Tibyan Automated Arabic Correction Using Machine-Learning in Detecting Syntactical Mistakes

Ashwag O. Maghraby, Nida N. Khan, Hosnia A. Ahmed, Ghufran N. Brohi, Hind F. Assouli, Jawaher S. Melibari

Abstract**—**The Arabic language is one of the most important languages. Learning it is so important for many people around the world because of its religious and economic importance and the real challenge lies in practicing it without grammatical or syntactical mistakes. This research focused on detecting and correcting the syntactic mistakes of Arabic syntax according to their position in the sentence and focused on two of the main syntactical rules in Arabic: Dual and Plural. It analyzes each sentence in the text, using Stanford CoreNLP morphological analyzer and machine-learning approach in order to detect the syntactical mistakes and then correct it. A prototype of the proposed system was implemented and evaluated. It uses support vector machine (SVM) algorithm to detect Arabic grammatical errors and correct them using the rule-based approach. The prototype system has a far accuracy 81%. In general, it shows a set of useful grammatical suggestions that the user may forget about while writing due to lack of familiarity with grammar or as a result of the speed of writing such as alerting the user when using a plural term to indicate one person.

Keywords**—**Arabic Language acquisition and learning, natural language processing, morphological analyzer, part-of-speech.

I. Introduction

ARABIC language is one of the most important languages that exist for thousands of years. Arabic now is the fourth widely spoken language worldwide; it is spoken by 315 million people around the world [1]. The biggest challenge in Arabic lies in its structure. Words in Arabic are graphically similar with each other and they may also contain more than one part-of-speech (POS) at the same word without separation like the word (أَنُلْزِمُكُمُوهَا); here we can find verb, a hidden pronoun and other POSs in one word. Due to the previous reason and more others, Arabic language is one of the most difficult languages even for those who consider it as their native language.

Nowadays, many researches have been done on Arabic syntax analysis, detecting and correcting spelling (syntax) mistakes. A few researches are concerned with Arabic syntactic analysis, in terms of detecting and correcting syntactical mistakes and those researches are still in the beginning. Authors in [2] use two approaches: learning-based approaches and rule-based approaches in order to analyze grammar in Arabic language sentences (إعراب الجمل) without detecting and correcting the Syntactical mistakes. GramCheck [3] detects syntactical errors in the sentences and provides a description for each error. However, the system cannot detect the free order of Arabic sentence and cannot modify the incorrect diacritic signs (علامات التشكيل) of the sentences.

This research is concerned with detecting and correcting Arabic syntactical mistakes including incorrect diacritic signs. It proposed a model of detecting Arabic grammatical errors using the SVM algorithm and correcting them using the rule-based approach. A prototype which is called "Tibyan" of the proposed system which detects and corrects Dual and Plural in Arabic sentences was implemented and evaluated. The prototype system has a far accuracy 81%.

II. Dual and Plural in Arabic Language

Dual and Plural Arabic is one of the most used dialects in daily conversations. Dual subject pronoun refers to two people or two things, and the plural subject pronoun refers to more than two people or things. The formation of dual and plural is difficult because there are many patterns through which the word can be formed in Arabic Language.

A. Dual

"The dual is used for pairs, namely for two individuals or things of the same kind or class" [3]. For example, two books (كتابان), two apples (تفاحتان), two rings (خاتمان), etc. The dual is formed by replacing the endings of the singular form with the two following suffixes:

- /ـان/in the nominative.
  Ex: الطالبان يكتبان الدرس
  The word (الطالبان) ended with /ـان / because it is nominative for being a subject of a nominal sentence.
- /ـين/ in the genitive and accusative.
  Ex: نسلمُ على المسافرين
  The word (المسافرين) ended with /ـين / because it is genitive, for coming after a preposition which is (على)

B. Plural

There are two plural types in Arabic [3]:

1. "The sound plural (الجمع السالم) may be compared to the English external plural or regular plural" [3]. For example, teachers (معلمون), engineers (مهندسون), readers (قارئون), containing two types:

a. The sound masculine plural (جمع المذكر السالم): It is formed by replacing the case endings of the singular form with the two following suffixes [3]:

- /ـون/ for the nominative.
  Ex: السائحون تنزهوا في المدينة
  The word (السائحون) ended with /ـون / because it is nominative for being a subject of a nominal sentence.

Ashwag. O. M is assistant professor in College of Computer Science and Information System, Makkah, UMM Al Qura University, Saudi Arabia (e-mail: aomaghraby@uqu.edu.sa).

Nida. N. K, Hosnia. A. A, Ghufran N. B, Hind F.A, and Jawaher S. M. are with College of Computer Science and Information System, Makkah, UMM Al Qura University.

- /ـين / for the accusative and genitive.

  Ex: نُسلم على المسافرين

  The word (المسافرين) ended with /ـين / because it is genitive, since it came after the preposition (على).

b. The sound feminine plural (جمع المؤنث السالم):

For example, teachers (معلمات), engineers (مهندسات), readers (قارئات). It is formed by adding the following two suffixes to the singular word ending [3]:

- /ـات ُ / for the nominative, Dammah is used on the last letter.

  Ex: الطالباتُ يكتبن الدرس

  The word (الطالباتُ) has Dammah on its last letter because it is nominative for being a subject of a nominal sentence.

- /ـاتِ / for the accusative and genitive, Kasrah is used under the last letter.

  Ex: نُسلمُ على المسافراتِ

  The word (المسافراتِ) has Kasrah under its last letter because it is genitive, for coming after a preposition which is (على).

2. The broken plural (جمع التكسير) "Broken plurals are formed from the singular by internal changes and/or specific increments according to some thirty different patterns. There are hardly any rules about how to form the broken plural from the singular." [3]. For example, books (كتب), students (طلاب), chairs (كراسي), Short vowels are used with the broken plural according to the position or case of each word.

- Ex: الطلابُ يكتبون الدرس

  The word (الطلابُ) has Dammah on its last letter because it is nominative for being the subject of a nominal sentence.

- Ex: قرأتُ الكتبَ

  The word (الكتبَ) has Fathah on its last letter because it is Accusative for being An Object of the sentence.

- Ex: مشيتُ في الطرقِ

  The word (الطرقِ) has Kasrah under its last letter because it is genitive, for coming after the preposition (في).

- Ex: مشيتُ في الطرقِ

  The word (الطرقِ) has Kasrah under its last letter because it is genitive, for coming after the preposition (في).

III. Tibyan Structure

Tibyan analyzes the Arabic text entered by the user, detects the syntactic mistakes and corrects them. Fig. 1 shows the system structure which is divided into two stages: Syntactic Mistakes Detection and Syntactic Mistakes Correction.

Syntactic Mistakes Detection includes a detection of the syntactic mistakes through three steps: (1) Segmentation, where the entered text is separated into sentences via Arabic punctuations marks; (2) Morphological Analysis, in that a POS is specified for each sentence; (3) Mistake Detection, which includes a detection of the syntactic mistakes using SVM algorithm. Syntactic Mistakes Correction corrects all the mistakes detected in the previous stage according to its label and POS tagger.

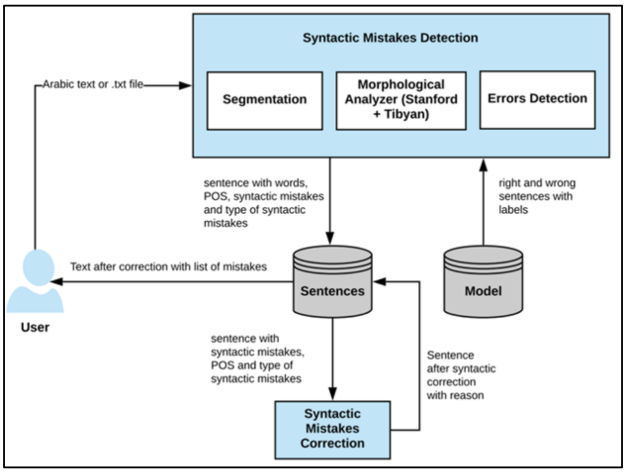

Syntactic Mistakes Detection
Arabic text or .txt file
Segmentation
Morphological Analyzer (Stanford + Tibyan)
Errors Detection
sentence with words, POS, syntactic mistakes and type of syntactic mistakes
right and wrong sentences with labels
Text after correction with list of mistakes
Sentences
Model
User
sentence with syntactic mistakes, POS and type of syntactic mistakes
Sentence after syntactic correction with reason
Syntactic Mistakes Correction

Fig. 1 Tibyan Overall Structure

A. Syntactical Mistake Detection

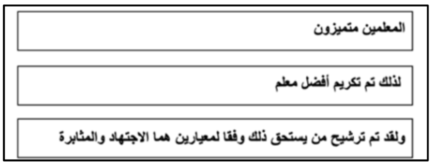
Syntactic cases are markings (حركات/Harakat) or letterings (حروف/Hurof) that are always attached to the end of a word in order to determine the syntactic case. In this stage, Tibyan will analyze the input text in order to classify the sentences to multiple categories depending in syntactic case using three steps:

1. Segmentation

This step is considered as a preprocessing for the next stage (Morphological Analysis). The inserted text consists of multiple sentences, which are separated into individual sentences in order to be stored in a database for the next stage. Therefore, in the segmentation stage, Tibyan will divide the whole input text into separated sentences according to Arabic punctuation marks which are {. ,؟ , … ,, , ؛, "",-,[],=, :}. The text in Fig. 2 is separated into individual sentences as shown in Fig. 3.

2. Morphological Analysis

There are various Morphological analyzer tools which are available to use. We used Stanford CoreNLP.

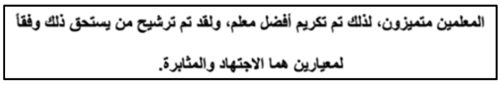


المعلمين متميزون

لذلك تم تكريم أفضل معلم

ولقد تم ترشيح من يستحق ذلك وفقا لمعيارين هما الاجتهاد والمثابرة

Fig. 2 Arabic Sentence example



المعلمين متميزون، لذلك تم تكريم أفضل معلم، ولقد تم ترشيح من يستحق ذلك وفقاً
لمعيارين هما الاجتهاد والمثابرة.

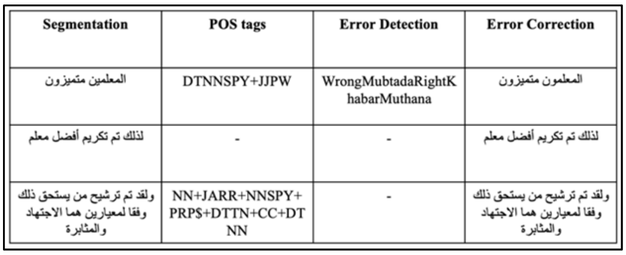
Fig. 3 Arabic Sentence after segmentation

TABLE I
WORDS OF SENTENCE”المعلمين متميزون” AFTER POS

Word	POS
المعلمين	DTNNSPY (stands for noun with determiner)
متميزون	JJPW (stands for adjective)

TABLE II
WORDS OF SENTENCE” وفقا لمعيارين هما الاجتهاد و المثابرة” AFTER POS

Word	POS
وفقا	NN (stands for noun)
لـ	JARR (stands for preposition)
معيارين	NNSPY (stands for proper noun or plural)
هما	PRP$ (Possessive pronoun)
الاجتهاد	DTNN (stands for noun with determiner)
و	CC (Coordinating conjunction)
المثابرة	DTNN (stands for noun with determiner)



Segmentation	POS tags	Error Detection	Error Correction
المعلمين متميزون	DTNNSPY+JJPW	WrongMubtadaRightKhabarMuthana	المعلمون متميزون
لذلك تم تكريم أفضل معلم	-	-	لذلك تم تكريم أفضل معلم
ولقد تم ترشيح من يستحق ذلك وفقا لمعيارين هما الاجتهاد والمثابرة	NN+JARR+NNSPY+PRP$+DTTN+CC+DTNN	-	ولقد تم ترشيح من يستحق ذلك وفقا لمعيارين هما الاجتهاد والمثابرة

Fig. 4 Methodology steps

There are various morphological analyzer tools which are available to use. We use Stanford CoreNLP tagger for many reasons: (1) Its simplicity and ease of use, (2) it can be integrated smoothly as a Python library (implementation language used to developed Tibyan), (3) it performs any of the morphological analyzing steps easily by using the appropriate function, (4) the most Arabic morphological analyzer with minimum number of wrong results compared to the others and finally, (5) its results can be improved by using it along with NLTK python library.

Tagging in Stanford CoreNLP is done according to each word's syntactical categories, such as verbs, nouns, particles, etc. However, Stanford CoreNLP tags did not serve our need enough so, we modified some of them and create *Tibyan Tagger which will be explained in the next section.* The text in Fig. 3 after the POS tagging is shown in Tables I and II.

3. Mistake Detection

In this step, Tibyan classifies each sentence as correct or incorrect according to the model that we have built and does the training by using the SVM. In our previous example, the two sentences that are shown in Tables I and II will be detected; the word المعلمين has syntactic mistake. Therefore, the result will be the sentence and the type of the syntactic mistake.

B. Syntactical Mistake Correction

The correction stage is the stage where Tibyan turns the wrong sentences into right ones. This stage follows the detection stage, so the result of the detection stage which includes the sentences that have syntactic mistakes as well as the type of the syntactic mistakes represents the input in this stage. The end of the wrong word will change according to the classification label and POS. In our example, the wordالمعلمين will be changed to be المعلمون. All methodology stages are illustrated in Fig. 4.

IV. Tibyan Tagger

To analyze Arabic sentences, Stanford CoreNLP cannot be used alone due to several problems related to its results: (1) It cannot distinguish between the dual and the different types of plural such as: the masculine plural (جمع المذكر السالم) and the feminine plural (جمع المؤنث السالم), in that Stanford assigns the same POS tags for each one of them. (2) Stanford cannot recognize the different syntactic cases of the same word. For example, in Arabic grammar the dual nominative case (المثنى المرفوع) of the word “**الكتاب**” in sentence “هذا **الكتاب** ممتاز” is “ هذان **الكتابان** ممتازان” and the dual accusative case (المثنى المنصوب) of the word “**الكتاب**” in sentence “قرأت هذا **الكتاب**” is “ قرأت هذين **الكتابين**”. However, Stanford assigns the same POS tags for both syntactic cases of the word “الكتاب”. (3) Stanford cannot recognize the different syntactic cases of the dual accusative case or genitive case (المثنى المنصوب أو المجرور) and the masculine plural accusative case or genitive case (جمع المذكر المنصوب أو المجرور). It assigns the same POS tags for both of them. (4) Stanford removes diacritic signs from the feminine plural (جمع المؤنث السالم) and assigns the same POS tags for diacritization and non- diacritization words. (5) It also does not differentiate between the types of Arabic tools used in the sentences, such as: conjunction letters or kana and inna and their sisters. Table III shows an example of POS tags for the sentence “ذهب المعلمون إلى المدرسة” after using Stanford CoreNLP.

In the light of these problems, we end up with creating the Tibyan Tagger. In the Tibyan Tagger, the Arabic sentences are initially analyzed using the Stanford CoreNLP in order to get the POS tags of the tokens, and then the work of Tibyan starts. Tibyan tagger works by adding a meaningful letter at the end of the Stanford.

TABLE III
EXAMPLE OF ARABIC SENTENCES USING STANFORD TAGGER

Word	Stanford POS tags
ذهب	VBD
المعلمون	DTNNS
إلى	IN
المدرسة	DTNN

TABLE IV
TIBYAN POS TAGGER

Type and Syntactic case	Tibyan POS tag
• المثنى المرفوع *the dual nominative case*	Stanford POS + DA
• جمع المذكر السالم المرفوع *the masculine plural nominative case*	Stanford POS +PW
• جمع المذكر السالم المنصوب والمجرور *the masculine plural accusative and genitive case*	Stanford POS +PY
• جمع المؤنث السالم المرفوع *the feminine plural nominative case*	Stanford POS +MW
• جمع المؤنث السالم المنصوب والمجرور *the feminine plural accusative and genitive case*	Stanford POS +MY

TABLE V
TIBYAN POS TAGGER FOR DIACRITIZATION WORDS

Type and Syntactic case	Tibyan POS tag
• جمع المؤنث السالم بلا حركة *the feminine plural without diacritic signs*	Stanford POS +M
• جمع المؤنث السالم بفتحة أو تنوين بالفتح *the feminine plural with Fateha*	Stanford POS +MR

TABLE VI
TIBYAN POS TAGGER FOR JARR LETTER AND KAANA AND INNA

Type and Syntactic case	Tibyan POS tag
• حروف الجر *the dual nominative case*	JARR
• كان وأخواتها *the dual accusative and genitive case*	KANA
• إن وأخواتها *the masculine plural nominative case*	INNA

TABLE VII
EXAMPLE OF ARABIC SENTENCES USING BOTH TAGGERS

Word	Stanford POS tags	Tibyan POS tags
ذهب	VBD	-
المعلمون	DTNNS	DTNNSPW
إلى	IN	JARR
المدرسة	DTNN	-

The Tibyan tagger performs three main functions: (1) Rechecking and editing the Stanford POS tags of Dual and Plural tokens according to their classification relating to their types and syntactic cases by adding the appropriate letters at the end of the tag. Table IV shows the changes that take place for the POS tags. (2) Rechecking and editing the Stanford POS tags for feminine plural tokens according to their classification relating to their types, diacritic signs, and syntactic cases by adding the appropriate letters at the end of the tag. Table V shows the changes that take place to the POS tags. (3) The Tibyan tagger also edits the POS tags for Arabic language tools of kana and inna and their sisters (كان وأخواتها وإن وأخواتها) as well as prepositions (حروف الجر). Table VI shows the changes that take place to the POS tags. The result of applying Tibyan tagger for the sentence "ذهب المعلمون إلى المدرسة" is shown in Table VII.

V. DATASET

In order to be able to do the syntactical detection and correction we needed to work with Arabic language dataset. In our case, we did not find any available dataset that is suitable for us. All the available datasets that we found for natural language processing (NLP) were either for other languages or built for other purposes, such as spelling checking. We collected the data by ourselves from Arabic books. We collected 25,157 sentences for the dual and plural cases from Arabic books, such as خرائط ,اسمي أحمر ,عداء الطائرة الورقية ,لا تحزن التيه, etc. [4]-[7].

All the sentences, we aimed to detect and correct, were related to one of the nine Arabic grammars:

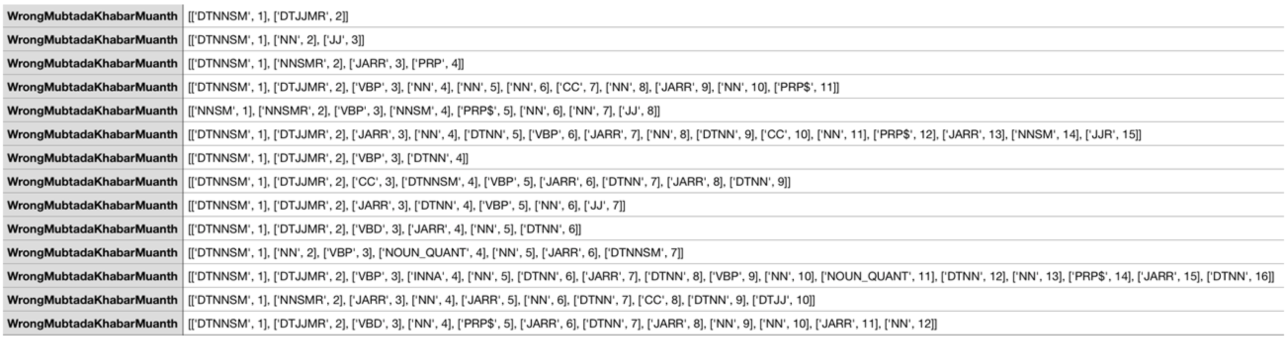
1. The subject (Al-Mubtada - المبتدأ).
2. The predicate (Al-Khabar - الخبر).
3. The active participle (Al-Fael – الفاعل).
4. The accusative object (Al-Mafoul bih - المفعول به).
5. The subject of Inn and its sisters (Isim inna wa akhwatuha - اسم إن وأخواتها).
6. The predicate of Inn and its sisters (Khabar inna wa akhwatuha - خبر إن وأخواتها).
7. The subject of kana and its sisters (Isim kana wa akhwatuha - اسم كان وأخواتها).
8. The predicate of Inn and its sisters (Khabar kana wa akhwatuha - خبر كان وأخواتها).
9. The noun after preposition (Isim Majrur - الاسم المجرور).

Finally, after collecting the correct sentences, we turned these sentences to a version that contains wrong syntactic cases in order to allow the model to learn both correct and wrong cases. We structured our dataset to contain words, the index of the word, POS tags as well as a label (wrong or right) of the sentence in order to give us the best prediction. Our dataset recovers 76 syntactic cases. The structure of the dataset is shown in Fig. 5.

Table VIII shows an example to restructure the sentence كان المتميزون رائعين in order to be appropriate in the dataset.

TABLE VIII
THE STRUCTURE OF THE SENTENCE IN DATASET

Words	Index of words	POS tags	Label of sentence
كان	1	KANA	
المتميزون	2	DTNNSPW	RightIsimKanaRightKhabarKana
رائعين	3	JJPY	
The result:	RightIsimKanaRightKhabarKana, "[['KANA', 1], ['DTNNSPW', 2], ['JJPY', 3]]"		



WrongMubtadaKhabarMuanth	[['DTNNSM', 1], ['DTJJMR', 2]]
WrongMubtadaKhabarMuanth	[['DTNNSM', 1], ['NN', 2], ['JJ', 3]]
WrongMubtadaKhabarMuanth	[['DTNNSM', 1], ['NNSMR', 2], ['JARR', 3], ['PRP', 4]]
WrongMubtadaKhabarMuanth	[['DTNNSM', 1], ['DTJJMR', 2], ['VBP', 3], ['NN', 4], ['NN', 5], ['NN', 6], ['CC', 7], ['NN', 8], ['JARR', 9], ['NN', 10], ['PRP$', 11]]
WrongMubtadaKhabarMuanth	[['NNSM', 1], ['NNSMR', 2], ['VBP', 3], ['NNSM', 4], ['PRP$', 5], ['NN', 6], ['NN', 7], ['JJ', 8]]
WrongMubtadaKhabarMuanth	[['DTNNSM', 1], ['DTJJMR', 2], ['JARR', 3], ['NN', 4], ['DTNN', 5], ['VBP', 6], ['JARR', 7], ['NN', 8], ['DTNN', 9], ['CC', 10], ['NN', 11], ['PRP$', 12], ['JARR', 13], ['NNSM', 14], ['JJR', 15]]
WrongMubtadaKhabarMuanth	[['DTNNSM', 1], ['DTJJMR', 2], ['VBP', 3], ['DTNN', 4]]
WrongMubtadaKhabarMuanth	[['DTNNSM', 1], ['DTJJMR', 2], ['CC', 3], ['DTNNSM', 4], ['VBP', 5], ['JARR', 6], ['DTNN', 7], ['JARR', 8], ['DTNN', 9]]
WrongMubtadaKhabarMuanth	[['DTNNSM', 1], ['DTJJMR', 2], ['JARR', 3], ['DTNN', 4], ['VBP', 5], ['NN', 6], ['JJ', 7]]
WrongMubtadaKhabarMuanth	[['DTNNSM', 1], ['DTJJMR', 2], ['VBD', 3], ['JARR', 4], ['NN', 5], ['DTNN', 6]]
WrongMubtadaKhabarMuanth	[['DTNNSM', 1], ['NN', 2], ['VBP', 3], ['NOUN_QUANT', 4], ['NN', 5], ['JARR', 6], ['DTNNSM', 7]]
WrongMubtadaKhabarMuanth	[['DTNNSM', 1], ['DTJJMR', 2], ['VBP', 3], ['INNA', 4], ['NN', 5], ['DTNN', 6], ['JARR', 7], ['DTNN', 8], ['VBP', 9], ['NN', 10], ['NOUN_QUANT', 11], ['DTNN', 12], ['NN', 13], ['PRP$', 14], ['JARR', 15], ['DTNN', 16]]
WrongMubtadaKhabarMuanth	[['DTNNSM', 1], ['NNSMR', 2], ['JARR', 3], ['NN', 4], ['JARR', 5], ['NN', 6], ['DTNN', 7], ['CC', 8], ['DTNN', 9], ['DTJJ', 10]]
WrongMubtadaKhabarMuanth	[['DTNNSM', 1], ['DTJJMR', 2], ['VBD', 3], ['NN', 4], ['PRP$', 5], ['JARR', 6], ['DTNN', 7], ['JARR', 8], ['NN', 9], ['NN', 10], ['JARR', 11], ['NN', 12]]

Fig. 5 Tibyan Dataset

VI. Tibyan Prototype

We used Python to develop Tibyan. Fig. 6 showS the flow from one screen to another. Screen 1 is the main interface for the software where users can choose to upload a file from their devices or type the text directly to the software. If the users choose to upload a file, they should select the file location on their computer using screen 2. The chosen file should be in the correct format (.txt). The content of the uploaded file should be written in Arabic; otherwise an error message is displayed. Whether the user chooses to upload a file or type a text directly, the screen 3 is displayed. Users can then click the button to start detecting errors. If the software detects any syntactic mistakes in the inserted text, screen 4 is shown. Each mistake is labeled in red color. The correction for each mistake is shown as a list in the left side of the interface. By default, all the corrections are selected to be in the corrected file. However, users can choose the mistake that should be included in the corrected file. If there is no mistake in the input text, screen 5 is displayed. After that, users can choose if they want to accept the corrections by clicking on the green button or they can reject them by clicking on the red button. If the user wants to accept the corrections, the software asks him/her if she/he wants to save the file or not in screen 6.

VII. Conclusion and Future Work

Unlike other languages, Arabic grammatical detection and correction researches and systems are still in the beginning. This project proposed a model of detecting Arabic grammatical errors using the SVM algorithm and correcting them using the rule-based approach. A prototype of the proposed system which detects and corrects Dual and Plural in Arabic sentences was implemented and evaluated.

As future work we would like to continue our work and improve Tiyban by extending it to detect all the Arabic syntactical mistakes. We would also like to explore using Recurrent Neural Network to correcting the syntactical mistakes. We are planning to develop our own neural network model and consider our problem as a translation-related problem. Also, we would like to add spelling correction for detect.

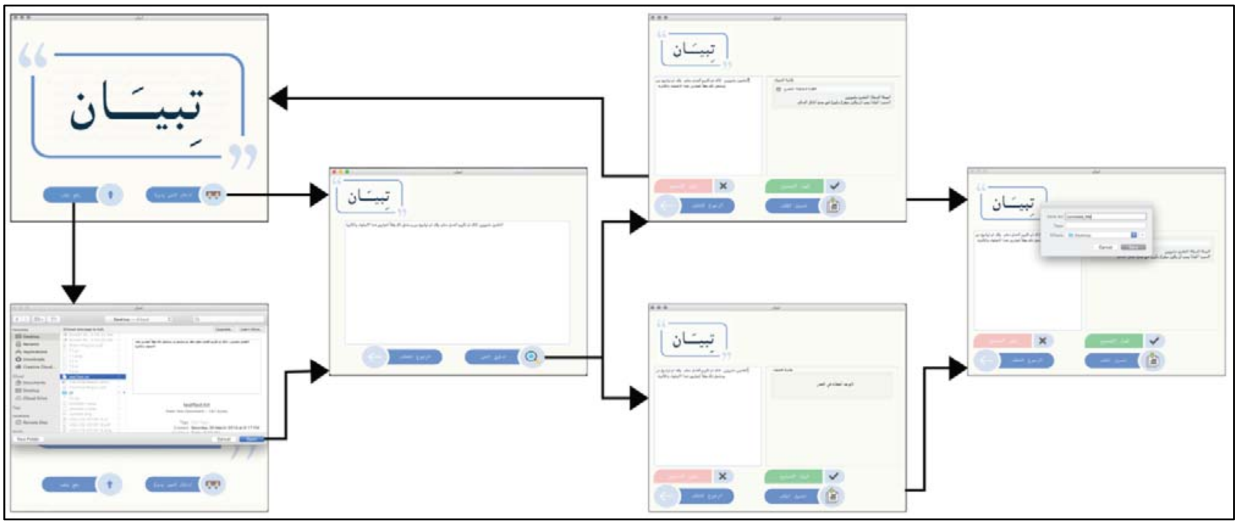


Fig. 6 Tibyan Porotype Interface

References

[1] Accredited Language Services. Accredited Language Services. [Online) Available from: https://www.accreditedlanguage.com/ (Accessed 7th November 2018).

[2] Ibrahim, M.N. Mahmoud, M.N. & El-Reedy, D. Bel-Arabi: Advanced Arabic Grammar Analyzer (Online) 6, 341-346. Available from: doi:10.7763/IJSSH.2016.V6.669 (Accessed 15th July 2018), 2016.

[3] Abu-chacra F, Arabic an essential grammar. (Online) Available from: https://www.pdfdrive.com/arabic-an-essential-grammar-e23160514.html (Accessed 14th October 2018), 2007.

[4] AlQarni, A, “لا تحزن”. (Online) Available from: http://www.waraqat.net/2008/11/3aeez_lat7zn.doc (Accessed in 12sd February 2019), 2003.

[5] Hesaine, K, “عداء الطائرة الورقية”. (Online) Available from: https://jamalon.com/ar/1007470.html (Accessed in 12sd February 2019), 2013.

[6] Bamoq, O, “اسمي أحمر”. (Online) Available from: https://www.goodreads.com/review/show/187028440 (Accessed in 12sd February 2019), 2006.

[7] Alisa, B, “خرائط التيه”. (Online) Available from: https://www.goodreads.com/book/show/25635774 (Accessed in 12sd February 2019), 2015.