A Proposed Adaptive Scheme for Arabic Part-of Speech Tagging

Mohammad Fasha
Department of Computer Science
King Abdulla II School for Information Technology
The University of Jordan, Amman

***Abstract*—This paper presents an Arabic-compliant part-of-speech (POS) tagging scheme based on using atomic tag markers that are grouped together using brackets. This scheme promotes the speedy production of annotations while preserving the richness of resultant annotations. The proposed scheme is comprised of two main elements, a new tokenization approach and a custom tool that enables the semi-automatic implementation of this scheme. The proposed model can serve in many scenarios where the user is in a need for better Arabic support and more control over the Part-of-Speech tagging process. This scheme was used to annotate sample narratives and it demonstrated capability and adaptability while addressing the various distinguishing features of Arabic language including its unique declension system. It also sets new baselines that are prospect for further exploration by future efforts.**

***Keywords*—Arabic natural language processing (ANLP); part-of-speech (POS) tagging; part-of-speech tokenization scheme; morpho-syntactic tagging; Arabic declension system**

I. INTRODUCTION

Part-of-Speech (POS) tagging is the process of classifying and labeling words in a sentence according to their grammatical categories, i.e., verbs, nouns, particles, … etc. [1]. It is considered as an important step in many Natural Language Processing (NLP) implementations [2] as it deliver a layer of abstraction over the vast variances of the lexical, syntactic and semantic content of natural language. This generalization process renders that vast amount of knowledge into controllable artifacts that are valuable for many related implementations.

In contrast to other languages, Arabic has several distinguishing and challenging features, more importantly, its rich morphology and highly inflectional nature. A single Arabic word can bear more meaning than it's English counterparts [3]. Therefore and more often, information is either lost or misrepresented using the conventional Part-of-Speech tagging schemes. Moreover, there is a noticeable shortage in terms of standards related to Arabic Part-of-Speech tagging schemes, whether for the used tagsets or for the tokenization process [4], [5].

To assist in mitigating some of these challenges, we propose a new Part-of-Speech tagging scheme that can provide rich annotations while being simpler and less demanding than the detailed parsing of corpora, which is cumbersome and time consuming [6]. The scheme we are proposing is based on using tagsets of atomic tag markers that can be aggregated and grouped together using brackets. Having such arrangements, users are provided with fundamental baselines that enable them to seamlessly commence with a rich morpho-syntactic annotation process for Arabic text.

The contributions of this work includes the definition of a declension system (نظام الاعراب) complaint morpho-syntactic tagging scheme that promotes simplicity, clarity and agility of the produced annotations as well as the tagging process itself. Further, to the best of our knowledge, this is one of the rare studies that surveys Arabic Part-of-Speech tagging schemes and discusses their pros and cons. This important subject needs further investigation due to the unique linguistic features of Arabic language, while most related work concentrates on establishing rule-based or statistical motivated Part-of-Speech taggers and morphology analyzers.

This paper is structured as follows. In Section II, we present a brief introduction about the distinguishing features of Arabic language. In Section III, we discuss the related previous work. Section IV presents some of the challenges that are related to the conventional Arabic Part-of-Speech tagging schemes. In Section V, the proposed tagging scheme is presented in more detail. Section VI presents the custom annotation tool. Section VII presents a sample narrative annotated using the proposed scheme and finally in Section VIII we present the conclusion and the suggested future work.

II. ARABIC DISTINGUISHING FEATURES

Arabic is a Semitic language spoken by over 300 million speakers in 22 Arabic countries, it has a liturgical importance as it is the language of Quran, the Holy book for over 1.2 billion Muslims around the world [7].

In contrast to many other languages i.e. Indo-European languages, Arabic has many distinguishing features. These features are related to its rich morphology, highly inflectional nature, subject dropping, free words order, short vowels omission, large lexicon and vocabulary and many others [8], [9]. Accordingly, it is quite often challenging to identify the correct Part-of-Speech of a given word under a certain context.

The rich morphology of Arabic can be related to its template nature where new words are derived from root ones by applying a set of fixed patterns. In addition, Arabic has a concatenate nature where words (nouns and verbs) are inflected to indicate different senses. For example, Arabic

nouns can be inflected to indicate number (singular, dual, plural), gender (masculine, feminine), definiteness (definite, indefinite) and case (nominative, accusative, genitive) as well as possession. Similarly, Arabic verbs are inflected to indicate aspect (perfective, imperfective, imperative), voice (active, passive), tense (past, present, future), mood (indicative, subjunctive, jussive), subject (person, number, gender) as well as object clitics. In addition, Arabic words can be prefixed with functional morphemes (single particles or prepositions) to indicate various senses (causality, conjunctions, assertion, inquiry, association … etc.).

To demonstrate the richness of Arabic language and the amount and diversity of information that can be convoyed in a single word, we consider the surface word (wa sa nokhberu hum, وسنخبرهم, and we shall inform them) as an example. This single word is comprised of the following constituents:

- The proclitic morpheme (wa, و, and) which indicates coordinating conjunction.
- The proclitic morpheme (sa,س , shall) which indicates a future event.
- The inflection particle (nun, ن) which indicates first voice plural speaker (us).
- The stem (khabara,خبر , tell) which is the verb itself.
- Finally, the enclitic morphemes (hum,هم , them) which is an attached pronoun that indicates a plural object of the verb.

In [10], the author provides a more detailed discussion about Arabic morphology and its distinguishing features. Nevertheless, annotating the previous sample word with a verb marker (VB) according to its grammatical category shall waste numerous information. Therefore, a viable Arabic part-of-speech tagging scheme has to possess the capacity to address Arabic distinguishing features and to accurately classify Arabic words without losing information or creating ambiguities. In order to be able to support the distinguishing features of Arabic language, the required part-of-speech tagging scheme has to be able to fully support Arabic's declension system (نظام الاعراب).

In the next section, we present a brief discussion about the related previous work and highlight their main challenges.

III. Related Work

A limited number of part-of-speech Taggers were presented for Arabic language [11]. Generally, these automated taggers can be classified under three main schemes: the statistical-based schemes, the rule-based schemes and the hybrid ones [2]. More importantly, reviewing the previous related work, we noticed an overlapping between part-of-speech tagging and morphology analyses process. For example, Stanford NLP toolkit uses the reduced Penn tagset, while others like the Buckwalter AraMorph incorporates the syntactic category of a given word within the generated morphology analyses results.

Nevertheless, in this work, we are interested in the part-of-speech annotating scheme and format that was implemented by every one of these tools. We start our listing with an early effort that was presented by [12] who introduced a hybrid algorithm for Arabic part-of-speech tagging. That algorithm used a custom tagset comprised of (130) fixed morpho-syntactic markers that were defined based on Arabic grammar rules. Each marker identifies the grammatical category and the inflections of a given word. For example, a perfect verb in the second person masculine plural form is annotated using the (VPPl2M) marker and a singular masculine accusative definite adjective is annotated using the (NACSgMAD) marker.

An interesting tagging scheme was presented in Arabic Treebank (ATB) project [13]. That tagging scheme was based on the well-known rule-based Buckwalter Arabic Morphology Analyzer (BAMA) [14]. (BAMA) uses around (70) basic tag markers that can be combined together to form a larger number of composite tags. For example, in (BAMA), the (IV_PASS) marker indicates imperfective passive verb, three types of information are aggregated together in that composite tag, i.e., imperfect, passive and verb. (BAMA) include tags for indicating person, voice, mood and aspect for verbs, and gender and number for their subjects. It also includes gender, number, case and state for different types of nominals [5].

Another important tagging scheme was introduced by the Prague Arabic Dependency Treebank (PADT) project that was presented in [15]. In that work, a multi-level annotation scheme for a selected corpus was implemented. The first level of annotation involved the morphology analyses of Arabic words. For that part, a morphology compliant tagset was used to construct a (15) slots structure covering the various morphological aspects of a given word i.e. gender, number, person, aspect … etc. In PADT, a single character represented each morphology feature. A challenge in (PADT) tagging scheme was that the meaning of the same character might differ according to a specific internal structuring procedure. For example, the letter (P) on the second position is to be read as Passive Particle if it was preceded by an (N, Noun), and as a Perfect if it was preceded by a (V, Verb). This arrangement requires specialized skills and knowledge to able to use and interpret (PADT) tagging scheme [16].

Similarly, CATiB project [17] presented an Arabic Treebanking scheme that was designed with the motivation of providing rich annotations while being simpler than other similar efforts i.e. ATB and PADT. The focus of CATiB was primarily on the speedy production of the manually annotated corpus while the inspiration was not to duplicate information that could be extracted or indicated by other means, i.e., by syntactic analysis. Consequently, CATiB introduced a succinct (POS) tagset comprised of (6) POS tags which are: NOM for nominals. PROP for proper nouns, VRB for active-voice verbs, and VRB-PASS for passive- voice verbs, PRT for particles and PNX for punctuations. Other markers were identified for the deeper level of syntactic-motivated annotations.

In [18], authors presented a functional based (POS) tagset where words are tokenized and (POS) tagged based on their grammatical functions rather than their morpho-syntactic structure. For example, the sentence (زمانها خلصت المسيرة, the march must have finished) is labeled as a modal (MD)

although the direct (POS) for the Arabic word (زمان, Time) is (NN, Noun).

A relatively recent effort was introduced by [11] who presented a systematic scheme for establishing Arabic compliant tagsets. In that work, a three level categorization of Arabic morpho-syntactic tagsets was defined. The first level was comprised of 7 tags, the inner level included 23 tags while lower level included 54 tags. Accordingly, the user of the system can use the depth of tagging that can better address his needs.

Finally, [2] and [5] presented interesting reviews on Arabic part-of-speech taggers and tagsets where the former concentrated on tagsets while the later presented a listing of the most prominent taggers along with a discussion about their challenges and limitations.

IV. Challenges Related to the Existing Arabic (POS) Schemes

The review process that was presented in the previous section revealed several challenges and limitations that are related to the existing tagging schemes. To further assess these schemes, we examined a number of the accessible taggers and morphology analyzer which included Stanford NLP toolkit [19], NLTK toolkit [20], AL-Khalil morphology analyzer [3], BAMA morphology analyzer [14] as well as MADAMIRA [21] and SAFAR platforms [22]. Table 1 below presents a listing of the results that were captured while examining these tools over a sample sentence. Analyzing the results from a linguistic perspective, we concluded to the following list of observations:

a) There is no standardized or a community adopted (POS) scheme for Arabic language. Our examination revealed that different (POS) tagsets were used by different (POS) taggers; some of these tagsets were generalized while others were more detailed to better address Arabic distinguishing features. The observation was also noted by [4]. Similarly, the tokenization scheme of the tag markers is also different in each tool.

b) The accuracy of the examined (POS) taggers was questionable. For example, Stanford NLP produced numerous errors in the generated tagging such as the noun (كرته, his ball) which was annotated as (NNP) or a proper noun. Similarly, and for a different sample sentence, MADAMIRA identified the word (شعر, felt) as a noun (poetry) rather than a verb, also the verb (حضر, came) was identified as a verb inflected for third person singular masculine while the correct interpretation according to the context was a third person plural masculine. Likewise, the verb (خدعتكم. I deceived you) was identified as a verb inflected with a third person singular feminine subject while it was masculine according to the context. Moreover, the library failed to analyze some words e.g. (مسرعين, in a hurry) which were tagged as NO-ANALYSIS.

TABLE I. Analysis Results of Arabic (POS) Taggers and Morphology Analyzers for A Sample Sentence

Arabic Sentence	وركل	كرته	فاتجهت	نحو الزجاج	فحطمته
English Translation	and he hit/inflected as Singular Masculine Subject Inflection	his ball/ attached possession pronoun	So it went Singular Feminine Subject Inflection	towards the window/glass	so it broke it/inflected as Singular Feminine Subject + Singular Masculine Object
Buckwalter Transliteration	wrkl	krth	fitajahat	naHow AlzujAj	fHTmt
Stanford NLP	وركل/VBD	كرته/NN	فاتجهت/VBD	نحو/NN الزجاج/DTNN	فحطمته /VBD
Al-Khalil	12 solutions, verbs and gerunds	17 solutions, verbs, nouns and gerunds	5 solutions, verbs	15 solutions for both	13 solutions, verbs and gerunds
BAMA	2 solutions including VERB_PERFECT and NOUN	6 solutions, NOUN	6 solutions including VERB_PERFECT With different subject inflections	4 solutions for naHowa and 5 solutions for AlzujAj	9 solutions VERB_PERFECT Different Subject and Object inflections
SAFAR	7 solutions, different subjects inflections	15 solutions, different subjects and objects inflections	4 solutions, different subject inflections	14 solutions for naHowa and 17 solutions for Al zujAj	12 solutions, different subjects and objects inflections
MADAMIRA	wa/CONJ+rakal /PV+a/PVSUFF _SUBJ:3MS	kur/NOUN+ap/NSUFF_FEM_SG +a/CASE_DEF_ACC+hu/POSS_ PRON_3MS	fa/CONJ+{it~ajah/PV+ at/PVSUFF_SUBJ:3FS	naHowa/PREP Al/DET+zujAj/NO UN+a/CASE_DEF _ACC	fa/CONJ+HaT~am/PV+at/PVSU FF_SUBJ:3FS+hu/PVSUFF_DO: 3MS

c) Some of the examined tools were not suitable for automated (POS) processing as they generate all the possible interpretations for a given word. This observation was noticed in BAMA and AL-Khalil morphology analyzers. Moreover, Al-Khalil does not employ any (POS) tokenization scheme,

rather, it generates all its results in plan Arabic text according to Arabic declension system, this features makes it unsuitable for any integration potentials.

d) Some of the investigated tools, i.e., SAFAR were a collection of other tools that were aggregated and compiled under a single platform. These tools were not stand-alone products by themselves and they did not introduce any original add-ins in terms of the Part-of-Speech tagging functionalities.

e) In many situations, words were tagged with an overly generalized version of tag markers where useful information was lost. This can be witnessed in Stanford (POS) tagger that employs the English Penn Treebank tagset for annotating Arabic words. That tagset lacks Arabic morphology features. Similarly, useful information is wasted as the examined tools are not fully compliant with Arabic declension system (نظام الاعراب). For example, gender information proper nouns, some adjectives and nouns were not included. Likewise, functional characters have an important role in Arabic language, yet the functional specificity for some Arabic particles was neglected such as the conditional (اذا, if).

f) The number of basic tag markers and the number of their possible combinations can reach large amounts that can complicate the tagging process. In [23], the authors identified over (2000) markers for Arabic while the combination of these markers can theoretically reach (33000) different tag combination [24].

g) Overlapping and duplications can be witnessed in some of the existing tagging schemes. Such overlapping can complicate string-based matching over the Part-of-Speech strings. For example, in the Penn Treebank tag markers presented below, we notice that the concept of feminine gender is represented using the single character (F), yet this same character appears as part of the (PVSUFF) marker in the same string.

*VERB_**P**A**SS**IVE+**P**V**S**U**FF**_**S**UBJ:3FS*

*VERB_**P**A**SS**IVE+**P**V**S**UFF_**S**UBJ:3**FS***

*VERB_**P**A**SS**IVE+**P**V**S**UFF_**S**UBJ:3M**P***

The same remark can be observed for the singular number marker (S) and the plural (P) as they overlap with characters in the word (PASSIVE).

h) In addition, we can observe that the same concept might be represented using different markers within the same scheme. For example, the tags markers presented below demonstrate how the singular number was represented using (SG) in the first sample and using an (S) in the second.

*ADJ+NSUFF_**FEM_SG***

*IV3**FS**+VERB_IMPERFECT*

The same is true for the feminine gender markers i.e. the (F) and (FEM). Such inconsistency can create confusion during the use of the markers and weakens the scheme's standardization potentials.

i) For generating morpho-syntactic tagging, it is required that we perform a full tokenization for sentences prior to the tagging process. Such requirement might be cumbersome and time consuming and it should be useful if we can develop a simpler scheme that can replace the explicit tokenization with an implicit one as the missing information can be recovered using algorithmic measures.

j) Considering the previously discussed challenges and limitations, manual intervention is often required to fine-tune the automatically generated annotations. This intervention is required to verify and/or extend the generated annotations and to validate their accuracy and adequacy for further stages of processing, which brings us to another challenge in this respect and that is the scarcity of available and accessible annotation tools that can enable and facilitate such functions of manual intervention.

In the next section, we present our proposed (POS) tagging scheme which might assist in addressing some of the aforementioned challenges as well as setting new perspectives for further exploration in future.

V. The Proposed Tagging Scheme

In this section, we present the proposed part-of-speech tagging scheme including its objectives, design principles, the initial tagset, the tagging process as well as the custom tool that was prepared to enable this scheme.

A. Objectives and Design Principles of the Proposed Scheme

The main objective of the proposed tagging scheme was to provide users with initial baselines that enable them to implement a rich morpho-syntactic declension-system compliant annotation for Arabic words in a clear, simple and agile manner. Using this scheme, users can experiment with different tag markers that are more compliant with Arabic language, and would be able to examine their influence on different Natural Language Processing (NLP) applications e.g. Information Extraction, Text Translation, Text Summarization … etc.

The clarity, simplicity and agility of the proposed scheme were established by allowing users to commence with the annotation process without the need for the explicit tokenization of words. Rather, the tokenization is achieved using different brackets as shall be presented later. The inspiration for this arrangement was motivated by the tagging scheme that was presented in [17]. In that work, the speedy production of annotations was enabled by eliminating the annotation of information that could be extracted by other means. For instance, case markers for nominals could be identified from syntax, therefore, the Part-of-Speech annotation scheme presented in [17] did not include such markers in its tagset.

The morpho-syntactic richness of the annotations is enabled by the support of different categories of tag markers that are compliant with Arabic declension system, this includes lexical categories of words; morphology related markers, functional markers as well as declension system specific ones.

To enable the aforementioned objectives, the proposed scheme was based on the following design principles:

a) All the defined tag markers in the scheme were standalone and atomic. Each marker is self-explaining and self-contained and clearly defines a single concept e.g. gender, number, case, mood…etc. This design principle promotes the clarity of markers and ensures that no duplication or overlapping between markers can occur. For example, if a marker indicates a certain concept e.g. FEM for feminine gender, this same marker will be used for all words categories that might be inflected to indicate gender i.e. nouns, verbs, adjectives, pronouns, relative pronouns…etc. No other marker will be used for the same concept regardless of the word category. Therefore, the challenges that were stated in items g)) and h)) of section IV cannot occur.

b) Composite markers are established as aggregates of the basic and atomic ones. For instance, a plural noun is represented using the (NN) marker and the (PLR) marker, not with a single marker i.e. (NNS), for both concepts. This design principle preserves clarity and allows extendibility using clear composition of markers; it also facilitates string-based matching operations that can be implemented over part-of-speech annotations.

B. Initial (POS) Tagset

The definition of a coherent Arabic-compliant tagset is out of the scope of our current work. In [11] and [25], the authors provided interesting guidelines that can assist in defining an Arabic-compliant tagset in a more systematic manner.

Nevertheless, for assessing our proposed model, we established an initial tagset to demonstrate the capability of the scheme and the diversity of markers that it can seamlessly support. This initial tagset (presented in Appendix A) classify the tag markers according to the following categories:

- Lexical markers:

This category includes the basic grammatical classification of words according to Arabic language rules. This includes the classification of nominals, verbs and particles, the three main Arabic word types along with their direct subsets.

- Morphology related markers:

This includes the markers that identify affixations and inflections related to nouns and verbs.

- Functional markers:

Functional markers include the tags that indicate the functional role of a given lexical entity. This includes senses of causality, modality, time and space relations, assertion, confirmation, negation, sequencing and conjunction coordination as well as others.

- Arabic declension system:

This category includes markers that are related to case definitions for Arabic nouns and mood definitions for Arabic verbs, as well as other features that signals specific insights that are related to Arabic language e.g. (Kana and its sisters, كان وأخواتها).

C. The Proposed Tokenization Scheme

A main objective of the proposed model was to better support Arabic declension system i.e. (نظام الاعراب) where the user is able to employ adequate combination of markers that can better satisfy his needs and his language proficiency.

Having an extended and diverse tag set, it was important to define an adaptive, dynamic and flexible tokenization scheme that can utilize these diverse markers in a simple, clear and agile manner.

Two types of brackets were employed to establish the proposed tokenization scheme, the round brackets or parenthesis "( )" and the braces or the curly brackets "{ }". Using these brackets, different levels of grouping and hierarchies could be established to annotate different word categories. The parentheses are used to establish word level groupings while the curly brackets are used to create token level annotations. This arrangement combines concepts from conventional Part-of-Speech tagging, morphology analysis as well as syntactic tree parsing as a single Arabic word can encompass a multi-token paragraph according to its morphology.

To demonstrate the proposed bracketing scheme, we consider the sample surface word that was presented in Section 2 (wa sa nokhberu hum, وسنخبرهم, and we shall inform them). Using the proposed scheme, this single word is annotated as following:

- {RP+WA+CC}: The proclitic morpheme (wa, و, and) which indicates coordinating conjunction particle.
- {RP+SA+FTR}: The proclitic morpheme (sa,س , shall) which indicates a future event particle.
- {PLRL+stV}: The inflection particle (nun, ن) which indicates first voice plural speaker (us).
- {VB}: The stem (khabara,خبر , tell) which is the verb itself.
- {PRN+SFX_OBJ+PLRL+MSC}: The enclitic morphemes (hum,هم , them) which is an attached pronoun that indicates a plural masculine object.

While the composite tag for this word is defined as following:

({RP+WA+CC}{RP+SA+FTR}{VB+PLRL+stV}{PRN+SFX+OBJ+PLRL+MSC})

D. Advantages of the Proposed Scheme

To demonstrate the advantages of the proposed tagging scheme over other available schemes, we performed several examinations for annotation sample words using Stanford (POS) tagger, MADAMIRA morphology analyzer and the proposed scheme.

TABLE II. COMPARING THE PROPOSED SCHEME AGAINST OTHER SCHEMES

	Annotation Scenario	Sentence Sample	Stanford (POS) Tagger	MADAMIRA Morphology Analyzer Scheme	Proposed Scheme
1	Composite words	**شاهدته** She saw him	VBD	{bw:$Ahad/PV+tu/PVSUFF_SUBJ:1S+hu/PVSUFF_DO:3MS}	({VBD+SNG+FEM+rdV}{SFX+OBJ+PRN+SNG+MSC})
2	Kana and its sister كان واخواتها	السماء ماطرة **كانت** The sky was raining	VBD	{bw:kAn/PV+at/PVSUFF_SUBJ:3FS}	({VBD+KANA+SNG+FEM+rdV})
3	ENNA and its sister إنَّ وأخواتها	تمطر بغزارة **انها** It is raining heavily	VBP	{bw:<in~a/FUNC_WORD+hA/PRON_3FS}	({IN+ENNA}{PRN+SNG+FEM+rdV})
4	إنْ: gloss: if/whether	**ان** تدرس تنجح If you study you succeed	IN	{bw:<in/FUNC_WORD}	({IN+CND})
5	Active Participle اسم الفاعل	ذاهبه **هي** I am going	JJ	{bw:*Ahib/ADJ+ap/NSUFF_FEM_SG}	{AP+SNG+FEM}
6	Passive Participle اسم المفعول	هو **مظلوم** He is oppressed	NNP	{bw:maZoluwm/ADJ}	{PP+SNG+MSC}
7	Relative Pronouns الاسماء الموصوله	الطفل **الذي** يبكي The baby that is crying	WP	{bw:Al~a*iy/REL_PRON}	({RPRN+SNG+MSC})
8	Demonstrative Pronouns ضمائر الاشارة	**هذا** كتابي This is my book **ذلك** كتابي That is my book	DT DT	{bw:h`*A/DEM_PRON_MS} {bw:h`*A/DEM_PRON_MS}	({PRN+SNG+MSC+NR}) ({PRN+SNG+MSC+FR})
9	Pronouns	**هي** تلعب بالكرة She is playing with the ball **انت** تلعبين بالكرة You are playing with the ball	PRN PRN	{bw:hiya/PRON_3FS} {bw:hiya/PRON_2FS}	({PRN+SNG+FEM+rdV} ({PRN+SNG+FEM+ndV}
10	Distinguish Prepositions	ذهبنا **الى** المدرسة We went to school جلسنا **على** المقعد We sat on the chair	IN IN	{bw:<ilaY/PREP} {bw:EalaY/PREP}	{RP+ELA} {RP+ALA}
11	Gender and Number Markers for Nouns	شاهدت **السماء**	DTNN	{bw:Al/DET+samA'/NOUN+u/CASE_DEF_NOM}	{DT+NN+SNG+FEM+CSN}
12	Adverbs of manner	ركض الولد **سريعا** The boy ran quickly	JJ	{bw:sariyE/ADV+AF/CASE_INDEF_ACC}	({RB+MNR})
13	Interrogative Nouns	كم How much متى When كيف How أين Where لمن Whose	WRB WRB WRB WRB WP$	{bw:kam/INTERROG_PART} {bw:mataY/INTERROG_PART} {bw:kayofa/INTERROG_PART} {bw:>ayona/INTERROG_PART} {bw:li/PREP+man/INTERROG_PART}	({WP+QTY}) ({WP+TIM}) ({WP+MNR}) ({WP+LOC}) ({WP+POSS})

Stanford tagger produces basic syntactic based tag markers for Arabic, while MADAMIRA provides a more extended version of markers that includes syntactic word classifications as well as the morphology analysis related ones. Table 2 below presents a listing of the gathered results.

As demonstrated in the table, the proposed scheme can deliver the same set of capabilities that are provided by the other models only it has the following additional advantages:

- The format of the proposed tagging scheme falls between the briefed Stanford format and the extended format of MADAMIRA. Nevertheless, the proposed scheme provides all the information that is delivered by those two schemes in a simplified manner that includes the syntactic word type classification as well as the morphology related ones.
- The use of brackets eliminates and substitutes the explicit tokenization of composite words. As demonstrated in the first sample, that composite word is comprised of two parts, the perfect verb and the attached pronoun. Curly brackets surround each of these two word parts and parenthesis surrounds the whole string. While in the other schemes, the aggregation is achieved by attaching characters together without any separators or using separators such as the underscore marker "_", the plus sign "+", the colons ":", as well as other approaches e.g. PV+PVSUFF_SUBJ:3MS.

- The proposed scheme does not use single-character markers as they can create ambiguities and overlaps. Rather, multi-character atomic tag markers are used to establish a self-explaining set of annotations.
- Also, unlike [12], [14], [16], [17], no aggregate markers are used in the proposed scheme, rather, all aggregations are established using the plus sign "+" character which is inserted between the atomic markers. Reference [26] presents an interesting listing for tokenization alternatives that are used by a number of different schemes. While in the previous efforts, different approaches where employed to achieve the same objective where a combination of the tokenization process, part-of-speech tagging and morphology analysis are all combined causing overlapping and ambiguity.
- Finally, the proposed scheme enables the introduction of different categories and types of tagsets and tag markers, whether they are related to basic syntactic and grammatical markers, functional markers, morphology related and semantic markers or any other type that might be needed for a specific objective. The expendability while maintaining clarity and simplicity is a powerful feature that maximizes the benefits of the proposed scheme. This can be observed in many samples in the previous table where explicit markers are used for different Arabic linguistic features e.g. active participle, passive participle, KANA and its sisters … etc. Using such explicit markers can facilitate later efforts such as information extraction since these explicit markers can signal the existence of specific types of information.

VI. The Custom Annotation Tool

To enable the proposed scheme, a Java based custom tool was prepared. We refer to this custom tool as the Bracket Based Arabic Annotation (B2A2) tool as it employs brackets to establish morpho-syntactic compliant part-of-speech annotations for Arabic language.

Fig. 1 below presents a screenshot of the (B2A2) tool that demonstrates the tagging hierarchies (left) and the available tag markers (right). To commence with a new tagging process, a newline-terminated text file is uploaded into this tool where it will be initially bootstrap annotated using Stanford (POS) tagger. Later, the user uses the custom tool to review the initial annotations and modify/extend them accordingly. As demonstrated in the figure, the tool is delivered with an initial tagset where markers are classified into a number of categories e.g. base or lexical tags, functional tags, Arabic specific … etc. These tags and tagsets can be easily modified and configured by the user who can introduce new tagsets or tag markers or modify the existing ones according to his needs. The modification for these markers can be introduced into the designated (tag_def) database table i.e. SQL Server database. The structure of the tag definition table is described in Table 3 next. The user can modify the markers themselves as well as their categorization. The custom tool dynamically incorporates any modifications on the markers or their categories during its initialization process. This dynamicity in marker definition as well as their utilization by the user allows users to use different formats for annotating the same word.

The variance in annotations is related to the defined tag markers, the required depth of coverage and richness of the annotation process as well as the user's linguistic proficiency.

Fig. 2 below demonstrates a screenshot of the (B2A2) tool, which clarifies how different annotations can be implemented for the same word according to the user's defined annotation guidelines.

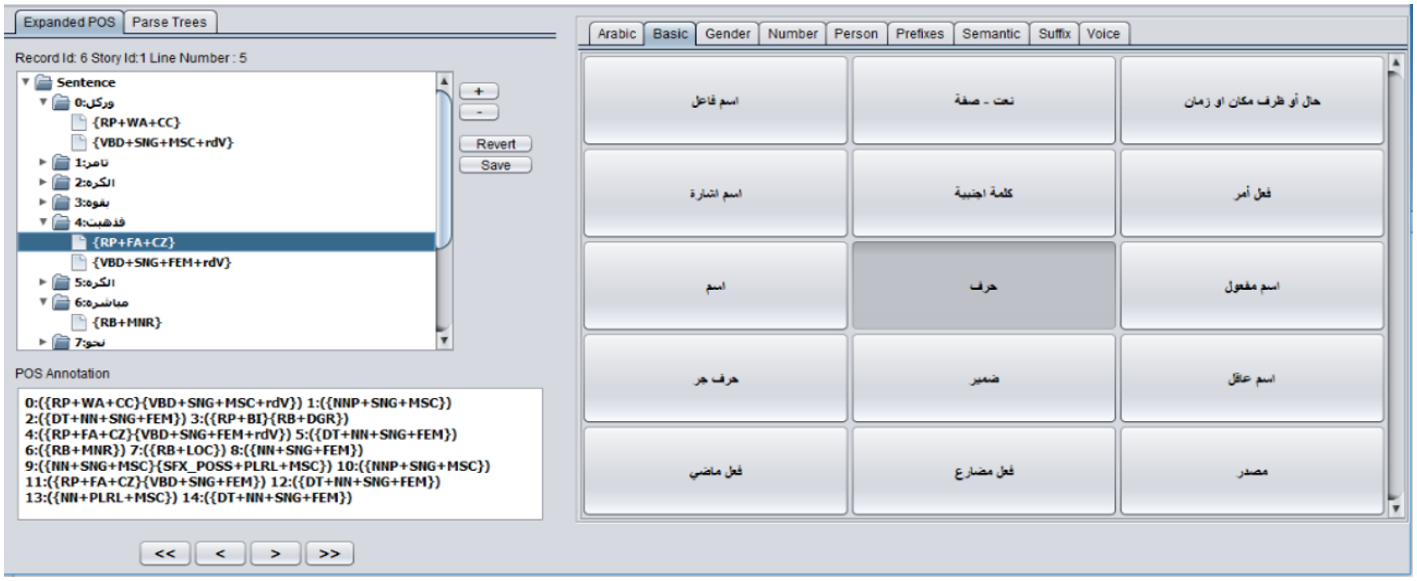

Fig. 1. The custom annotation tool.

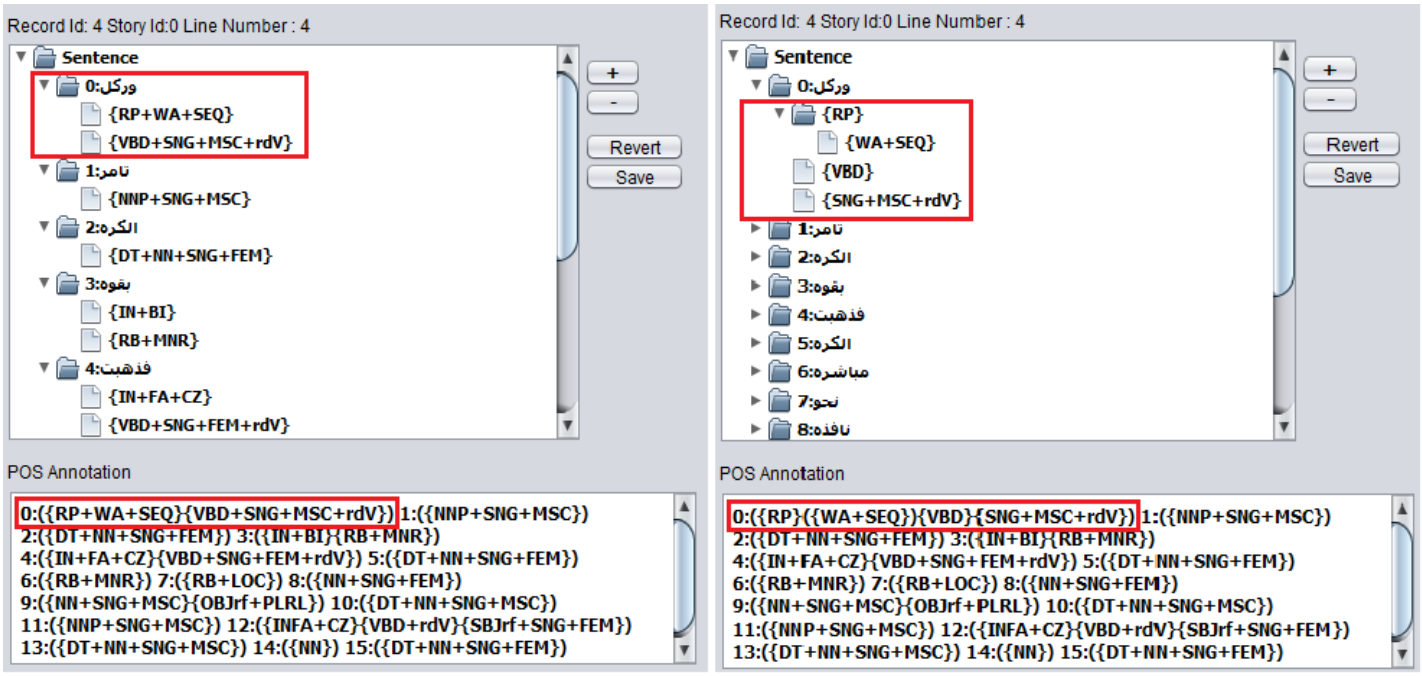

Fig. 2. Words and their constituents can be annotated different according to the user's definitions and requirements.

TABLE III. TAGS DEFINITION TABLE

Column Id	Explanation	Example
id	Unique identifier for the tag	
tag_order	The order or the precedence of the tag within a composite annotation	10
tag	The tag marker – acronym	RB
short_name	The short name for the marker	Adverb
english_description	English description for the marker	An adverb is a word that changes or qualifies the meaning of a verb.
arabic_description	Arabic description for the marker	حال أو ظرف مكان او زمان
tag_category	The category where this marker belongs i.e. appears in the custom tool	Lexical Markers

VII. ANNOTATING A SAMPLE NARRATIVE

To assess the proposed scheme in action, we used the (B2A2) tool to annotate a sample narrative comprised of a few sentences. As discussed in the previous section, (B2A2) scheme provides different alternatives for annotating text in terms of the tag markers that can be used as well as their arrangement and grouping using brackets. In this respect, the following guidelines were defined and enforced during the annotation process:

- Verbs annotations were extended with number, gender and person markers.
- Verb prefixes were distinguished using custom particles tagging i.e. (واو WAW, فاء FA, سين SA…etc.).
- Nominals tagging was extended using number and gender markers.
- Noun and adjectives prefixes were distinguished using custom particles tagging i.e. (واو WAW, كاف KA,باء BI … etc.).
- A more precise tag set was used to annotate propositions i.e. (في FEE, من MIN, الى ELA … etc.).
- Prefix particles, propositions and affixes where separated and grouped using dedicated brackets.
- Arabic (KANA, كان واخواتها, Was) was annotated using a custom tag (VBD+KANA) so that it can be better identified for future purposes.
- Occasionally during the annotation process, Arabic declension system was used in order to determine the correct grammatical analyses of some words and phrases so that ambiguous interpretations are resolved.

TABLE IV. A SAMPLE STORY ANNOTATED USING THE CUSTOM SCHEME

Line #	Text and (POS) Annotation
1	اهدت:0 ليلى:1 شقيقها:2 تامر:3 كره:4 جديده،:5
	0:({VBD+SNG+FEM+rdV}) 1:({NNP+SNG+FEM}) 2:({NN+SNG+MSC}{POSS+SNG+FEM}) 3:({NNP+SNG+ MSC}) 3:({NN+SNG+ FEM}) 5:({JJ+SNG+FEM})
2	وكان:0 تامر:1 سعيدا:2 جدا:3 بها.:4
	0:({WA+RP+CC}{KANA+VBD+SNG+MSC+rdV}) 1:({NNP+SNG+MSC}) 2:({JJ+SNG+MSC}) 3:({RB+MNR}) 4:({IN+BI}{SFX_SUBJ+SNG+FEM})
3	ركض:0 تامر:1 نحو:2 الحديقه:3 ليلعب:4 بكرته:5 الجديده.:6
	0:({VBD+SNG+MSC+rdV}) 1:({NNP+SNG+MSC}) 2:({RB+LOC}) 3:({DT+NN+SNG+FEM}) 4:({IN+LI+CZ}{VBP+SNG+MSC+rdV}) 5:({IN+BI}{NN+SNG+FEM}{POSS+SNG+MSC}) 6:({DT+JJ+SNG+FEM})
4	وركل:0 كرته:1 بقوه:2 فذهبت:3 الكره:4 مباشره:5 نحو:6 نافذه:7 جارهم:8 السيد:9 عادل:10 فتحطم:11 زجاج:12 النافذه.:13
	0:({RP+WA+CC}{VBD+SNG+MSC+rdV}) 1:({DT+NN+SNG+FEM}{POSS+SNG+MSC}) 2:({IN+BI}{RB+MNR}) 3:({IN+FA+CZ}{VBD+SNG+FEM+rdV}) 4:({DT+NN+SNG+FEM}) 5:({RB+MNR}) 6:({RB+LOC}) 7:({NN+SNG+FEM}) 8:({NN+SNG+MSC}{POSS+PLRL+MSC}) 9:({DT+NN+SNG+MSC}) 10:({NNP+SNG+MSC}) 11:({IN+FA+CZ}{VBD+SNG+FEM+rdV}) 12:({DT+NN+SNG+FEM}) 13:({NN}) 14:({DT+NN+SNG+FEM})
5	راى:0 تامر:1 السيد:2 عادل:3 قادما:4 الى:5 منزله:6
	0:({VBD+SNG+MSC+rdV}) 1:({NNP+SNG+MSC+rdV}) 2:({DT+NN+SNG+MSC}) 3:({NNP+SNG+MSC}) 4:({AP+SNG+MSC}) 5:({IN+ELA}) 6:({NN+SNG+MSC}{POSS+SNG+MSC})
6	فخاف:0 واختبا:1 خلف:2 شجره.:3
	0:({IN+FA+CZ}{VBD+SNG+MSC+rdV}) 1:({RP+WA+CC}{VBD+SNG+MSC+rdV}) 2:({RB+LOC}) 3:({NN+SNG+FEM})
7	راى:0 السيد:1 عادل:2 تامر:3 مختبئا:4 وابتسم:5 وقال:6 لا:7 تختبىء:8 يا:9 صغيري،:10
	0:({VBD+SNG+MSC+rdV}) 1:({DT+NN+SNG+MSC}) 2:({NNP+SNG+MSC}) 3:({NNP+SNG+MSC}) 4:({AP+SNG+MSC}) 5:({RP+WA+CC}{VBD+SNG+MSC+rdV}) 6:({RP+WA+CC}{VBD+SNG+MSC+rdV}) 7:({RP+DMND+NEG}) 8:({VBP+SNG+MSC+ndV}) 9:({RP+YAA}) 10:({JJ+SNG+MSC}{POSS+SNG+MSC})
8	اخبرني:0 الحقيقه:1 ولا:2 تخف.:3
	0:({VMP+SNG+MSC}{PRN+SNG+MSC}) 1:({DT+NN+SNG+FEM}) 2:({RP+WA+CC}{RP+DMND+NEG}) 3:({VBP+SNG+MSC+ndV})
9	خرج:0 تامر:1 من:2 مخبأه:3 واخبر:4 السيد:5 عادل:6 الحقيقه:7 وهي:8 ان:9 كرته:10 تسببت:11 في:12 تحطيم:13 زجاج:14 النافذه.:15
	0:({VBD+SNG+MSC+rdV}) 1:({NNP+SNG+MSC}) 2:({IN+MEN}) 3:({NN+SNG+MSC}{POSS+SNG+MSC}) 4:({RP+WA+SEQ}{VBD+SNG+MSC}) 5:({DT+NN+SNG+MSC}) 6:({NNP+SNG+MSC}) 7:({DT+NN+SNG+FEM}) 8:({RP+WA+CC}{PRN+SNG+FEM+rdV}) 9:({RP+ASRT}) 10:({NN+SNG+FEM}{POSS+SNG+MSC}) 11:({VBD+SNG+FEM+rdV}) 12:({NN}) 13:({DT+NN+SNG+FEM})
10	وتاسف:0 ووعده:1 الا:2 يكرر:3 هذا:4 الفعل:5 مره:6 اخرى.:7
	0:({RP+WA+CC}{VBD+SNG+MSC}) 1:({NNP+SNG+MSC}) 2:({RP+WA+CC}{VBD+SNG+MSC+rdV}{SFX_OBJ+SNG+MSC}) 3:({RP+CNF}{RP+NEG}) 4:({VBP+SNG+MSC+rdV}) 5:({RPRN+SNG+MSC}) 6:({NN+SNG+FEM}) 7:({NN})
11	تبسم:0 السيد:1 عادل:2 وقال:3 يا:4 بني:5 لقد:6 احسنت:7 صنعا:8 بقولك:9 الحقيقه:10 وتاسفك،:11
	0:({VBP+SNG+MSC+rdV}) 1:({DT+NN+SNG+MSC}) 2:({NNP+SNG+MSC}) 3:({RP+WA+CC}{VBD+SNG+MSC+rdV}) 4:({RP+YAA}) 5:({NN+SNG+MSC}{POSS+SNG+MSC}) 6:({RP+CNFRM}) 7:({VBD+SNG+MSC+ndV}) 8:({VN}) 9:({IN+BI}{VBG+SNG+MSC+ndV}) 10:({DT+NN+SNG+FEM}) 11:({RP+WA+CC}{VN+SNG+MSC}{POSS+SNG+MSC})
12	والان:0 خذ:1 الكره:2 وارجو:3 ان:4 لا:5 يتكرر:6 مثل:7 هذا:8 الفعل:9 مجددا.:10
	0:({RP+WA+CC}{RB+TIM}) 1:({VMP+SNG+MSC+ndV}) 2:({DT+NN+SNG+FEM}) 3:({RP+WA+CC}{VBP+SNG+stV}) 4:({IN+CNFRM}) 5:({RP+NEG}) 6:({VBP+SNG+MSC}) 7:({NN}) 8:({DT+SNG+MSC+NR}) 9:({DT+VN+SNG+MSC}) 10:({JJ})

The result of annotation the sample narrative is presented in Table 4 above. For example, the noun (شقيقها, shaqequha, her brother) was annotated using two segments, the first one belongs to the noun part along with its inflection, and the second is related to the attached pronoun suffix. The first part is annotated using {NN+SNG+MSC} tag group while the second part is annotated using the {POSS+SNG+FEM} tag group. As presented, each part is identified using a pair of curly brackets while the whole word (multi-token word) is grouped using a pair of parenthesis.

The annotation process demonstrated the efficiency of the proposed tagging scheme in representing the required syntactic and morphological information in simple yet rich manner. Further, the (B2A2) tool provided an enabling framework that accelerated the process of revising the automatically generated Part-of-Speech tagging and facilitated extending it using the proposed tagging scheme.

The proposed framework (the proposed Part-of-Speech tagging scheme and the B2A2 tool) can serve in numerous scenarios where the user is in a need to annotate a given corpus using a rich morpho-syntactic annotation while that labeled corpus can be used later for different Natural Language Processing (NLP) implementations e.g. Information Extraction from text.

VIII. Conclusion and Future Work

This paper presented a proposed scheme for Arabic-compliant part-of-speech tagging (POST).

Acknowledging the complexity and the richness of Arabic language, along with the shortages in the related standardizations, efforts and resources, the proposed (POST) scheme presented new perspectives that might assist in enhancing Arabic-based part-of-speech tagging process as well as opening doors for new perspectives and insights to regular such efforts.

The theme of the proposed model is relatively simple and straightforward yet powerful and capable in representing different types of information specific to Arabic language and its declension system. This scheme is based on: 1) using well-defined atomic part-of-speech markers; and 2) grouping these markers using two types of brackets, the curly brackets for sub-word level and the parenthesis for the word level of groupings.

A custom tool that is bootstrapped using Stanford (POS) tagger enabled the initial version of the proposed (POST) scheme. This tool is freely available online and it can assist users to commence with a rich Part-of-Speech tagging process in a controllable and seamless manner.

The next work we intend to implement is to examine the benefits that can be achieved by using the proposed scheme in information extraction implementations. In addition, we intend to investigate the bootstrapping of the enabling tool using a morphology aware part-of-speech tagging library, e.g., MADAMIRA.

References

[1] S. Alqrainy, “A Morphological - Syntactical Analysis,” 2008.

[2] R. A. Abumalloh, H. M. Al-Sarhan, O. Bin Ibrahim, and W. Abu-Ulbeh, “Arabic Part-of-Speech Tagging,” J. Soft Comput. Decis. Support Syst., vol. 3, no. 2, pp. 45–52, 2016.

[3] A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane, M. Ould Abdallahi Ould Bebah, and M. Shoul, “Alkhalil Morpho SYS1: A Morphosyntactic Analysis System for Arabic Texts,” Int. Arab Conf. Inf. Technol., pp. 1–6, 2010.

[4] S. Alqrainy, A. Ayesh, and H. Almuaidi, “Automated Tagging System And Tagset Design For Arabic Text,” vol. 1, no. 2, pp. 55–62, 2010.

[5] A. M. S. Alosaimy and E. S. Atwell, “A review of morphosyntactic analysers and tag-sets for Arabic corpus linguistics,” Corpus Linguist. 2015, pp. 16–19, 2015.

[6] N. Habash, R. Faraj, and R. Roth, “Syntactic Annotation in the Columbia Arabic Treebank,” Proc. MEDAR Int. Conf. Arab. Lang. Resour. Tools, Cairo, Egypt, pp. 125–132, 2009.

[7] F. Al-shargi and O. Rambow, “DIWAN : A Dialectal Word Annotation Tool for Arabic,” pp. 49–58, 2015.

[8] M. Attia, “An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks,” Challenges Arab. NLP/MT Conf. …, pp. 48–67, 2006.

[9] M. Sawalha, E. Atwell, and M. a. M. Abushariah, “SALMA Standard Arabic Language Morphological Analysis,” Proceedings ICCSPA International Conference on Communications, Signal Processing, and their Applications. Sharjah, UAE, 2013. [Online]. Available: http://www.comp.leeds.ac.uk/eric/sawalha13iccspa.pdf. [Accessed: 23-Dec-2015].

[10] N. Y. Habash, “Introduction to Arabic Natural Language Processing,” Synth. Lect. Hum. Lang. Technol., vol. 3, no. 1, pp. 1–187, 2010.

[11] Y. O. M. Elhadj, A. Abdelali, R. Bouziane, and A. H. Ammar, “Revisiting Arabic Part of Speech Tagsets,” Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA, vol. 2014, pp. 793–802, 2014.

[12] S. Khoja, “APT : Arabic Part-Of-speech Tagger,” Proc. Student Work. NAACL, pp. 20--25, 2001.

[13] M. Maamouri and A. Bies, “Developing an Arabic treebank: methods, guidelines, procedures, and tools,” Proc. Work. Comput. Approaches to Arab. Script-based Lang., pp. 2–9, 2004.

[14] T. Buckwalter, “Arabic Morphological Analyser Version 1.0,” Linguist. Data Consort. numéro LDC2002L49, 2002.

[15] J. Hajič, O. Smrž, P. Zemánek, J. Šnaidauf, and E. Beška, “Prague Arabic Dependency Treebank Development in Data and Tools,” Proc. {{}NEMLAR{}} Int. Conf. {{}A{}}rabic Lang. Resour. Tools, pp. 110–117, 2004.

[16] O. Smrž, J. Šnaidauf, and P. Zemánek, “Prague Dependency Treebank for Arabic: Multi-Level Annotation for Arabic Corpus,” Adv. Math. (N. Y)., vol. 217, no. 6, pp. 2401–2442, 2008.

[17] N. Habash and R. M. Roth, “CATiB: the Columbia Arabic Treebank,” ’09 Proc. ACL-IJCNLP 2009 Conf., no. August, pp. 221–224, 2009.

[18] R. Al-sabbagh and R. Girju, “Supervised POS Tagger for Written Arabic Social Networking Corpora,” Proc. KONVENS 2012 (Main track oral Present. Vienna, Sept. 19, vol. 2012, pp. 39–52, 2012.

[19] C. D. Manning, J. Bauer, J. Finkel, S. J. Bethard, M. Surdeanu, and D. McClosky, “The Stanford CoreNLP Natural Language Processing Toolkit,” Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. Syst. Demonstr., pp. 55–60, 2014.

[20] S. Bird, S. Bird, and E. Loper, “NLTK : The natural language toolkit NLTK : The Natural Language Toolkit,” Proc. ACL-02 Work. Eff. tools Methodol. Teach. Nat. Lang. Process. Comput. Linguist. 1, no. March, pp. 63–70, 2016.

[21] A. Pasha, M. Al-badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. M. Roth, “MADAMIRA : A Fast , Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic,” Proc. 9th Lang. Resour. Eval. Conf., pp. 1094–1101, 2014.

[22] Y. Souteh and K. Bouzoubaa, “SAFAR platform and its morphological layer SAFAR platform and its morphological layer,” no. December 2011, 2015.

[23] A. Bies and M. Maamouri, “Penn Arabic Treebank Guidelines,” Draft January, vol. 28, no. December, p. 2003, 2003.

[24] N. Habash, O. Rambow, and R. Roth, “MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization,” Proc. Second Int. Conf. Arab. Lang. Resour. Tools, no. November 2015, pp. 102–109, 2009.

[25] G. Leech and A. Wilson, “EAGLES recommendations for the morphosyntactic annotation of corpora,” Version of March, 1996.

[26] N. Habash, O. Rambow, and R. Roth, “MADA+TOKAN Manual,” 2010.

Appendix A. Initial Part-of-Speech Tagset

Lexical Markers			**Functional Markers – Semantic Driven**		
NN	Noun	اسم	LOC	Location	دلالة مكانية
JJ	Adjective	نعت - صفة	TIM	Time	دلالة زمانية
RB	Adverb	حال أو ظرف مكان او زمان	CZ	Cause	دلالة سبب

RP	Particle	حرف	EFCT	Effect	دلالة النتيجة
IN	Preposition	حرف جر	SEQ	Sequence	دلالة تتابع
PRN	Pronoun	ضمير	BGN	Begin of	دلالة بداية وقت
DT	Determiner	اسم اشارة	END	End of	دلالة نهاية وقت
VBP	Verb Present	فعل مضارع	CND	Condition	دلالة شرط
VBD	Verb Past	فعل ماضي	CNF	Confirmation	دلالة توكيد
NNP	Proper Name	اسم عاقل	ASRT	Assertion	دلالة اخبار
FW	Foreign Word	كلمة اجنبية	CC	Conjunction	دلالة عطف
VN	Verbal Noun	مصدر	INTR	Interrogative	دلالة استفهام
PP	Passive Participle	اسم مفعول	QTY	Quantity	دلالة كميات
AP	Active Participle	اسم فاعل	NEG	Negation	دلالة نفي
VMP	Imperative	فعل أمر	EXP	Explanation	دلالة تفسير
RPRN	Relative Pronoun	اسم موصول	DMN	Demand	دلالة طلب
WP	Wh-pronoun	اسم استفهام	PRD	Predicate	خبر
Number Markers			PRD	Predicate	WHY
SNG	Single	مفرد	WHN	When	استفسار متى
DUAL	Dual	مثنى	HOW	How	استفسار كيف
PLRL	Plural	جمع	WHO	Who	استفسار من
			SWR	Swearing	دلالة قسم
Gender Markers			SWR	Swearing	MNR
MSC	Masculine	مذكر	DGR	Degree	درجة الفعل
FEM	Feminine	مؤنث	NR	Near	دلالة القرب
			FR	Far	دلالة البعد
Voice Markers			**Arabic Declension System Specifics**		
stV	First Voice	First Voice Verb	YAA	YAA	يا النداء
ndV	Second Voice	Second Voice Indicator	KANA	KANA	كان واخواتها
rdV	Third Voice	Third Voice Indicator	INNA	INNA	ان واخواتها
Active-Passive Markers			ZRFL	Locative Adverb	ظرف مكان
PSV	Absent Person	صيغة الغائب	ZRFZ	Temporal Adverb	ظرف زمان
ATV	Present Person	صيغة الحاضر	CSA	Accusative Case Ending	علامة النصب
Suffix Markers			CSN	Nominative Case Ending	علامة الرفع
SFX	Attached Pronoun	ضمير متصل	CSG	Genitive Case Ending	علامة الكسر
POSS	Possession	مؤشر على الملكية	CSNU	Nunation Case	علامة التنوين
OBJ	Object Reference	مؤشر على المفعول به	AAN	AAN	عن
Prefix Markers – Functional Particles			ALA	ALA	على
BI	BI	باء	FEE	FEE	في
LI	LI	لام	MEN	MEN	من
FA	FA	فاء	HATTA	HATTA	حتى
SA	SA	سين	ELA	ELA	الى
WA	Waw	واو	SBJ	Subject Reference	مؤشر على الفاعل