Preposition | ADP Prep Dependency Relation | ADP Prep case |

Examples of basic Arabic verbal sentences are shown in Fig. 6. The basic verbal sentence of the verb and subject is shown in part (a). An example of a verbal sentence with a verb, subject, and object is shown in part (b). In part (c), a verbal sentence is shown with a complement to the sentence. The jussive tool with a verbal sentence is shown in part (d).

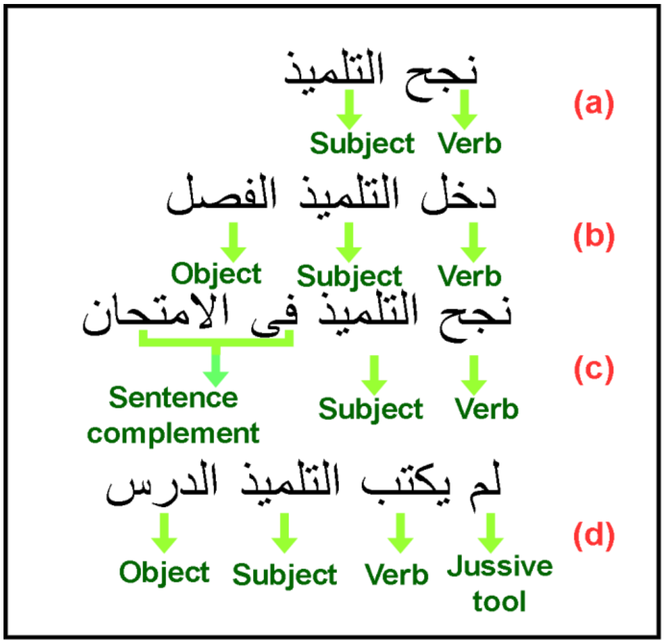


Fig. 6. Examples of Arabic verbal sentences: a) A verbal sentence with a verb and a subject, b) A verbal sentence with a verb, a subject, and an object, c) A verbal sentence with a verb and a subject, and a semi-sentence as a complement of the sentence, d) A jussive tool with a verbal sentence.

An example of the generation of entries for the DAGC-ATS database for input Arabic text is shown in Fig. 7. The input of the Arabic sentences is shown in part(a) of Fig. 7. Part (b) of Fig. 7 shows the sentence segmentation stage, which separates the input text into isolated sentences. Then each sentence is separated into clauses as shown in part(c) of Fig. 7. Next, each clause is segmented into words as shown in part(d) of Fig. 7. The features of each word are represented as shown in part (e) of Fig. 7. The features of the words that make up the summary of a sentence are collected as shown in part(f) of Fig. 7.

## F. Arabic Grammar Errors Generation

In the third stage, all AGE probabilities are generated using FPC. The fundamental idea of FPC is a far simpler method of determining the total number of combinations by applying the basic counting principle to make informed choices from all available choices. Finding each combination in a particular scenario is made easier by the Fundamental Principle of Counting. Suppose that the Arabic word is described by three features that cause grammatical errors, as shown in Fig. 8. These features are gender, numbers, and grammar. There is a need to create all probabilities of Arabic grammatical errors for that word by applying FPC. This is done by multiplication of the number of features of each variable, which is equal to 3*3*2=18.

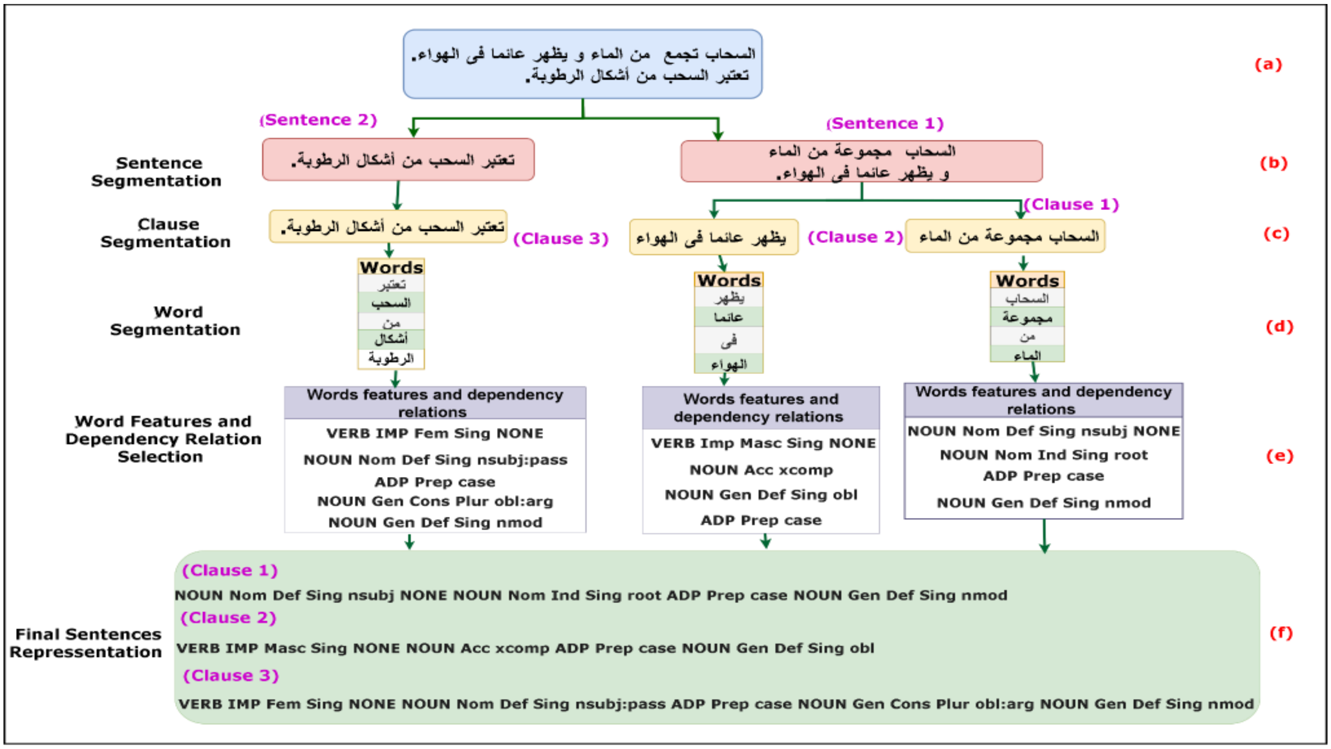

السحاب تجمع من الماء و يظهر عائما فى الهواء.
تعتبر السحب من أشكال الرطوبة.
(a)
(Sentence 2)
(Sentence 1)
Sentence Segmentation
تعتبر السحب من أشكال الرطوبة.
السحاب مجموعة من الماء
و يظهر عائما فى الهواء.
(b)
Clause Segmentation
تعتبر السحب من أشكال الرطوبة.
(Clause 3)
يظهر عائما فى الهواء
(Clause 2)
السحاب مجموعة من الماء
(Clause 1)
(c)
Word Segmentation
Words
تعتبر
السحب
من
أشكال
الرطوبة
Words
يظهر
عائما
فى
الهواء
Words
السحاب
مجموعة
من
الماء
(d)
Word Features and Dependency Relation Selection
Words features and dependency relations
VERB IMP Fem Sing NONE
NOUN Nom Def Sing nsubj:pass
ADP Prep case
NOUN Gen Cons Plur obl:arg
NOUN Gen Def Sing nmod
Words features and dependency relations
VERB Imp Masc Sing NONE
NOUN Acc xcomp
NOUN Gen Def Sing obl
ADP Prep case
Words features and dependency relations
NOUN Nom Def Sing nsubj NONE
NOUN Nom Ind Sing root
ADP Prep case
NOUN Gen Def Sing nmod
(e)
Final Sentences Representation
(Clause 1)
NOUN Nom Def Sing nsubj NONE NOUN Nom Ind Sing root ADP Prep case NOUN Gen Def Sing nmod
(Clause 2)
VERB IMP Masc Sing NONE NOUN Acc xcomp ADP Prep case NOUN Gen Def Sing obl
(Clause 3)
VERB IMP Fem Sing NONE NOUN Nom Def Sing nsubj:pass ADP Prep case NOUN Gen Cons Plur obl:arg NOUN Gen Def Sing nmod
(f)

Fig. 7. An explanation of Arabic clause representation in the DAGC-CTS database: a) The original text. b) Sentence segmentation of the text. c) Clause segmentation of the text. d) Text segmentation into words. e) Features representation of words in the text. f) Final sentences representations.

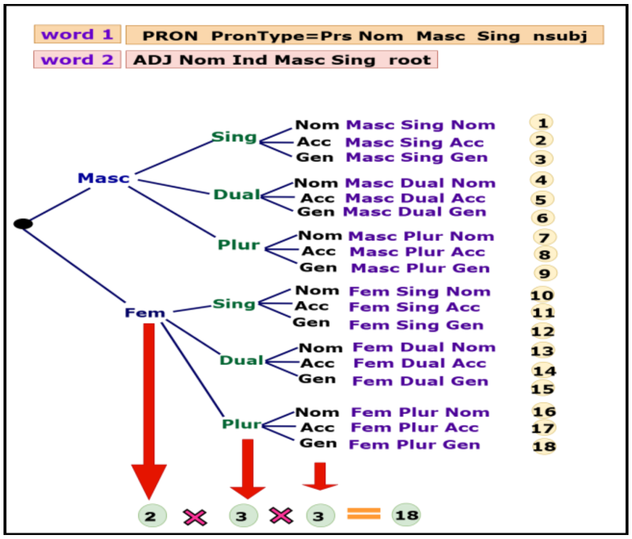


Fig. 8. An explanation of the generation of all combinations of AGE using FPC.

## G. Generating the DAGC_ATS Database Entries and Statistics

Finally, generated Arabic sentence representations are collected into two separate files, one for inputting incorrect Arabic sentences. The other file is for the correct output of Arabic sentences. The statistics of the DAG-CTS database are reported. In this section, the data collection process is discussed, and different statistics are reported. The database covers 24 grammar cases for simple sentences with all probabilities of Arabic errors. The database consists of the structure of basic Arabic sentences that have meaning. Every entry in the DAGC-ATS database has a different sentence structure. The DAGC-ATS database is grouped into four groups: nominal sentence, verbal sentence, “kan” and her sisters' sentence, and “en” and his sister's sentences in Arabic. Each entry in every group is different from the other entries in that group in gender, number, and case. Every entry in the database is different from the other entries and covers one or more errors and their corrections.

## V. Database Statistics

The DAGC-ATS database contains 9309888 entries. These entries represent all cases for Arabic grammar errors, including differences in number, gender, definiteness, and grammatical cases. The DAGC-ATS database consists of two files. One file for input data entries contains different combinations of incorrect Arabic clauses in addition to sentence complements. The other file involves the corrected version of the clauses for the input file. It combines 687 Arabic sentence structures. The statistics for the database are shown in Table V and Fig. 9.

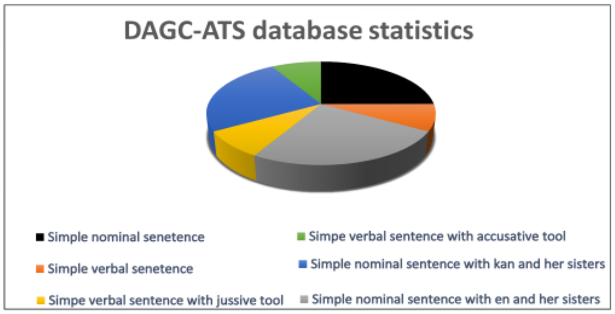


Fig. 9. The statistics of the DAGC-ATS databases.

The challenge facing the DAGC-ATS database is the inability of the Stanza toolkit to identify “kan” and “en” and their sisters. Also, the lack of distinction between the jussive and accusative tools for the present tense and normal verbs is important. This case is solved by adding a list of these tools to recognize them. In addition to the lack of identification of the broken plural.

TABLE V. THE DAGC-ATS DATABASE STATISTICS

| Type of sentence | Number of sentences |
|---|---|
| Nominal sentence | 2363604 |
| Verbal sentence | 739692 |
| En tool with nominal sentence | 2363604 |
| Kan tool with nominal sentence | 2363604 |
| Jussive tool with verbal sentence | 739692 |
| Acoustic tool with verbal sentences | 739692 |
| Total | 9468324 |

The challenge facing the DAGC-ATS database is the inability of the Stanza toolkit to identify kan, en, and her sisters. Also, the lack of distinction between the jussive and accusative tools for the present tense and normal verbs is important. This case is solved by adding a list of these tools to recognize them. In addition to the lack of identification of the broken plural.

## A. Results

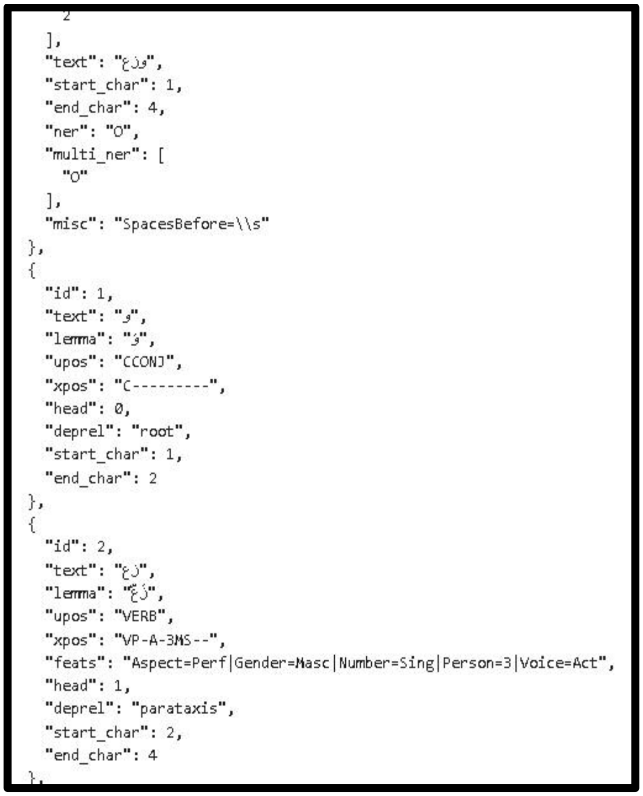
The goal of the initial QALB shared task, QALB-2014, was to fix mistakes made by native Arabic speakers in online comments. QALB-2015, the second shared task, expanded the task to include a track to fix mistakes in texts written by Arabic language learners. If the system produces a sentence that is syntactically correct, the output is deemed correct. The evaluation metrics used in all GECs are used to evaluate the suggested database. Eq. (1)- Eq. (3) are used to calculate the precision, recall, and F-measure for the suggested GEC system. A confusion matrix is a simple table that contrasts the classification model's predictions with the actual results [34]. That division is made into four categories: accurate predictions for both classes (True Positives (TP) and True Negatives (TN)) and incorrect predictions (False Positives (FP) and False Negatives (FN)). The matrix displays the number of instances in the model produced on the test data. The ratio of True Positives to all Positives is known as precision. The metric of our model that accurately detects True Positives is called recall [41].

$$precision\ (p) \frac{TP}{TP+FP} \tag{1}$$

$$Recall(R) = \frac{TP}{TP+FN} \tag{2}$$

$$F\ 1Score(F\ 1) = 2\ \frac{PR}{P+R} \tag{3}$$

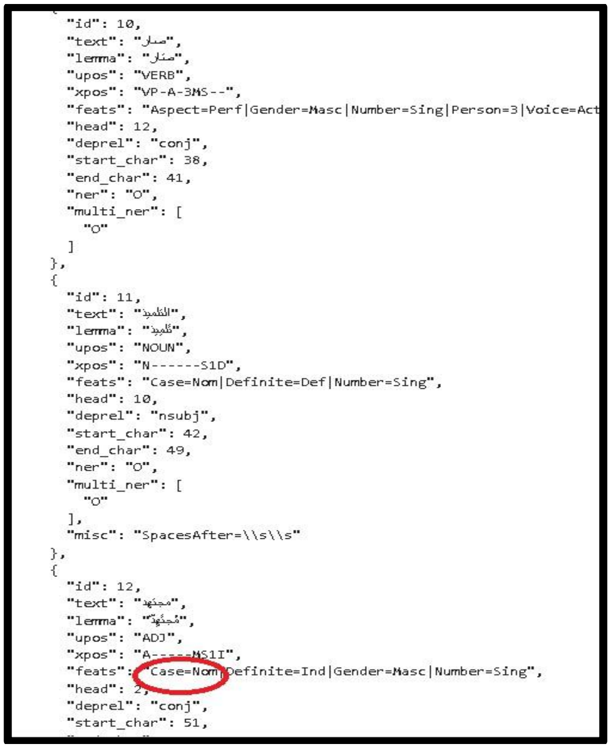
Part of the DAGC-ATS is used for testing AGEC in the work [42]. The system is tested using the QALB dataset and achieved 96.9, 94.80, and 95.83 for precision, recall, and F-measure. In addition, the same system is tested using the EASC database using 785 summaries, and it obtained 99.73%, 95.90%, 97.77% for precision, recall, and F-measure. The challenge facing the AGEC using this database is the inability of stanza to differentiate between the sisters of "kān" and the sisters of "en" from other verbs. This problem was solved by adding all the sisters of "kān" and "en" to a list and adding them as "En" and "kan" in the database. This problem also failed in the stanza to give the correct diacritical marks to the subject and predicate of "kān" and "en." The stanza was also unable to differentiate between past tense and imperative tense verbs, and it was treated as past tense verbs. In some cases, the letter "waw" in the verb was counted as a letter, and the verb was incorrectly divided. The system was unable to give the jussive and accusative grammatical case to jussive and accusative verbs in the present tense, due to its inability to recognize all the jussive and accusative tools. Fig. 10 shows the separation of the verb into segments because of the “waw” letters.



```
      2
    ],
    "text": "وزع",
    "start_char": 1,
    "end_char": 4,
    "ner": "O",
    "multi_ner": [
      "O"
    ],
    "misc": "SpacesBefore=\\s"
  },
  {
    "id": 1,
    "text": "و",
    "lemma": "وَ",
    "upos": "CCONJ",
    "xpos": "C---------",
    "head": 0,
    "deprel": "root",
    "start_char": 1,
    "end_char": 2
  },
  {
    "id": 2,
    "text": "زع",
    "lemma": "زَع",
    "upos": "VERB",
    "xpos": "VP-A-3MS--",
    "feats": "Aspect=Perf|Gender=Masc|Number=Sing|Person=3|Voice=Act",
    "head": 1,
    "deprel": "parataxis",
    "start_char": 2,
    "end_char": 4
  },
```

Fig. 10. The wrong verb separation.

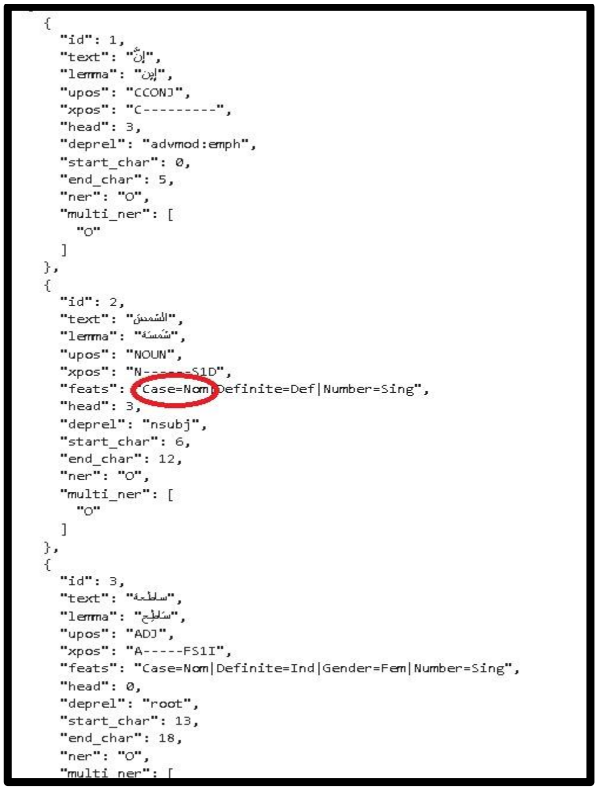
Fig. 11 shows the wrong grammatical case for the predicate of the sister of “kan”. The grammatical case is supposed to be accusative, not nominal, in the predicate of the sister of “kan.”.



```
    "id": 10,
    "text": "صار",
    "lemma": "صَار",
    "upos": "VERB",
    "xpos": "VP-A-3MS--",
    "feats": "Aspect=Perf|Gender=Masc|Number=Sing|Person=3|Voice=Act
    "head": 12,
    "deprel": "conj",
    "start_char": 38,
    "end_char": 41,
    "ner": "O",
    "multi_ner": [
      "O"
    ]
  },
  {
    "id": 11,
    "text": "التلميذ",
    "lemma": "تِلمِيذ",
    "upos": "NOUN",
    "xpos": "N------S1D",
    "feats": "Case=Nom|Definite=Def|Number=Sing",
    "head": 10,
    "deprel": "nsubj",
    "start_char": 42,
    "end_char": 49,
    "ner": "O",
    "multi_ner": [
      "O"
    ],
    "misc": "SpacesAfter=\\s\\s"
  },
  {
    "id": 12,
    "text": "مجتهد",
    "lemma": "مُجتَهِد",
    "upos": "ADJ",
    "xpos": "A-----MS1I",
    "feats": "Case=Nom|Definite=Ind|Gender=Masc|Number=Sing",
    "head": 2,
    "deprel": "conj",
    "start_char": 51,
```

Fig. 11. The inaccurate grammatical case example.

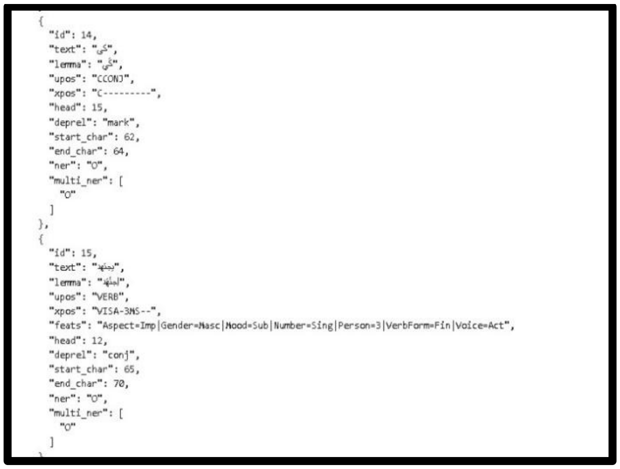
Fig. 12 displays an example of an inaccurate grammatical case for the subject of the sister "en". The grammatical case for the subject of the sister "en" is supposed to be accusative.



```
  {
    "id": 1,
    "text": "إنّ",
    "lemma": "إِن",
    "upos": "CCONJ",
    "xpos": "C---------",
    "head": 3,
    "deprel": "advmod:emph",
    "start_char": 0,
    "end_char": 5,
    "ner": "O",
    "multi_ner": [
      "O"
    ]
  },
  {
    "id": 2,
    "text": "الشمس",
    "lemma": "شَمسَة",
    "upos": "NOUN",
    "xpos": "N------S1D",
    "feats": "Case=Nom|Definite=Def|Number=Sing",
    "head": 3,
    "deprel": "nsubj",
    "start_char": 6,
    "end_char": 12,
    "ner": "O",
    "multi_ner": [
      "O"
    ]
  },
  {
    "id": 3,
    "text": "ساطعة",
    "lemma": "سَاطِع",
    "upos": "ADJ",
    "xpos": "A-----FS1I",
    "feats": "Case=Nom|Definite=Ind|Gender=Fem|Number=Sing",
    "head": 0,
    "deprel": "root",
    "start_char": 13,
    "end_char": 18,
    "ner": "O",
    "multi_ner": [
```

Fig. 12. Example of wrong grammatical case of the subject of the sister of" en".

The grammatical case is not recognized in the existence of jussive and accusative tools in addition to the verbs in the present tense, as shown in Fig. 13. In the "feats" of the verb, there is no case.



```
  {
    "id": 14,
    "text": "كي",
    "lemma": "كَي",
    "upos": "CCONJ",
    "xpos": "C---------",
    "head": 15,
    "deprel": "mark",
    "start_char": 62,
    "end_char": 64,
    "ner": "O",
    "multi_ner": [
      "O"
    ]
  },
  {
    "id": 15,
    "text": "يجتهد",
    "lemma": "اِجتَهَد",
    "upos": "VERB",
    "xpos": "VISA-3MS--",
    "feats": "Aspect=Imp|Gender=Masc|Mood=Sub|Number=Sing|Person=3|VerbForm=Fin|Voice=Act",
    "head": 12,
    "deprel": "conj",
    "start_char": 65,
    "end_char": 70,
    "ner": "O",
    "multi_ner": [
      "O"
    ]
```

Fig. 13. The non-recognition of the grammatical case of the verb in the present tense after accusative tools.

A comparison with recent work in correcting Arabic errors is shown in Table VI. All the works mentioned in the table cited the problem of a lack of databases for all fields, or a large amount of training data covering all fields and dialects. A work mentioned methods to solve this problem using data augmentation [47]. Other work uses a confusion method to generate training data [45]. Another work uses a noising method to get training data. All of the works in Table VI suffer from domain difference. The proposed database creates another unconventional approach that focuses on representing correct and incorrect Arabic sentences with a set of characteristics that represent their words, without regard to the spelling of Arabic words, in order to identify errors in Arabic grammar. This database is used as a training database.

TABLE VI. RECENT WORKS COMPARISON FOR CORRECTION OF ARABIC ERRORS

| Authors | Database / Dataset | Precision | Recall | F1 Score | Year |
|---|---|---|---|---|---|
| Alrehili & Alhothali [12] | Tibyan Corpus | 97 | 97 | 97 | 2025 |
| AlOyaynaa & Kotb [14] | Public Arabic Sets | 0.97 | 0.99 | 0.98 | 2023 |
| Abdelaal et al. [43] | QALB 2014/2015 | 0.51 | 0.4 | 0.45 | 2024 |
| S. Mahmoud et al. [44] | ZAEBUC | 85.1 | 75.4 | 74.7 | 2024 |
| Solyman et al. [45] | QALB 2014/2015 | 79.06 | 70.43 | 74.18 | 2022 |
| Z. Mahmoud et al. [46] | QALB 2014/2015 | 64.37 | 45.51 | 53.32 | 2023 |
| Solyman et al. [47] | QALB 2014/2015 | 75.99 | 58.29 | 65.98 | 2023 |
| The proposed database | QALB | 96.9, | 94.80 | 95.83 | 2026 |

## VI. CONCLUSION

The Arabic language is complex with a lot of morphology. In addition to the lack of available and open domain databases for AGEC. So, AGEC is a big challenge. This study presents a new open-domain database called DAGC-ATS for AGCEC. It is based on the description of the BAS structure using word type and features, in addition to the dependency relations between words. It contains 9309888 entries covering nominal and verbal sentences, en or kan and their sisters with nominal and verbal sentences, and jussive and accusative tools with verbal sentences. It contains 687 Arabic sentence structures. It has 9468324 different entries.

## VII. Future Work

We will expand the database to include all types of simple sentences, and we may also address complex sentences. I also believe that by studying syntactic dependency relations in UD_Arabic-PADT, we can solve other Arabic language error problems. And by studying the features of words in UD_Arabic-PADT, we can differentiate between special cases of Arabic words. Always keep an eye out for updates on the database website. We will increase the number of Arabic sentence structures in the database to include most of the structures in the Arabic language, and we will test the database on a number of other databases for AGEC.

## Data Availability

The DAGC-ATS dataset is available: \url{https://github.com/nadaessa369/dagc_ats-database}. Samples of verbal and nominal sentences are shown in verbal sentence samples.xlsx and nominal sentence samples.xlsx. In addition to the probabilities of error in gender, number, type of verbs, and definiteness, the total number of probabilities in each sample is also presented. An explanation of the more important things to understand the database is also presented.

## References

[1] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: An overview," Journal of King Saud University-Computer and Information Sciences, vol. 33(5), pp. 497-507, 2021.

[2] A. Farghaly, and K. Shaalan, "Arabic natural language processing: Challenges and solutions," ACM Transactions on Asian Language Information Processing, TALIP, vol. 8(4), pp.1-22, 2009.

[3] A. P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, and D. R. I. M Setiadi, "Review of automatic text summarization techniques & methods," Journal of King Saud University-Computer and Information Sciences, vol. 34(4), pp.1029-1046, 2022.

[4] M. Mamnunah, M. Abdurrahman, and A. Sopian, "The Error Analysis of Arabic Grammar in The Kalamuna Book," Arabi: Journal of Arabic Studies, vol. 6(2), pp.158-166, 2021

[5] M. E. Saleh, Y. M. Wazery, and A. A. Ali, "A systematic literature review of deep learning-based text summarization: Techniques, input representation, training strategies, mechanisms, datasets, evaluation, and challenges,"Expert Systems with Applications, vol. 252, pp.124153, 2024.

[6] G. Alselwi, and T. Taşcı, "Extractive Arabic text summarization using PageRank and word embedding," Arabian Journal for Science and Engineering, vol. 49(9), pp.13115-13130, 2024.

[7] M. A. Salam, M. Aldawsari, M. Gamal, H. F. Hamed, and S. Sweidan, "MSG-ATS: Multi-level semantic graph for arabic text summarization,"IEEE Access, vol. 12, pp.118773-118784, 2024.

[8] N. Madi, and H. Al-Khalifa, " Error detection for Arabic text using neural sequence labeling," Applied Sciences, vol. 10(15), pp.5279, 2020.

[9] A. A. ElSabagh, S. S. Azab, and H. A. Hefny, "A comprehensive survey on Arabic text augmentation: approaches, challenges, and applications,"Neural Computing and Applications, pp.1-34, 2025.

[10] K. Ismail, S. Abdou, M. Farouk, and A. Salem, "Transformers to the rescue: alleviating data scarcity in arabic grammatical error correction with pre-trained models," Neural Computing and Applications, vol. 37(18), pp.13011-13038, 2023

[11] A. Ahmed, N. Ali, M. Alzubaidi, W. Zaghouani, A. A. Abd-alrazaq, and M. Househ, "Freely available Arabic corpora: A scoping review," Computer Methods and Programs in Biomedicine Update, vol. 2, pp. 100049, 2022.

[12] A. Alrehili, and A. Alhothali, "Tibyan corpus: balanced and comprehensive error coverage corpus using ChatGPT for Arabic grammatical error correction," PeerJ Computer Science, vol. 11, pp. e2724, 2025.

[13] C. Moukrim, T. Abderrahim, and A. Tarik, "An innovative approach to autocorrecting grammatical errors in Arabic texts," Journal of King Saud University-Computer and Information Sciences, vol. 33(4), pp.476-488, 2021.

[14] B. Mohit, "QALB: Qatar Arabic language bank," In Qatar Foundation Annual Research Forum, vol. 2013, No. 1, pp. ICTP-032, Hamad bin Khalifa University Press (HBKU Press), 2013.

[15] Z. Althafir, and R. Ghnemat, "A hybrid approach for auto-correcting grammatical errors generated by non-native Arabic speakers," In 2022 International Conference on Emerging Trends in Computing and Engineering Applications (ETCEA), pp. 1-6, IEEE, 2022

[16] S. AlOyaynaa, and Y. Kotb, "Arabic grammatical error detection using transformers-based pretrained language models," In ITM Web of Conferences , vol. 56, p.p. 04009, EDP Sciences, 2023.

[17] A. Solyman, W. Zhenyu, T. Qian, A. A. M. Elhag, M. Toseef, and Z. Aleibeid, "Synthetic data with neural machine translation for automatic correction in arabic grammar," Egyptian Informatics Journal, vol. 22(3), pp.303-315, 2021.

[18] Nerabie, Abdul Munem, et al. "The impact of Arabic part of speech tagging on sentiment analysis: A new corpus and deep learning approach." Procedia Computer Science 184 (2021): 148-155.

[19] I. Zeroual, D. Goldhahn, T. Eckart, and A. Lakhouaja, "OSIAN: Open source international Arabic news corpus-preparation and integration into the CLARIN-infrastructure," In Proceedings of the fourth arabic natural language processing workshop, pp. 175-182, 2019.

[20] S. Mahmoud, E. Nabil, and M. Torki, " Automatic Scoring of Arabic Essays: A Parameter-Efficient Approach for Grammatical Assessment," IEEE Access, 2024.

[21] N. Habash, and D. Palfreyman, "ZAEBUC: An annotated Arabic-English bilingual writer corpus," In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 79-88, 2022,

[22] A. Rozovskaya, N. Habash, R. Eskander, N. Farra, and W. Salloum, "The columbia system in the qalb-2014 shared task on arabic error correction," In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pp. 160-164, 2014.

[23] M. Nawar, "CUFE@ QALB-2015 shared task: Arabic error correction system," In Proceedings of the second workshop on Arabic natural language processing , pp. 133-137, 2015,

[24] A. S. Kaye, "Arabic," In The world's major languages, pp. 577-594, Routledge, 2018.

[25] K. Versteegh, "Arabic language," Edinburgh University Press, 2014.

[26] Y. Suleiman, " Arabic grammar and linguistics," Routledge, 2013.

[27] Wright, and C. P. Caspari, "A grammar of the Arabic language," Cosimo, Inc., 2011.

[28] M. A. Alqarni, and M. S. Alanazi, "The Syntax of Nominal Appositions in Modern Standard Arabic," Theory and Practice in Language Studies, vol. 12(8), pp.1669-1689, 2022.

[29] A. Moubaiddin, A. Tuffaha, B. Hammo, and N. Obeid, "Investigating the syntactic structure of Arabic sentences." In 2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA), pp. 1-6, IEEE, 2013.

[30] S. R. Olaniyi, "A Survey of Al-Jumal Al-Ashartiyyah (The Conditional Sentences) in Arabic Language," African Journal of Humanities & Contemporary Education Research, vol. 2, pp.138-144, 2022.

[31] M. T. Alhawary, "Modern standard Arabic grammar: A learner's guide," John Wiley & Sons, 2011.

[32] M. S. Al-Rabiah, and A. Al-Salman, "An XML-based semantic parser for traditional Arabic," In 2010 4th International Universal Communication Symposium, pp. 312-319, IEEE, 2010.

[33] N. Essa, M. M. El-Gayar, and E. M. El-Daydamony, "Enhanced model for abstractive Arabic text summarization using natural language generation and named entity recognition," Neural Computing and Applications, vol. 37(10), pp.7279-7301, 2025.

[34] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," arXiv preprint arXiv:2003.07082, 2020.

[35] N. Hardeniya, J. Perkins, D. Chopra, N. Joshi, and I. Mathur, "Natural language processing: python and NLTK," Packt Publishing Ltd, 2016.

[36] A. Ibrahim, T. Elghazaly, and M. Gheith, "A novel Arabic text summarization model based on rhetorical structure theory and vector space model," International Journal of Computational Linguistics and Natural Language Processing, vol. 2(8), pp.480-485, 2013.

[37] A. M. Nerabie, M. AlKhatib, S. S. Mathew, M. El Barachi, and F. Oroumchian, " The impact of Arabic part of speech tagging on sentiment analysis: A new corpus and deep learning approach," Procedia Computer Science, vol. 184, pp.148-155, 2021.

[38] Y. Marton, N. Habash, and O. Rambow, " Dependency parsing of Modern Standard Arabic with lexical and inflectional features," Computational Linguistics, vol. 39(1), pp.161-194, 2013.

[39] J. Hajič, E. Hajičová, M. Mikulová, and J. Mírovský, " Prague dependency treebank," In Handbook of Linguistic Annotation, pp. 555-594, Dordrecht: Springer Netherlands, 2017.

[40] A. Zeldes, and M. Abrams, "The coptic universal dependency treebank," In Proceedings of the second workshop on universal dependencies (UDW 2018), pp. 192-201, 2018.

[41] S. R. Goyal, V. S. Kulkarni, R. Choudhary, and R. Jain, "A comparative analysis of efficacy of machine learning techniques for disease detection in some economically important crops," Crop Protection, vol. 190, p.p.107093, 2025.

[42] N. Essa, M. Elgayar, and E. El-Daydamony, "Arabic Grammar Correction for Arabic Text Summaries," Mansoura Journal for Computer and Information Sciences, vol. 20(2), pp.1-16, 2025.

[43] A. Abdelaal, A. A. Medhat, M. Elsayad, M. Foad, S. Khaled, A. Tamer, and W. Medhat, "Text Correction for Modern Standard Arabic," Procedia Computer Science, vol. 244, pp. 371–377, 2024.

[44] S. AlOyaynaa and Y. Kotb, "Arabic Grammatical Error Detection Using Transformers- based Pretrained Language Models," EDP Sciences, vol. 56 , pp. 1–14, 2023.

[45] A. Solyman, Z. Wang, Q. Tao, A. Abdulgader, M. Elhag, R. Zhang, and Z. Mahmoud, "Automatic Arabic Grammatical Error Correction based on Expectation- Maximization routing and target-bidirectional agreement," Knowledge-Based Systems , vol. 241, pp. 1–13, 2022.

[46] Z. Mahmoud, C. Li, M. Zappatore, A. Solyman, A. Alfatemi, A. O. Ibrahim, and A. Abdelmaboud. "Semi-supervised learning and bidirectional decoding for effective grammar correction in low resource scenarios," Computer Science , vol. 9.e1639 , pp. 1–25.2023

[47] A. Solyman, M. Zappatore, W. Zhenyu, Z. Mahmoud, A. Alfatemi, A. O. Ibrahim, and L. A. Gabralla. "Optimizing the impact of data augmentation for low-resource grammatical error correction," *Journal of King Saud University-Computer and Information Sciences, vol.*35, 6, pp. 1–15, 2023.

## ABBREVIATIONS

The following abbreviations are used in this manuscript:

TABLE VII. DAGC-ATS & DATABASE FOR ARABIC GRAMMAR CORRECTION

| NLP | Natural Language Processing |
|---|---|
| BPE | Byte Pair Encoding |
| MSA | Modern Standard Arabic |
| VSO | Verb-Subject-Object |
| SVO | Subject-Verb-Object\ |
| AGEG | Arabic Grammar Errors Generation |
| BAS | Basic Arabic Sentences |
| FPC | Fundamental Principle of Counting |
| RST | Rhetorical Structure Theory |
| POS | Part-Of-Speech |
| PADT | Prague Arabic Dependency Treebank |
| NLTK | Natural Language Toolkit |
| ARETA | Arabic Error Type Annotation tool |
| OSIAN | Open-Source International Arabic News |

## APPENDIX A

TABLE VIII. SOME ARABIC LANGUAGE TERMS AND WHAT THEY MEAN

| Arabic terms | Their meaning\ examples |
|---|---|
| Taa marbuta | ة |
| Preposition | من فى إلى على |
| Conjuncts | و فى |
| jussive tools | لم لما |
| Accusative tools | ان لن كى |
| En and her sisters | إن أن لكن لعل |
| kan and her sisters | كان صار ظل |

## APPENDIX B

TABLE IX. DIFFERENT ARABIC WORD TYPES AND THEIR POS TAGS AND DEPENDENCY RELATIONS IN DAG-CTS DATABASE

| Word type, features, and dependency relation | Universal UD in the DAG-CTS database |
|---|---|
| Verb | VERB |
| Noun | NOUN |
| Adjective | ADJ |
| Pronoun | Pron |
| Proper noun | X |
| Unknown word | X |
| Adjectival modifier | amod |
| Case marking | case |
| Clausal complement | ccomp |
| Determiner | det |
| Indirect object | iobj |
| Nominal modifier | nmod |
| Object | obj |
| Subject | nsubj |
| Oblique nominal | obl |
| Open clausal complement | Xcomp |
| Root | Root |
| Nominative | Nom |
| Accusative | Acc |
| Genitive | Gen |
| Feminine | Fem |

| | |
|---|---|
| **Masculine** | **Masc** |
| **Singular** | **Sing** |
| **Dual** | **Dual** |
| **Plural** | **Plur** |
| **imperfect** | **Imp** |
| **Perfect** | **Perf** |
| **subjunctive** | **Sub** |
| **coordinating conjunction** | **cc** |
| **conjunct** | **conj** |
| **unspecified dependency** | **dep** |
| **passive nominal subject** | **nsubj:pass** |