Metadata of the chapter that will be visualized in SpringerLink

Book Title	Intelligent Systems and Applications	
Series Title		
Chapter Title	Arabic Tag Sets: Review	
Copyright Year	2019	
Copyright HolderName	Springer Nature Switzerland AG	
Corresponding Author	Family Name	**Alian**
	Particle	
	Given Name	**Marwah**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Hashemite University
	Address	Zarqa, Jordan
	Division	
	Organization	Princess Sumaya University for Technology
	Address	Amman, Jordan
	Email	Marwah2001@yahoo.com
Author	Family Name	**Awajan**
	Particle	
	Given Name	**Arafat**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Princess Sumaya University for Technology
	Address	Amman, Jordan
	Email	Awajan@psut.edu.jo

Abstract	Labeling a word with a suitable tag based on its context and its grammatical category is a major step in many applications of natural language processing. Constantly, there is an effort for inventing a set of these tags for Arabic language. In this research, a review for the existing Arabic tag sets is presented. A description for their features and limitations is also introduced.

Keywords (separated by '-')	Tag - Tag set - Arabic tag set

Arabic Tag Sets: Review

Marwah Alian[1,2(✉)] and Arafat Awajan[2]

[1] Hashemite University, Zarqa, Jordan
Marwah2001@yahoo.com
[2] Princess Sumaya University for Technology, Amman, Jordan
Awajan@psut.edu.jo

Abstract. Labeling a word with a suitable tag based on its context and its grammatical category is a major step in many applications of natural language processing. Constantly, there is an effort for inventing a set of these tags for Arabic language. In this research, a review for the existing Arabic tag sets is presented. A description for their features and limitations is also introduced.

AQ1

AQ2

Keywords: Tag · Tag set · Arabic tag set

1 Introduction

Part of Speech Tagging is the process of assigning proper tag for each word in a text representing its grammatical and morphological syntactic feature for a word [1]. Further, a tag is a code that holds simple or complex information that represent a word features and it labels the word in a text [2]. The development of a tag set that consist of representative tags at early stages is important for diacritical based tagging system. The need for such a tag set comes from the fact that Arabic language does not have a standard or complete tag set [3].

The approaches used for Part Of Speech (POS) tagging are classified into three main approaches; the first approach is the Rule-based Approach which sometimes called linguistic approach or Knowledge-Based Approach [4, 5], this approach use a set of linguistic rules during the process of tagging. The second approach is the Statistical Approach that is also called Probabilistic Approach or Stochastic Approach [6, 7]. This approach depends on building a statistical language model by gathering statistics from existing tagged corpora. The third approach is the hybrid approach in which rule-based and statistical approaches are involved [8, 10]. In the hybrid approach, both rule-based and statistical approaches are combined. On the other hand, some systems use other approaches, like machine learning algorithms, neural networks and decision trees [5, 9]. The existing Arabic tag sets vary in size from 6 tags to 2,000 detailed tags.

Some of these tag sets follow the same standards adopted in the tag set design for English, but these tag sets may be inappropriate for Arabic. Also, there are some morphological features that are common between Arabic tag sets like number, gender, case, person, definiteness and mood. However, the attributes are not uniformed among the morphological features [15]. In this research, a review for the existing Arabic tag sets is presented with their features and limitations.

K. Arai et al. (Eds.): IntelliSys 2018, AISC 868, pp. 1–15, 2019.
https://doi.org/10.1007/978-3-030-01054-6_43

2 Arabic Tag Sets

2.1 El-Kareh and Al-Ansary Tag Set

It consists of 72 tags and it was used in their semi-automatic tagger. El-Kareh and Al-Ansary tagger [8] was constructed from traditional Arabic grammar and it gives an accuracy of 90%. They classify words into three major classes: Verb, Noun, and Particle. However, Verbs were divided into three subclasses: Nouns into 46 subclasses and particles into 23 subclasses [1].

AQ3

2.2 Khoja Tag Set

Khoja [10] depends on ancient Arabic grammar to design a morphosyntactic tag set and she did not follow Indo-European tag sets which depend on Latin. All subcategories in Khoja tag set are derived from the parent categories therefore the tag set hold language generalization. In this tag set the noun is classified into: common, proper, numeral, adjective, and pronoun which have three subclasses: personal, relative, and demonstrative. However particle classified into: preposition, adverbial, conjunctions, interjections, explanations, exceptions, answer, negatives, and subordinates. Figure 1 shows an example of tagging a part of text using Khoja tag set.

بعث_VPSg3M خادم_NCSgMNI الحرمين_NCDuMAD الشريفين_NCDuMGD الملك_NCSgMND فهد_NP بن_
عبد NCSgMAI_ العزيز NCSgMAD_ آل R_ سعود NP_ برقية NCSgFNI_ تهنئة NCSgFGI_ الى PPr_
فخامة NCSgFGI_ الرئيس NCSgMGD_ الكسندر RF_ كواسنيفيسكي RF_ رئيس NCSgMNI_ جمهورية NCSgFGI_
بولندا RF_ بمناسبة NCSgFGI_ اليوم NCSgMAD_ الوطني NCSgMND_ لبلاده PPr_NCPlFGI_NPrPSg3M_

Fig. 1. Tagging a text using Khoja tag set [10].

Even that the work of Khoja has been the first comprehensive designed Arabic tag set and it is highly used in several applications, it has limitations [11].

2.3 Buckwalter Tag Set

Buckwalter tag set has two types: the first one is the untokenized tag set and the second is the tokenized tag set which contains around 500 tags. However, Buckwalter tag set consist of 70 sub tags that are possible to be combined to make around 170 complex tags with features for verbs, nominal and subject. For example, person, voice, aspect, and mood are included as verbs features [12]

The untokenized Buckwalter tag set contains around 485 tags and use the same basic 70 sub tags such as DET for determiner, ACC for accusative and ADJ for adjective. As an example for tagging the word الصين ‘China’ using Buckwalter tag set would be: DET+NOUN_PROP+CASE_DEF_NOM [13, 14].

Moreover, this tags set was reduced and called reduced Bukwalter in which the number of tags included in the new tag set is about 220 tags where case, mood and state.

2.4 Reduced Tag Set (RTS)

Linguistic Data Consortium (LDC) tag set was developed by the LDC team which consists of 24 tags and it was the reason behind introducing the reduced tag set (RTS) which goal was to increase the efficiency and performance of syntactic parsing for Arabic. RTS consists of 25 tags and follows the English tag set that where designed especially for Wall Street Journal. Also, RTS marks some features such as gender, person, definiteness, case and mood [4, 15].

2.5 Extended RTS

RTS was expanded to include 75 tags with adding only labeled morphological features on words. This tag set is called extended reduced tag set (ERTS) which holds the same features in RTS plus additional marked morphological features of number, nominal, definiteness, gender [4]. Table 1 shows some examples for Full Bukwalter, reduced RTS, and Extended ERTS tag sets.

Table 1. Buckwalter, reduced RTS and ERTS example

			Full	RTS	ERTS
حصيلة	HSylp	'result'	NOUN+NSUFF_FEM_SG +CASE_IND_NOM	NN	NNF
نهائية	nhA}yp	'final'	ADJ+NSUFF_FEM_SG +CASE_IND_NOM	JJ	JJF
حادث	HAdv	'accident'	NOUN+CASE_DEF_ACC	NN	NNM
النار	AlnAr	'the-fire'	DET+NOUN+CASE_DEF_GEN	NN	DNNM
الجماعي	AlimAEy	'group'	DET+ADJ+CASE_DEF_GEN	JJ	DJJM
شخصين	$xSyn	'two-persons'	NOUN +NSUFF_MASC_DU_GEN	NN	NNMDu

2.6 Penn Arabic Treebank (PATB)

PATB used widely as a tag set for Arabic [10] and it provides tokenization, complex POS tags, and syntactic structure. Also it provides discretization, empty categories, lemma and some semantic tags [7]. This tag set has over 400 tags that specify details about Arabic word morphology like definiteness, number, gender, person, voice, case and mood [16]. However, twenty dash tags are used for syntactic and semantic functions where syntactic tags have TPC and OBJ while semantic tags cover time (TMP) and location (LOC). Some tags in PATB can be used to label syntactic or semantic purpose like SBJ that is used to label semantic subject of adverb or syntactic subject of a verb [7]. The total number of tags used in PATB reaches 2,200 tag types to specify several issues and features for Arabic word morphology [16, 18] including 114 basic tags [10]. PATB is invented by the Linguistic Data Consortium in the University of Pennsylvania and it has four published versions and it provide a morphological analysis level of annotation where in PATB a morphological analyzer is used to produce a set of candidate analyses for every word then the linguists select the best

analyses for the word according to its context [4]. Figures 2 and 3 provide examples for PATB annotation.

```
(S  (NP-TPC-1 Huquwq+u حُقُوقُ
            (NP Al+<inosAn+i الإنْسَان ))
    (VP ta+qaE+u  تَقَعُ
       (NP-SBJ-1 *T*)
       (PP Dimona ضِمْنَ
          (NP <ihotimAm+i+nA إهْتِمامِنا )
       )))
```

حُقُوقُ الإنْسَانِ تَقَعُ ضِمْنَ إهْتِمامِنا

human rights exist within our concern

Fig. 2. PATB annotation example.

2.7 PADT Tag Set

The PADT tag is designed to have two main parts. The first part represents the part of speech using two letters while the second part represents features and it consists of seven letters to encode the values of the following features: Mood (Indicative, Subjunctive, Jussive or D for ambiguous), Voice (Active or Passive), Person (1 speaker, 2 addressee, 3 others), Gender (Masculine or Feminine), Number (Singular, Dual or Plural), Case (1: nominative, 2: genitive or 4: accusative) and Definiteness (Indefinite, Definite, Reduced or Complex) [7]. For example, the tag NNIS7 ----A---- means: noun, common noun, masculine inanimate, singular, 7th case which is instrumental, and negativeness is affirmative" [2]. Figure 4 shows another example using PADT tagging.

2.8 Alshamsi and Guessom Tag Set

This tag set is designed for the Hidden Marcov Model part of speech tagger system that Alshamsi and Guessom [6] invented to extract Name Entity. Their tag set consists of 55 tags and it is not a fine-grained tag set since they identify the NOUN category and subcategories by a number of tags such as NOUN (noun), adjective (ADJ), pronoun (PRON), proper noun (PNOUN), definite noun (DEF) and indefinite noun (INDEF). These tags are needed for their tagger that is used for Name Entity extraction.

Alshamsi and Guessom subdivide both noun and particle class into four subclasses. They point out, there is no need to have fine-grained a tag set since their tagger was intended to be for Named Entity extraction. For verb category, they use the tags: IVERB to represent imperfect verb, PVERB for perfect verb, CVERB for imperative verb, MOOD-.I for indicative, MOOD∼J for subjunctive, SUFF∼UBJ for suffix subject and FUTURE for future Imperative. Moreover for particle category they use the tags: INTERROGATE to represent interrogation, NEGATION to represent negation, CONJ to represent conjunction and PREP to represent preposition particles. Also, they

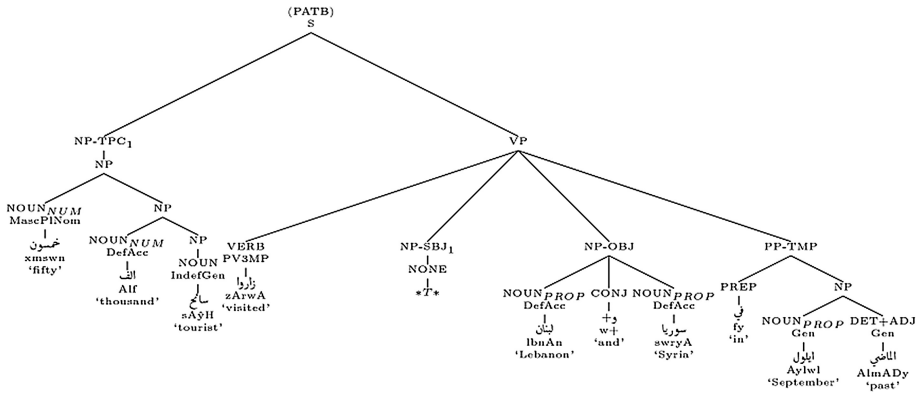

Fig. 3. In Penn Arabic Treebank (PATB) خمسون الف سائح زارو لبنان وسوريا في أيلول الماضي

Form	*al-lA-silkIyaTu*	*al-lā-silkīyatu* اَللَّاسِلْكِيَّةُ
Morph	al >| lA >| FiCL |< Iy |< aT |<< "u"	
Tag	A-----FS1D	
Form	*lA-silkIy*	*lā-silkīy* لَاسِلْكِيّ
Morph	lA >| FiCL |< Iy	
Root	"s l k"	
Reflex	wireless, radio	
Class	adjective	

Fig. 4. PADT example.

use features like person, number and gender with tag names in order to describe the morphology analysis of a word [6]. Figure 5 represents two examples for tagging two sentences with Alshamsi and Guessom tag set.

Word	Meaning	POS Tag
الحادي	The first of	DEF+ADJ
عشر	Ten	NOUN
من	From	PREP
أكتوبر	October	PNOUN

Word	فرنسي	شخص	ال	هذا
Transliteration	faransiyy	shakhS	al	haadhaa
Meaning	French	person	is	This
POS Tag	ADJ	NOUN	DEF	DPRON_MS

Fig. 5. Alshamsi and Guessom tagging for two sentences.

2.9 ARBTAGS Tag Set

ARBTAGS tag set has a hierarchy that makes it different than other tag sets. In this hierarchy, the noun is divided in to sixteen sub-classes while verb is classified into three sub-classes: perfect, imperfect and imperative. Particle class has seven sub-classes: preposition, vocative, conjunction, etc. Also, it has one punctuation tag. This tag set has a new additional tag that is used to present foreign words. Therefore, the size of general tags in ARBTAGS is 28 tags but it consists of 161 detailed tags: 101 tags for noun, 50 tags for verb, 9 for particle and one for punctuation. These tags have more features information that considered inflectional. For example; the word كتب is tagged by ARBTAGS as [VePeMaSnThSj] where the tag means [Perfect Verb, Masculine Gender, Singular Number, Third Person, and Subjunctive Mood [24]. Figure 6 shows the specifications of ARBTAGS formula.

The tagset has the following main formula:
[T , S , G , N , P , M , C , F] , Where:
T (Type) = {Verb, Noun, Particle}
S = Sub-Class {Common, Demonstrative, Relative,
Personal, Adverb,Diminutive,Instrument,
Conjunctive, Interrogative, Proper
and Adjective}
G (gender)= {Masculine, Feminine, Neuter}
N (Number) = {Singular, Plural, Dual}
P (Person) = {First, Second, Third}
M (Mood) = {Indicative, Subjunctive, Jussive}
C (Case) = {Nominative, Accusative, Genitive}
F (State) = {Definite, Indefinite}

Fig. 6. ARBTAGS tag formula [24].

2.10 CATiB Tag Set

Habash and Both construct the Columbia Arabic Treebank (CATiB) [23] which is a database of syntactic analysis used for Arabic text. CATiB differ than other Arabic Treebanking approaches in the constraints that it involves on linguistic richness and its focus on speed. CATiB approach has two basic ideas; the first one is to avoid annotation of redundant information while the second idea is to use terminology and representations inherited from traditional Arabic syntax. Thus, a simple parsing approach can produce grammar analysis [21].

CATiB uses MADA&TOKAN toolkit for initial tokenization and POS tagging. However, CATiB tokenization scheme is the same schema for PATB and PADT but the number of tags in CATiB is very small compared to the tags in PATB's.

CATiB has six POS tags: NOM, PROP, VRB, VRB-PASS, PRT, and PNX where NOM for non-proper nominals include nouns, pronouns, adjectives and adverbs, PROP for proper nouns, VRB for active-voice verbs, VRB-PASS for passive voice verbs, PRT for particles including prepositions or conjunctions and PNX for punctuation [23]. Figure 7 shows an example sentence that is tagged by CATiB.

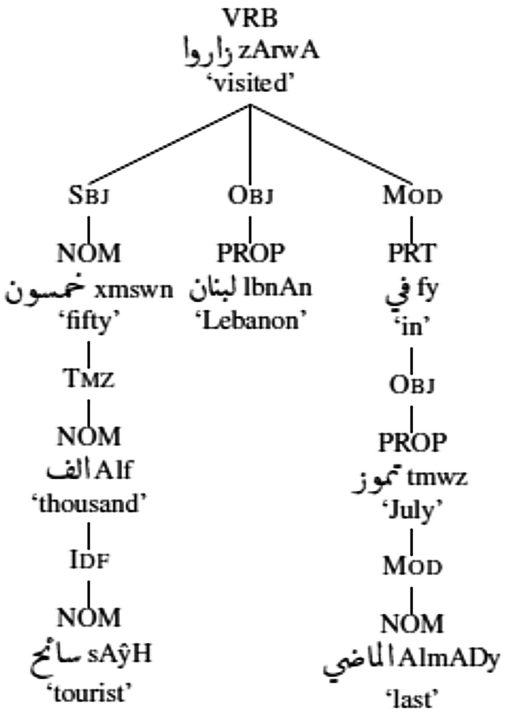

Fig. 7. CATiB example.

2.11 Yahya Elhadj Tag Set

Elhadj et al. [7] define 16 features as the main correlated morpho-syntactic features for a word then they add more fine grain features to get 71 features that are labeled with proper tags. In this tagging system the tag set covers three levels: the first level consists of 7 tags and the second level consists of 23 tags while the lower level has 54 tags. According to the requirements the tagging system can be used in any of the three levels. Figure 8 shows an example for using Elhadj tag set to tag Sect. 4 from Chapter al fateha in Quran.

2.12 SALMA Tag Set

SALMA Tag is introduced by Sawalha [15] to encode 22 morphological feature categories for each morpheme. The first character in the tag represents the word's class: verb, noun, particle or punctuation. The second character represents the noun subclass while the third character represents verb subclass while the fourth character represents the particle subclass. The fifth and sixth character are used for others and punctuations. The next characters from the seventh to the eighteen character in the tag represent morphological features as follows: gender in the seventh character, number in the eighth position, person in the ninth, inflectional morphology in the 10th, case and mood in the 11th, case and mood marks in the 12 position, definiteness in the 13, the 14

السورة	الأية	القطعة	النوع أو القسم	العدد	التركيب	التعريف	الجنس	الزمن	الوجه	الحالة الإعرابية	صيغة البناء	الشخص	تعدي	البناء	علامة الإعراب	وزن	الصحة	تجرد	جمود
الفاتحة	4	مالك	AJV	NS		DT	GM									PA	SS	FI	IT
الفاتحة	4	ي	MNV							CG					MI				
الفاتحة	4	يوم	NNN	NS	CU	DT	GM												
الفاتحة	4	ي	MNV							CG					MI				
الفاتحة	4	ل	MNV			DF								DJ					
الفاتحة	4	دين	NNN	NS	CU		GM												
الفاتحة	4	ي	MNV							CG					MI				

Fig. 8. Elhadj tag set tagging section (verse) 4 form chapter Alfateha in Quran [7].

character represent voice, emphasized and non-emphasized are presented in the 15 position, transitivity is presented in the 16, rational in the 17 character, declension and conjugation in the 18 position. The final four characters represent morphological information that is used in the analysis of Arabic text; the 19th character represent unaugmented and augmented, number of root letters represented in the 20th character, verb root in the 21 character and noun type in the final character. This tag set is utilized in Qutuf analyzer for Arabic morphology and part of speech tagger [15]. Figures 9 and 10 show examples for SALMA tagging.

Word	Morphemes			Tag
wa waaṣṣaynā And We have enjoined	وَوَصَّيْنَا	وَ *wa*	*And*	p--c------------------
		وَصَّيْ *waṣṣay*	*Have enjoined*	v-p---mpfs-s-amohvtt&-
		نَا *nā*	*We*	r---r-xpfs-s----hn----
al-'insāna (on) man	الْإِنسَانَ	الْ *al-*	*The*	r--d------------------
		إِنسَانَ *'insāna*	*man*	nq----ms-pafd---htbt-s
bi- wālidayhi His parents	بِوَالِدَيْهِ	بِ *bi*	*To*	p--p------------------
		وَالِدَ *wālida*	*Parents*	nu----md-vgki---htot-s
		يْ *y*	*Both*	r---r-xdts-s----------
		هِ *hi*	*His*	r---r-msts-k----------
ḥusn^{an} Kindness	حُسْنًا	حُسْنَ *ḥusn*	*kindness*	ng----ms-vafi---ndst-s
		اً ^{an}		r---k------f----------

Fig. 9. SALMA example 1 [15].

2.13 Ahmed H. Aliwy Tag Set

The main tags in this tag set are Noun, Verb, Particle, Residual and Punctuation where Noun has 17 subclasses with the features: Number, Gender, Case and Structured. Verb class has three subclasses: Past (Pst), Present (Prt), Imperative (Imv). While verb attributes are: Gender, Number, Person, Mood, Certainty, Structured, and Voice.

The subclasses of the particle are defined according to its working. In this tag set there are 21 subclasses for particles: for Jussive Jus, For Reduction, preposition Red, For Conjunction Cnj, for Accusative Acu, Not working or Preventive, NonCopier Cop

وَوَصَّيْنَا	وصي	فَعَّلْنَا	وَ p--c----------
			وَصَّيْ v-p---mpfs-s-amohvtt&-
			نَا r---r-xpfs-s----------
الْإِنْسَانَ	أنس	فِعْلَان	الْ r---d----------
			إِنْسَانَ nq----ms-pafd---hdbt-s
بِوَالِدَيْهِ	ولد	فَاعِل	بِ p--p----------
			وَالِدَ nq----ms-pafd---hdbt-s
			يْ r---r-xdts-s----------
			هِ r---r-msts-k----------
حُسْنًا	حسن	فُعْل	حُسْنَ ng----ms-vafi---ndst-s
			اً r---k------f----------

Fig. 10. SALMA tag example 2 [15].

and Prevent Prv. Residuals can be SymbolRSym, Abbreviations and Acronym RAbc or Not Classified RNcl. But there are no features for residuals and punctuations.

This tag set has 3552 tags since the combinations of some tags are impossible. The Noun tag has the form: N+POS_ Number+Gender+Case+Structured while Verb tag has the form: V +POS_Person+Number+Gender+Case+Structured+Certinity+Voice. Particles tag has the form: P+Working_Meaning. For Residual it can be one of three tags: ROth, RSys or RAcb, and the Punctuation tag is CPnc.

Figure 11 gives an example of the use of Aliwy Tag set for the Arabic text: "مرة وقبل سنتين كتبت عن العراق" [19].

It is considered as a multilevel tag set that is compatible with Classical Arabic and Modern Standard Arabic. Aliwy argued that his tag set has almost all Arabic features and classes where classes and features are selected carefully and his tag set has no interleaving [19].

3 Limitations in Arabic Tag Sets

Available Arabic tag sets do not have a standard scheme for correlating each word to its morpheme and they join the tagging of both morphemes and words. Moreover, many reports about these tag sets do not give a detailed description for their design aspects [22].

The existing tag sets have a limitation in covering all the features of Arabic language which leads to missing features.

However the analysis used for texts in designing existing tag sets is not competent for Arabic features and characteristics. Also a number of tagging systems involve a small number of tags that gives a narrow view about the text and they do not explain more about particles and verbs [7].

Clitics and word base			Tag
Token	Transliteration	Translation	
مرة	mrp	Once ,Time	NNou_SFNN
،	,	,	CPnc
و	w	And	PNon_Non
قبل	qbl	before	NAdv_SMAN
سنتين	sntyn	Two years	NNou_DFGN
،	,	,	CPnc
كتبت	ktbt	I wrote	VPst_3SMOYNA
عن	En	About	PRed_Adv
ال	Al	The	PNon_Def
عراق	ErAq	Iraq	NPrp_SMGN

Fig. 11. An example for the use of Aliwy tag set.

For example, in [17] they used only 24 tags and Catib [23] used only 8 tags. Such tagging systems may be insufficient for more general needs.

Even though the tag sets with large number of tags are complete and efficient for advanced tasks, they look very hard to be predicted while small tag sets tend to be more predictable and appropriate for many applications [20].

4 Summary

Since 2000, researchers have been introducing new Arabic tag sets or enhancing previously proposed tag sets. However, between 2010 and 2012, less attention was given to Arabic tag set until 2013 where Salma tag set was introduced by Sawalha [15] and Aliwy tag set was introduced by Ahmad Aliwy [19]. Table 2 shows a comparison between Arabic tag sets that were introduced in this report. It is shown that many tag sets has no particle attributes.

Existing tag sets shares common and basic tags while they differ in the number of levels they include in their morphological analysis and therefore they combine basic tags to get more complex tags as a feature for a word not as new tags. The tags details in any tag set depend on the application that the tag set was conducted specially for.

Table 2. Comparison between Arabic Tag Sets

Year	Tag set	Developed by	No. of tags	Simple/Complex	Tags details	Particle	Limitations
2000	El-Kareh S, Al-Ansary	El-Kareh S, Al-Ansary	72 tags	–	Words are classified into three main classes, Verbs, Noun and Particle. Each class is divided into subclasses, Verbs into 3 subclasses; Nouns into 46 subclasses and Particles into 23 subclasses.	23 sub-classes of the main class particle.	–
2001	Khoja	Shereen Khoja	177 tags	Simple	103 nouns, 57 verbs, 9 particles, 7 residual, and 1 punctuation.	No attributes	Many of Arabic classes are not taken into account.
2002	Buckwalter	Tim Buckwalter	485 tags-untokenized Thousands – tokenized	Complex	Form-based Very rich for many computational problems Tag set that can be used for tokenized and untokenized text Buckwalter's Tag set (170 morphemes, 500 tokenized tags, 22 K untokenized tags)	No attributes	No distinction between categories and features for POS
2004	Reduced Buckwalter tag sets BIES	Ann Bies and Dan Bikel	24	Very simple	Inspired by the Penn English Treebank	No attributes	The nouns, verbs and particles have no attributes.

(*continued*)

Table 2. (*continued*)

Year	Tag set	Developed by	No. of tags	Simple/Complex	Tags details	Particle	Limitations
2004-	The Extended Reduced Tag Set (ERTS)	Used in Amira system	72	–	A subset of the full Buckwalter morphological set	Added the explicit or marked morphological features of gender, number and definiteness on nominals	–
2004	Penn Treebank: PATB	Mohamed Maamouri and Ann Bies	2,000 tag types. This includes combinations of 114 basic tags.	Detailed tag set	Follows Arabic traditional grammer. Tags specify details about word morphology such as definiteness, number, case, person, voice, gender and mood.	No attribute	With some kinds of words, the PATB morphology systematically fails to determine many of the contextual and lexical parameters
2006	Alshamsi and Guessom	Alshamsi and Guessom	55	Very specific	Specific for Name Entity Take into account the structure of Arabic sentence		–
2008	ARBTAGS	AlQrainy & Ayesh	161 detailed tags and 28 general tags	Simple	Based on ancient Arabic grammar. 101 nouns, 50 verbs, 9 particles, 1 punctuation	No attributes	Punctuations and foreign words are not covered
2009	CATiB	Nizar Habash and Ryan M. Roth	6	The simplest tag set,	Using representations and terminology inspired by traditional Arabic syntax	No attributes.	Many classes and features are missed.

(*continued*)

Table 2. (*continued*)

Year	Tag set	Developed by	No. of tags	Simple/Complex	Tags details	Particle	Limitations
2009	Yahya Elhadj	Yahya Elhadj		Simple with respect to Noun, and Verb	Three classes (Noun, Verb, Particle), to cover the three categorization levels: 7 tags for the upper level, 23 for the inner level, and 54 for the lower level; 71 features	No attributes	No features for verbs.
2013	Salma	Majdi Sawalha	22 character for a tag	Simple	7 tags for the upper level, 23 for the inner level, and 54 for the lower level;	No variation	More theoretical Redundant tags
2013	Aliwy	Ahmad Aliwy	3552 detailed tags and 45 main tags.	Complex	17 tags for noun classes, 3 tags for verb, 21 tags for particles, 3 tags for Residuals and one for punctuation	21 tags for particles.	–

References

1. Abumalloh, R., Al-Sarhan, H., Ibrahim, O., Abu-Ulbeh, W.: Arabic part-of-speech tagging. J. Soft Comput. Decis. Support. Syst. **3**(2), 45–52 (2016)
2. Böhmová, A., Haji, J., Hajiová, E., Hladká, B.: The prague dependency treebank: a three level annotation scenario. In: Treebanks: Building and Using Parsed Corpora. Springer (2003)
3. Alqrainy, S., Ayesh, A., Almuaidi, H.: Automated tagging system and tagset design for arabic text. J. Comput. Linguist. Res. **1**(2), 55–62 (2010)
4. Maamouri, M., Bies, A.: Developing an Arabic treebank: methods, guidelines, procedures, and tools. In: Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (COLING), Geneva, pp. 2–9 (2004)
5. Alqrainy, S.: Morphological - syntactical analysis approach for Arabic textual tagging. Ph.D. thesis, De Montfort University (2008)
6. Al Shamsi, F., Guessoum, A.: A hidden Markov model–based POS tagger for Arabic (2006)
7. Elhadj, Y., Abdelali, A., Bouziane, R., Ammar, A.H.: Revisiting Arabic part of speech tagsets. In: Proceedings of 11th International Conference on Computer Systems and Applications (AICCSA), pp. 793–802 (2014)
8. El-Kareh, S., Al-Ansary, S.: An Arabic interactive multi-feature POS tagger. In: Proceedings of International Conference on Artificial and Computational Intelligence for Decision, Control, and Automation in Engineering and Industrial Applications (CIDCA), Monastir, Tunisia, pp. 204–210 AQ4
9. ElHadj, Y., Al-Sughayeir, I.A., Al-Aansari, A.M.: Arabic part-of-speech tagging using the sentence structure. In: Proceedings of 2nd International Conference on Arabic Language Resources and Tools. Cairo, pp. 241–245 (2009)
10. Khoja, S., Garside, R., Knowles, G.: A tagset for the morphosyntactic tagging of Arabic. In: Proceedings of Corpus Linguistics, Lancaster, pp. 341–353 (2001)
11. Abuzed, M., Arteimi, M.: Using the Brill of speech tagger for modern standard Arabic. In: The International Arab Conference on Information Technology (ACIT), Amman (2005)
12. Alosaimy, A.M.S. Atwell, E.S.: A review of morphosyntactic analysers and tag-sets for Arabic corpus linguistics. In: Corpus Linguistics, Lancaster, pp. 16–19 (2015)
13. Buckwalter, T.: Issues in Arabic orthography and morphology analysis. In: Proceedings of the Workshop on Computational Approaches to Arabic Script-Based Languages, pp. 31–34. COLING, Geneva (2004)
14. Alkuhlani, S., Habash, N., Roth, R.: Automatic morphological enrichment of a morphologically underspecified treebank. In: Association for Computational Linguistics (NAACL-HLT), Atlanta [s.n.], pp. 460–470 (2013)
15. Sawalha, M., Atwell, E., Abushariah, M.A.M.: SALMA: standard arabic language morphological analysis. In: Proceedings of 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA), Sharjah, pp. 1–6 (2013)
16. Smrž, O., Bielický, V., Kouřilová, I., Kráčmar, J., Hajič, J., Zemánek, P.: Prague Arabic dependency treebank: a word on the million words. In: Proceedings of the LREC Workshop on HLT & NLP within the Arabic World: Arabic Language and Local Languages (2008)
17. Diab, M., Kadri, H., Daniel, J.: Automatic tagging of Arabic text: from raw text to base phrase chunks. In: Proceedings of Human Language Technology-North American Association for Computational Linguistics (HLT-NAACL) (2004)
18. Diab, M.: Towards an optimal POS tag set for modern standard arabic processing. In: Proceedings of Recent Advances in Natural Language Processing (RANLP), Borovets (2007)

19. Aliwy, A.: Arabic morphosyntactic raw text part of speech tagging system. University of Warsaw, Faculty of Mathematics, Informatics and Mechanics (2013)
20. Habash, N.: Introduction to Arabic Natural Language Processing. Morgan & Claypool Publishers Series, San Rafael (2010)
21. Ibrahim, M.N.: Statistical Arabic grammar analyzer. In: Proceedings of 16th International Conference in Computational Linguistics and Intelligent Text Processing (CICLing), Cairo, pp. 187–200 (2015)
22. Sawalha, M., Atwell, E.: A standard tag set expounding traditional morphological features for Arabic language part-of-speech tagging. Word Struct. **6**(1), 43–99 (2013)
23. Habash, N., Roth, R.M.: CATiB: the Columbia Arabic treebank. In: Proceedings of the Association for Computational Linguistics (ACL-IJCNLP), pp. 221–224 (2009)
24. Alqrainy, S., Ayesh, A.: Developing a tagset for automated POS tagging in Arabic. In: Proceedings of the 10th WSEAS International Conference on COMPUTERS, Athens, pp. 956–961 (2006)

Author Query Form

Book ID : **473257_1_En**

Chapter No : **43**

Springer
the language of science

Please ensure you fill out your response to the queries raised below and return this form along with your corrections.

Dear Author,
During the process of typesetting your chapter, the following queries have arisen. Please check your typeset proof carefully against the queries listed below and mark the necessary changes either directly on the proof/online grid or in the 'Author's response' area provided below

Query Refs.	Details Required	Author's Response
AQ1	Please confirm if the corresponding author is correctly identified. Amend if necessary.	
AQ2	Per Springer style, both city and country names must be present in the affiliations. Accordingly, we have inserted the city names "Zarqa and Amman" and country name "Jordan" in affiliations "1 and 2". Please check and confirm if the inserted city and country names are correct. If not, please provide us with the correct city and country names.	
AQ3	Please confirm if the section headings identified are correct.	
AQ4	Please supply the year of publication for Ref. [8].	

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	··· under matter to remain	New matter followed by
Insert in text the matter indicated in the margin		or (x2)
Delete	through single character, rule or underline or through all characters to be deleted	or (x2)
Substitute character or substitute part of one or more word(s)	through letter or through characters	new character or new characters
Change to italics	under matter to be changed	
Change to capitals	under matter to be changed	
Change to small capitals	under matter to be changed	
Change to bold type	under matter to be changed	
Change to bold italic	under matter to be changed	
Change to lower case	Encircle matter to be changed	
Change italic to upright type	(As above)	
Change bold to non-bold type	(As above)	
Insert 'superior' character	through character or where required	or under character e.g. or
Insert 'inferior' character	(As above)	over character e.g.
Insert full stop	(As above)	
Insert comma	(As above)	
Insert single quotation marks	(As above)	or and/or or
Insert double quotation marks	(As above)	or and/or or
Insert hyphen	(As above)	
Start new paragraph		
No new paragraph		
Transpose		
Close up	linking characters	
Insert or substitute space between characters or words	through character or where required	
Reduce space between characters or words	between characters or words affected	