Available on: https://jsshrs.ep-journals.org/ E-ISSN: 3051-3561, P-ISSN: 3051-3553

Journal of Social Science and Human Research Studies

Volume : 02 Issue : 08 August-2026

Grammatical Tagging of Arabic Word for Corpus Analysis

Zainur Rijal Abdul Razak
Fakulti Pengajian Bahasa Utama, Universiti Sains Islam Malaysia, 71800 Nilai, Negeri Sembilan, Malaysia

Yuslina Mohamed
Fakulti Pengajian Bahasa Utama, Universiti Sains Islam Malaysia, 71800 Nilai, Negeri Sembilan, Malaysia

Published Date: August 31, 2026 **Article DOI: 10.65150/EP-jsshrs/V2E8/2026-31**

ABSTRACT : Word tagging is one of the important processes in computational linguistics studies. It provides detailed information about the grammatical part of speech (PoS) to the researcher. However, tagging process is not an easy process particularly when relating to Arabic text. One of the main reasons that led to this situation is the fact that many Arabic word characters (space between characters) could contain more than one word. Thus, this paper aims to revisit our experience in tagging Arabic words and to highlight critical issues related to the process. It also aims to validate the possibility of using computer analyzer software, namely Wordsmith 5.0 for Arabic texts. Using several Arabic texts from newspapers as examples, this paper reveals how the tagging process was executed using a self-constructed tag set. It also discusses the issues and linguistic factors considered in constructing the tag sets. This includes the differences that exist between Arabic word characteristics in writing and their articulation. This paper finally stresses the more complicated issues in tagging Arabic words in relation to the fact that Arabic has the affixation and diacritics systems. The application of short vowels and orthographic diacritics in Arabic language also contributes to the phenomenon. Despite this fact, this paper also proves that a researcher can make the tagging process simpler by focusing on certain grammatical features, aiming the tagging for word disambiguation and applying an appropriate analyzing software.

KEYWORD: Grammatical tagging, Arabic automatic analysis, corpus linguistics, computational linguistics, tag set.

1. INTRODUCTION

Word tagging in text is very important for various applications in natural language processing (NLP) (Suvanto Minerva, 2025; Abdelhamid, Abdellah and Aouragh, 2011) including the automatic computational analysis studies as it enables the researcher to analyze Arabic words in texts given in high accuracy. According to Nizar Habash (2010), tagging process or Part-of-speech (PoS) tagging is the task of assigning a contextually appropriate morpho-syntactic tag to every word in a sentence. Without the process, the researcher would not know what grammatical features that exist in the analyzed words as they are treated separately from contexts by the analysis software. As a result, the systematic analysis on language used could not be performed accordingly. The reason is that the computer software normally will consider the analyzed word as a single character regardless of its context. According to Baker, Hardie and McEnery (2006), the tagged corpus is a corpus with the additional information added to the corpus data that enables the researcher to carry out a more sophisticated analysis.

Therefore, the more detailed tagging process is performed, the more accurate grammatical information is obtained by the researcher. However, it will depend on the aim and objective of a particular study to what extent the study is carried out in language use. This is because tagging is a pre-step process before the next step language style analysis and the aim of the analysis has great contribution in deciding on how deep the tagging is needed in the corpus.

2. TYPE OF TAGGING

Generally, there are two types of corpora tagging. Both play roles in giving specific information of the analyzed words in the corpus but in different perspectives. The first type of corpus is called the tagged corpus, which is a corpus additional information added to the corpus data that enables the researcher to carry out a more sophisticated analysis (Baker et al, 2006). The type of tagging is normally assigned to words according to their grammatical categories. Specific tag sets represent the morphological, syntactical or semantic characteristics of words or the characteristics of constituents or sentences. To fully use the computer's potential, the linguistics nature of the text must be stored in the computer to analyze and process large amount of natural language data (Vinay Dwivedi, (2024).

The second kind of tagging is called a parsed corpus. The parsed corpus is somewhat like the first kind of tagging, especially in terms of purpose and method of application. However, there are slight differences between them as the corpus is parsed to indicate its syntactic structure. The words tagged in parsed corpus are therefore not treated as separate grammatical categories, as in the tagged corpus, but are judged from the

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

start and end points of units such as noun phrases, verb phrases and clauses. The parsed corpus may also add information about how the syntactic units relate to one another (Baker et al, 2006).

3. AUTOMATIC TAGGING AND MANUAL TAGGING

Tagging the corpus normally is not an easy task especially when the researcher decides to tag them manually. In normal practice, automatic tagging is preferred as it would save a lot of time and avoid human error, while the manual tagging is only reasonable for small sized data. As the corpus normally contains thousands or even millions of words that need to be tagged, the automatic tagger is preferable to all researchers.

Nevertheless, the fact does not deny the role of manual tagging in providing sufficient grammatical information to the researcher especially when the data is very small and untagged before. In such a situation, human manual tagging is inevitably called for directly tagging each token (Antonio Branco and Joao Silva, n.d.). The data containing several thousand words might not be very hard to tag manually. However, it should be stressed that the smaller the size of the corpus, the smaller its probability to be generalized to represent the real language use. Realizing that the aim of this paper was to revisit our experience and highlight how the tagging process plays role in Arabic word analysis through examples from several texts and some related issues, the size of the corpus is not a big deal.

The automatic tagging, although has more benefits for this kind of studies, the researcher or corpus compiler first must make sure the automatic tagger software is available to him or her. A lot of preliminary studies must be done to decide what type of software is compatible for his or her study. It should also be mentioned that several taggers are not compatible with Microsoft Windows (WS) or MAC but only operate on Linux operating system. Therefore, the in-depth knowledge relating to the computer operating system is also highly required before embarking on the use of automatic tagging software. There are several automatic taggers that can do Part of Speech (PoS) tagging with more than 90% accuracy, such as MADA+TOKAN, Automatic Part of Speech Tagger (APT), Support Vector Machine (SVM), TAGARAB and others (Zainur Rijal, 2011). Today, systems for automatic tagging of word groups (morphological analysis, word-class tagging) and automatic tagging of sentence fragments (syntactic analysis, parsing) have been developed (Abdurashetovna, 2023)

In contrast, manual tagging is more flexible to the researcher with limited knowledge about computer programming and operating systems. Besides, manual tagging will enable him or her to set how detailed tagging is needed according to the aim of studies. For instance, if the researcher decided to analyze the verb in terms of perfect and imperfect, he or she might not need the tagging in terms of transitive and intransitive verbs, while the automatic taggers will tag them all in detail without exception. As the data of corpus is stored in Microsoft Word, the process of manual tagging can be done in the same operating system.

4. ISSUES IN ARABIC WORD TAGGING

Arabic language is a Semitic language with rich templatic morphology (Mona Diab, Kadri Hacioglu and Daniel Jurafsky, 2004). It uses a clitic sound system which is implicitly embedded to the Arabic written text. Because of this character in the language, many Arabic words are similar in writing but different in pronunciation. This leads to the problem of ambiguity which proves the extreme need of grammatical tagging. According to Tlili-Guissa Yamina (2005), ambiguity in Arabic is enormous at every level; the absence of the representation of short vowels in normal texts increases dramatically the number of ambiguities.

However, this problem does not only exist in Arabic language, t relatively in all world languages. However, the fact that Arabic language faces this more complicated problem than other languages, it justifies the need of tagging in computational linguistics studies. This fact is due to several words in Arabic are non-vocalized text that may have more than one tag due to ambiguous words and unknown words in which the classification may depend on the word meaning (Boudchiche & Mazroui, 2015). Therefore, Nizar Habash and Owen Rambow (2005) have listed at least four issues relating to Arabic language problem in word tagging. They notified that Arabic is a morphological complex language that needs around 330,000 theoretically possible completely specified morphological analysis (Nizar, 2010), while the English normally need around 50 tags only. Sandra Kubler and Emad Mohamed (2011) strengthen the fact by saying that there are a high number of inflections per word; and the tag set is larger than the typical tag set for English.

This issue arises because some functional grammatical words such as conjunctions and pronouns can be combined to either noun or verb in one word character. The word ليسألهم (*to ask them*), for instance, is built by three words هم - يسأل - لـ (*to – ask – them*) that are combined in one character. This would cause a different tagging from the word ليسأله (to ask him) due to differences in pronoun (هم) and (هـ). Secondly, they added that the Arabic orthographic rules caused some parts of words to be deleted of modified when cliticization occurs. For example, the *Ta-Marbuta* (ة) appears as a regular Ta (ت) when followed by a pronominal clitic. Simple segmentation of the pronominal clitic without recovering the *Ta-Marbuta* could cause unnecessary ambiguity or add to the sparsity problem. Arabic also is written with optional diacritics that specify short vowels and germination (consonant doubling) and they are usually absent in written Arabic, which contribute to the ambiguity. Finally, the writing system also shows different levels of specificity in spelling some letters, for example, *Alif-Hamzah* above (أ) can be spelled without the *Hamzah* as (ا) and *ya* (ي) can be spelled without the dots as *Alif-Maqsura* (ى).

As our work in tagging the text is specifically for the use of specific analysis software, some of the problems mentioned above are automatically solved by the special feature of the software. For instance, the problem of difference between *ya* (ي) with dots and *ya maqsura* (ى) without dots would be automatically separated by the software because of the different writing physically. The first and second problems are solved when the separation process is used to separate between one character that contains more than one word such as ليسألهم (to-ask-them) above. However, there still needed tagging process to be done especially to solve the ambiguity issue in the third problem according to Habash and his colleagues' views. The tagging process therefore puts only the basic grammatical feature to cover the aim of this paper to demonstrate how the tagging process plays its role.

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

5. TAGGING PROCESS

As our job in our original corpus and example of analysis in this paper is exactly similar, we found that tagging process in both involved two steps, firstly separating between words into one character by putting spaces between them, and secondly, putting grammatical tags for the ambiguous words which expectedly could cause confusion by the analyzer software.

5. 1 Separation between words

This process is, however, only related to the single character that contains more than one word. In order to do this, we must define accurately what the word is. According to (David Crystal, 2003), words in writing are defined as lexical units in a string of writing (bounded by spaces). However, this identification is not always accurate as some lexical units are made up of more than one word. In English, for example, there is sometimes more than one word between two spaces or blanks: e.g. the word 'didn't' contains two words 'did' and 'not'.

As discussed, Arabic words are full of connected groups of letters that contain more than one word and the software only identifies words by picking out clusters between blanks. The tagging process was therefore mostly focused on separating the words in such groups, and this was done by placing a space between the individual word components to enable the software to recognize each word separately. As mentioned, the problem in Arabic is more complicated because individual words are very often not separated by spaces. This occurs more frequently in Arabic than in English. As mentioned before, the word ليسألهم (*to ask them*), for instance, actually contains three words: هم - يسأل - لـ (*to – ask – them*).

For the automatic analysis, using computational software on word with such this feature in Arabic language could cause unreliable output for our specific purpose and creates bigger trouble during the analysis process. This is because the software we use (Wordsmith) software was built to recognize words by identifying characters between spaces. Some of the problems that the stemmer faces are that some letters that appear to be affixes are in fact part of the word. Another problem is that some letters (for instance the long vowels) may change to other letters when an affix is added, and so the letters should be changed back when that affix is removed.

In relation to this issue, Abdulkarim (1990) suggested solving this problem by dividing the word up and placing a space between the individual components. Thus, instead of treating ليسألهم as one word, the researcher should write هم - يسأل - لـ to make the components accessible to computer software. However, al-Jaburi (1988) suggested transcribing Arabic into Latin fonts, as computer programs have many problems with Arabic fonts. Another solution proposed by 'Abdulkarim is to set up word lists in which every word is represented with all the possible prefixes and suffixes, which can be manually removed afterwards (Van Mol, 2003, 148). However, this method is impractical and time consuming because the researcher did not realize that so many combinations of different prefixes and suffixes were possible.

Despite all these issues, word identification by recognizing characters between spaces is the best method for the analysis as the software applied uses this method for analyzing. Since the selected software can identify words by recognizing groups of characters between spaces, the most suitable approach is to apply Abdulkarim's first suggestion: to separate all the components of a word by placing spaces between them. It should be noticed that this is only feasible when the capacity of the corpus (data) is small while the large data practically uses the automatic tagging.

The discussion above indicates that the encoding or tagging process would take place manually for this corpus. The researcher is aware that there are some automatic tokenization and tagging software programs that can do Part of Speech (PoS) tagging with more than 90% accuracy as mentioned before, such as MADA+TOKAN, Automatic Part of Speech Tagger (APT), Support Vector Machine (SVM), TAGARAB and others. However, the researcher found that some of these automatic taggers are not easily available online and some of them are not applicable with window Microsoft. We also found that most of the taggers are too detailed for extracting Arabic words into grammatical tagging in text. Extremely detailed tagging could cause the analysis software (Wordsmith Software) to extract the words contrary to our aims of study. One of the aims in our original work was to highlight the trend in word usage in Arabic media without giving more focus on grammatical features except in a few cases. Therefore, the researcher demonstrates again all works that had been done in this paper by giving examples of analysis to explain several issues we faced in tagging process.

5.2 Tagging the ambiguous words

Tagging was applied with the specific intention of ensuring that the corpus was compatible with the software, so that it could statistically analyze the language aspects that were being targeted. In our view, the amount of detail required in the tagging of the corpus depends on the objective of its compilation. Even though this present corpus is categorized as a raw corpus, it still needs simple encoding, although without involving the extent of grammatical tagging used in a tagged corpus as mentioned before. The purpose of this simple tagging is merely to enhance the capability of the software, which is unable to differentiate between words with similar spellings in different word categories, or what is called in corpus linguistics words with ambiguity.

Any corpus may be exposed to the problem of word ambiguity, which could affect the results. Tagging, besides providing additional grammatical information, is the main tool to avoid problems of ambiguity. According to Habash and Rambow (2005), Arabic words are often ambiguous in their morphological analysis because of Arabic's rich system of affixation and diacritics and the omission of disambiguating short vowel and orthographic diacritics in standard orthography (Abdelkader et al., 2024; Al-Thaani & Abu al-Rub, 2009; Habash and Rambow, 2005). The ambiguity problem occurs when there is a choice of two potential tags. For example, the word تحدث (*t-h-d-th*) might be a noun, depending on the context and the (unseen) vocalization. Word analysis software would normally treat it as one word without differentiating the variations. If this problem is not solved at an earlier stage, it will create defects in the analysis results (Scott, n.d).

The tagging process, therefore, focuses on solving this crucial problem by tagging several words using a very simple tag set in to make ambiguous words distinguishable by the software. That is because, although it has been mentioned that Arabic actually, needs about 330,000 tags for a complete tag set, reducing the tag set is always a focus of Arabic researchers. This explains why many researchers prefer to work with smaller reduced sets as long as the analyzer software able to predict the grammatical features in analyzed word accurately (Nizar, 2010). Our aim of tagging merely to solve the ambiguity problem helps us reduce the size of the tag set. The internal process of tagging takes place by putting a hyphen between the word and the tagset. In the previous example تحدث might be several forms of a verb (*tahaddatha*, *tahduthu*, *tuhdithu*) or a noun

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

(*tahadduth*). It was tagged as تحدث_فمعم for *tahaddatha*, تحدث_فضدث for *tahduthu,* تحدث_فضما for *tuhdithu* and تحدُّث_اعدف for *tahadduth*. The table below shows an example of possible ambiguous words with their tags.

Table 1: The example of word with possible tagging

Word	Ambiguous word	Grammatical feature	Tag
ت – ح – د – ث	تحدث – tahaddatha	verb – past – augmented – Ain فعل – ماضي – مزيد – معلوم	تحدث_فمعم
	تحدث – tahduthu	Verb – present – unaugmented - triliteral فعل – مضارع – مجرد – ثلاثي	تحدث_فضدث
	تحدث – tuhaddithu	Verb – present – augmented – Ain فعل – مضارع – مزيد – معلوم	تحدث_فضمع
	تحدث - tahadduth	Noun – general – infinitive - singular - اسم – عام – مصدر – مفرد	تحدث_اعدف
	تحدث – tuhdithu	Verb – present – augmented – Alif فعل – مضارع – مزيد – الألف	تحدث_فضما

Previous studies shown that noted that any tag set should reflect the word syntactic category based on its context for the purposes of resolving lexical ambiguity (Jamilu Awwalu et al., 2020; Fatma Al-Shamsi & Ahmed Guessoum, 2006). In line with the fact, our tag sets explained in Table 1 were set up by four Arabic alphabets to represent each grammatical feature beginning from the general word classification to the more specific. For the first letter, ف (*fa*) stands for فعل (*verb*), ا (*alif*) for اسم (*noun*) and ح (*ha*) for حرف (*particle*). These letters come first right after the hyphen in to differentiate between these three major categories of Arabic words. The second letter stands for the tense, in the case of a verb – either perfect or imperfect while the third letter stands for unaugmented (*mujarrad*) and augmented (*mazid*) verb (please see appendix for the full tagset).

At this stage, the possibility of ambiguous word occurrence still exists although it becomes slimmer which justifies the need of the fourth letter. That is because; several augmented verbs still have similar writing even though they might be different in pronunciation. For example, the verb *tuhaddithu* and *tuhdithu* above show that both are augmented verbs. To solve the problem, the fourth letter comes to differentiate between augmented verbs by clarifying the additional letter (العين for the verb *tuhaddithu* and الألف for the verb *tuhdithu*).

The case of nouns is even simpler because most of the ambiguous words can be differentiated with only the first three letters in the tag set. However, the standardization between nouns and verb makes the tag set in nouns contain four letters as occurred in verbs. A similar principle was applied to the tag set forming in noun starting from general noun classification to the specific. It starts with clarifying the noun representing by vowel (ا) that stand for اسم (*noun*), then explanation noun grammatical features such general noun or specific and the third letter comes to represent mostly infinitive or adjective, meanwhile, the last letter represents the singular, dual and plural. The representation of each letter on grammar classification is shown in Table 2 below:

Table 2: The representation of Arabic letters to word classification in tag set building

No	Representativeness	Word classification	English
1	ا	اسم	Noun
2	ف	فعل	Verb
3	خ	خاص	Specific
4	ع	عام	General
5	م	مكان	Place
6	غ	غير	Other
7	ص	صفة	Adjective
8	ف	مفرد	Singular
9	ث	مثنى	Dual
10	ج	جمع	Plural
11	د	مصدر	Infinitive
12	ض	مضارع	Present
13	م	ماضي	Past
14	ا	أمر	Affirmative
15	د	مجرد	Inaugmented
16	م	مزيد	Augmented
17	ج	مجهول	Passive
18	م	معلوم	Active
19	ذ	مذكر	Masculine
20	ن	مؤنث	Feminine
21	ل	فاعل	Subject
22	ت	تفضيل	Comparative

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

Table 2 above shows the representation of Arabic alphabets to grammatical features. As we can see that not all alphabets represent the first sound of words, they represent e.g. alphabet ف is for مفرد. The most important thing, the table shows that some grammatical classes share the same alphabet, for example, the sound م stands for ماضي (*past*), مزيد (*augmented*) and مكان (*place*), while the sound ج represents the grammatical class of مجهول (*passive*) and جمع (*plural*). However, it is not the case as these similar alphabets either located in different places in the tag set or already differentiated by their major word categories before in the first or second alphabet. For instance, the sound م for ماضي comes second after the sound ف for فعل (verb), the representative of مزيد comes only at third in verb category and the representative of مكان comes third in noun category.

In short explanation, this situation can be described in Table 3 below:

Table 3: The example of similar representativeness to different tagging

No	Representativeness	Grammatical Classification	Example of Tag	Tag Representative
1	م	ماضي	فمدث	فعل – **ماضي** – مجرد – ثلاثي
2	م	مكان	اخمف	اسم – خاص – **مكان** – مفرد
3	م	مزيد	فضمع	فعل – مضارع – **مزيد** - العين

Table 3 shows that all sounds of م were used in different places and situations as explained before. The classification of مكان and مزيد which both come third were already separated by a major word classification in the first alphabet between verb and noun.

However, the analysis using the Wordsmith software gives the researchers an alternative to clarify between the ambiguous words rather than merely depending on tagging. The concordance tool in the software, which can search and display all sentences containing a selected word, was used to make it possible to clarify the category of the word by looking at its context. The software recognizes words by their physical appearance, not their semantic meaning. Using the concordance to find what the neighboring words in the sentence are, is a useful method for solving the ambiguity problem besides tagging.

5.3 Example of the Role of Tagging in Arabic Text Analysis

Nevertheless, the researcher strongly believes that the tag sets built have capability to give enough information on grammatical features in Arabic texts for this kind of study especially relating to verb and noun. Although the encoding process may appear to be applicable only for the researcher's analysis, it is believed it can be improved for the use of other researchers working in other aspects of linguistics and that it will be useful for Machine Learning research, as a training corpus to develop an automatic system to learn tokenization that can be applied to new texts, especially with the existence of established tag sets.

For the time being, the larger corpora which normally contains thousands or millions of words would have to make use of automatic tagger. It is almost impossible to tag them manually as it is time consuming and exposed to human error. To make the analysis simpler, tagging process did not involve all particles as they were not included in the study objectives, even they may cause ambiguity either, for example the particle ل للتعليل and ل – حرف جر were left untagged. Since the tagging process is done manually, the researcher had tried to minimize the process by excluding all particles from the analysis.

The example of the tagging process involves 5 newspaper articles from online *al-Ahram* newspapers. The use of newspaper articles on such this study is very effective to reveal the language as they report the most valuable events around the world. Apart from that, the news article provides the field of idea sharing, suggestions and opinions that have been delivered through various types of newspaper article starting from current issues section to politics, sports, entertainment, literature, and other (Hishamudin Isam and Norsimah Mat Awal, 2011). To prove the capability of the researcher's tagging in separating the ambiguous words, the analysis results from the Wordsmith software below explain the fact:

Table 4: The example of ambiguous words extracted according to tagging

N	Word	Frequency	Percentage	Text	Percentage
356	أمر_اعدف	1	0.04	1	20.00
357	أمر_فمدم	1	0.04	1	20.00

Table 4 above shows that the word أمر had been used in two variations; firstly, as noun and secondly as perfect verb. Both words use the same spelling but different indication. Therefore, the tagging was used to separate between the same-spelling words to make it available for automatics analysis by Wordsmith software. Without tagging, the computer software will treat all words with similar spelling as one word regardless of their different indication because it operates by recognizing the physical characteristics of the word. The tagging was built specifically for this simple analysis by putting hyphen between the ambiguous words and the tag without spacing.

Table 5: The example of ambiguous words extracted according to tagging

N	Word	Frequency	Percentage	Text	Percentage
164	تحمل_فضدم	2	0.08	1	20.00
431	تحمل_اعدف	1	0.04	1	20.00

Table 5 above shows a similar situation where ambiguous problems have been solved by putting tag between the same-spelling words. The word تحمل had been used in 2 variations in the corpus, firstly as an imperfect verb and secondly as noun. The table also shows that the verb was used more frequently than the noun with 2 and 1 frequency each. The difference in frequency of use affected both ranking in the word frequency list

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

with the verb located in the 164th and the noun down to the 431st in ranking. So, the table proves that both words had been separated accordingly by the software.

Table 6: The example of ambiguous words extracted according to tagging

N	Word	Frequency	Percentage	Text	Percentage
69	توصل_اعدف	4	0.17	1	20.00
98	توصل_فضمم	3	0.13	1	20.00

Table 6 indicates the example of ambiguous words in the word توصل where the first word is used as noun and the second word is imperfect verb. By analyzing the frequency of both words, it can be determined they are quite popular words in the data compared to words in table 4 and 5. This table might be useful to prove how important the tagging process to this kind of analysis in separating between ambiguous words at high frequency. Therefore, the use of this software is also very useful to analyze the large data which contains millions of words and extract them into frequencies precisely if the data had been pre-tagged before the analysis.

Table 7: The example of ambiguous words extracted according to tagging

N	Word	Frequency	Percentage	Text	Percentage
73	خفض_اعدف	4	0.17	1	20.00
516	خفض_فمدم	1	0.04	1	20.00

Table 7 above shows another example of ambiguous words used in the analyzed data. The word خفض as noun was used more frequently than the perfect verb خفض with 4 and 1 frequency each. As other tables before, it proves the availability of this software to sort out the ambiguous words with preliminary work of tagging.

Table 8: The example of ambiguous words extracted according to tagging

N	Word	Frequency	Percentage	Text	Percentage
813	يفتح_فضدج	1	0.04	1	20.00
814	يفتح_فضدم	1	0.04	1	20.00

Table 8 proves similar fact about the role of tagging process in automatics Arabic word analysis. It shows that the software is capable to differentiate between verb classification is long as different tagging were given to each class of word. The example above shows that both words are verb but the first verb was used in passive, and the second verb is active.

6. CONCLUSION

In conclusion, it could be stated that several critical issues appeared in tagging Arabic texts. The existence of Wordsmith software that is capable to differentiate the word characters between spaces requires only minimal tagging. However, it still needs a lot of work in separating between words in one character as explained before. This problem occurs more complicatedly in Arabic compared to in other languages. This problem causes other problems which make the tagging and separating process more complicated as discussed previously. Despite this situation, this manual tagging is only preferable for the small corpus that was compiled for specifics aims like this present study. It is not suitable for a large corpus that was built for the use of other researchers generally with various aims and scopes.

The determination on what extent the tagging process will go on in manual tagging specifically depends on researcher's aim of study. That is because executing on manual tagging in a very detailed Part of Speech (PoS) as the automatics tagging does is time consuming. Therefore, it is advice that the researcher only studies selected PoS and does tagging on that area. As a result, the automatics software analyzer only will come out with the targeted output. The basis of tagging process is the researcher must try to make the ambiguous words physically different by putting variation of tag set to make the analysis software treat them as a different character.

Generally, tagging process for this present corpus involves two steps. The first step is to separate words into one character. The purpose of this process is to enable the analyzer software to compile the frequency of similar words in one entry. The second step is to place the grammatical tagging into the ambiguous words. If this process is not done, the analyzer software would compile all words that are physically similar into one entry regardless of their grammar classification, for example the verb عَقَدَ and the noun عَقْدُ into one entry which is defect for language analyzing.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest

FUNDING STATEMENT

No funding was received for this research

REFERENCES

1) Abdelkader, D.O., Mohamed Yagi, S. & Sawalha, M.S. (2024). Source of syntactic ambiguity of contemporary Arabic: A corpus study. *Jordanian Education Journal*, *9*(3), 27-46. https://doi.org/10.46515/jaes.v9i3.869

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

2) Abdul Razak, Z.R. (2011). Modern Media Arabic: Study of Word Frequency in World Affairs and Sports Sections in Arabic Newspapers. Unpublished Phd Thesis. Birmingham: University of Birmingham.
3) Abdurashetovna, Manzura, A. (2023). Tagging and annotation of corpus units. *International Journal of Language Learning and Applied Linguistics, 2*(12), 103-107.
4) Al-Shamsi, F. and Guessoum, A. (2006). A Hidden Markov Model-Based POS Tagger for Arabic. *8es Journees Internationales d'Analyse Statistique des Donnees Textuelles*, 31-42.
5) Al-Thaani, A. and Abu al-Rub, S. (2009). A Rule-Based Approach for Tagging Non-Vocalized Arabic Words. *The International Arab Journal of Information Technology* 6 (3), 320-328.
6) Awwalu, J., Abdullahi, S.Y & Evwickpaefe, A.E. (2020). Part of speech tagging: A review of techniques. FUDMA Journal of Science, 4(2), 712-721. https://doi.org/10.33003/fjs-2020-0402-325
7) Baker, P, A Hardie and T McEnery. (2006). *A Glossary of Corpus Linguistics.* Edinnburgh: Edinburgh University Press.
8) Boudchiche, M. & Mazroui, A. (2015). Evaluation of the ambiguity caused by the absence of diacritical marks in Arabic texts: Statistical study. Proceeding of the 2015 5th International Conference on Information & Communication Technology and Accessibility (ICTA), Marrakech, Morocco. https://doi.org/10.1109/ICTA.2015.7426904
9) Branco, A. and Silva, J. (n.d.). EtiFAC: A Facilitating Tool for Manual Tagging. Accessed on February 28, 2012, from http://www.di.fc.ul.pt/~ahb/BrancoSilva.
10) Crystal, D. (2003). *A Dictionary of Linguistics and Phonetics*, 5th ed. Oxford: Blackwell.
11) Diab, M., K. Hacioglu and D. Jurafsky. (2004). Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. *Proceedings of HLT-NAACL 2004*, East Stroudsburg, Pennsylvania.
12) Dwivedi, V. (2024). Natural language processing. In *Data Science for Beginners*. EDSOL Informatics.
13) Habash, N. (2010). *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.
14) Habash, N. and O. Rambow. (2005). Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguition in One Fell Swoop. *Proceeding of the 43rd Annual Meeting of the ACL 2005*, Michigan: Association for Computational Linguistics.
15) Isam, H. and Mat Awal, N. (2011). Analisis Berasaskan Korpus Dalam Menstruktur Semula Kedudukan Makna Teras Leksikal Setia. *GEMA Onlinetm Journal of Language Studies*, 11, (1), 143-158.
16) Kubler, S. and Mohamed, E. (2011). Part of Speech Tagging for Arabic. *Cambridge Journal Online.* Accessed on February 28, 2012, from http://journals.cambridge.org/action.
17) McEnery, T. and A. Wilson. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
18) Minerva, S. (2025). *Enhancing glass-box methods for part-of-speech tagging and text classification.* [Doctoral dissertation, , Chalmers University of Technology]. Chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://research.chalmers.se/publication/546192/file/546192_Fu
19) Scott, M. (n.d) Wordsmith tools. https://lexically.net/wordsmith/research/ (Accessed on August 24, 2026.
20) Van Mol, M. (2003). *Variation in Modern Standard Arabic in Radio News Broadcasts: A Synchronic Descriptive Investigation into the Use of Complementary Particles*. Paris: Peeters and Department of Oriental Studies.
21) Yamina, T.G. (2005). Tagging by Combining Rules-Based Method and Memory-Based Learning. *Journal of World Academy of Science, Engineering and Technology* 6, 110-114.

Appendix

Proposed Tag set

DESCRIPTION	TAG
NOUN	**الاسم**
اسم خاص إنسان مفرد	اخاف
اسم خاص إنسان مثنى	اخاث
اسم خاص إنسان جمع	اخاج
اسم خاص مكان مفرد	اخمف
اسم خاص مكان مثنى	اخمث
اسم خاص مكان جمع	اخمج
اسم خاص غير مفرد	اخغف
اسم خاص غير مثنى	اخغث
اسم خاص غير جمع	اخغج
اسم عام صفة مفرد	اعصف
اسم عام صفة مثنى	اعصث
اسم عام صفة جمع	اعصج
اسم عام فاعل مفرد	اعلف
اسم عام فاعل مثنى	اعلث
اسم عام فاعل جمع	اعلج
اسم عام غير مفرد	اعغف
اسم عام غير مثنى	اعغث
اسم عام غير جمع	اعغج

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/

اعظم	اسم عام ظرف مكان
اعظز	اسم عام ظرف زمان
اعمع	اسم عام موصول
اعمث	اسم عام موصول مثنى
اعمج	اسم عام موصول جمع
اعدف	اسم عام مصدر مفرد
اعدث	اسم عام مصدر مثنى
اعدج	اسم عام مصدر جمع
اعتذ	اسم عام تفضيل مذكر
اعتن	اسم عام تفضيل مؤنث

الفعل	**VERB**
فضدم	فعل مضارع مجرد معلوم
فضدج	فعل مضارع مجرد مجهول
فضمع	فعل مضارع مزيد معلوم
فضمج	فعل مضارع مزيد مجهول
فضدث	فعل مضارع مجرد مثنى
فضمث	فعل مضارع مزيد مثنى
فمدم	فعل ماض مجرد معلوم
فمدج	فعل ماض مجرد مجهول
فممم	فعل ماض مزيد معلوم
فممج	فعل ماض مزيد مجهول
فمدث	فعل ماض مجرد مثنى
فممث	فعل ماض مزيد مثنى
فادف	فعل أمر مجرد مفرد
فادج	فعل أمر مجرد جمع
فادث	فعل أمر مجرد مثنى
فامف	فعل أمر مزيد مفرد
فامث	فعل أمر مزيد مثنى
فامج	فعل أمر مزيد جمع

الحرف	**PARTICLE**
حنا	أنّ
حنب	أنْ
حنت	إنّ
حنث	إنْ
حشر	إما

License: This is an open access article under the CC BY 4.0 license: https://creativecommons.org/licenses/by/4.0/