A standard tag set expounding traditional morphological features for Arabic language part-of-speech tagging[1]

Majdi Sawalha and Eric Atwell

Abstract

The SALMA Morphological Features Tag Set (SALMA, Sawalha Atwell Leeds Morphological Analysis tag set for Arabic) captures long-established traditional morphological features of grammar and Arabic, in a compact yet transparent notation. First, we introduce Part-of-Speech tagging and tag set standards for English and other European languages, and then survey Arabic Part-of-Speech taggers and corpora, and long-established Arabic traditions in analysis of morphology. A range of existing Arabic Part-of-Speech tag sets are illustrated and compared; and we review generic design criteria for corpus tag sets. For a morphologically-rich language like Arabic, the Part-of-Speech tag set should be defined in terms of morphological features characterizing word structure. We describe the SALMA Tag Set in detail, explaining and illustrating each feature and possible values. In our analysis, a tag consists of 22 characters; each position represents a feature and the letter at that location represents a value or attribute of the morphological feature; the dash '-' represents a feature not relevant to a given word. The first character shows the main Parts of Speech, from: noun, verb, particle, punctuation, and Other (residual); these last two are an extension to the traditional three classes to handle modern texts. 'Noun' in Arabic subsumes what are traditionally referred to in English as 'noun' and 'adjective'. The characters 2, 3, and 4 are used to represent subcategories; traditional Arabic grammar recognizes 34 subclasses of noun (letter 2), 3 subclasses of verb (letter 3), 21 subclasses of particle (letter 4). Others (residuals) and punctuation marks are represented in letters 5 and 6 respectively. The next letters represent traditional morphological features: gender (7), number (8), person (9), inflectional morphology (10) case or mood (11), case and mood marks (12), definiteness (13), voice (14), emphasized and non-emphasized (15), transitivity (16), rational (17), declension and conjugation (18). Finally there are four characters representing morphological information which is useful in Arabic text analysis, although not all linguists would

Word Structure 6.1 (2013): 43–99
DOI: 10.3366/word.2013.0035

www.euppublishing.com/word

count these as traditional features: unaugmented and augmented (19), number of root letters (20), verb root (21), types of nouns according to their final letters (22). The SALMA Tag Set is not tied to a specific tagging algorithm or theory, and other tag sets could be mapped onto this standard, to simplify and promote comparisons between and reuse of Arabic taggers and tagged corpora.

1. Introduction: part-of-speech tagging and part-of-speech tag sets

Part-of-speech taggers are used to enrich a corpus by adding a part-of-speech category label to each word, showing the broad grammatical class of the word, and morphological features such as tense, number, gender, etc. The list of all grammatical category labels is called the tag set. The design of the tag set is an important prerequisite to this annotation task. The task requires a tagging scheme, where each tag or label is practically defined by showing the words and contexts where each tag applies; and a tagger, a program responsible for assigning a tag to each word in the corpus by implementing tag set and tagging scheme in a tag-assignment algorithm (Atwell 2008).

Automatic taggers have been used from the early years of Corpus Linguistics. TAGGIT in 1971 achieved an accuracy of 77% tested on the Brown corpus. In the late 1970s, CLAWS1, a data-driven statistical tagger was built to carry out the annotation of the Lancaster/ Oslo-Bergen corpus (LOB), and had an accuracy rate of 96–97%. Later tagger development included systems based on Hidden Markov Models (HMM); HMM taggers have been made for several languages. The Brill tagger (Brill 1995) is an example of data-driven symbolic tagger. The ENGCG and EngCG-2 are based on a framework known as Constraint Grammar (CG) (Voutilainen 2003).

Recently, many new systems based on a variety of Markov Model and Machine Learning (ML) techniques have appeared for many languages. Hybrid solutions have also been investigated (Voutilainen 2003). ACOPOST,[2] A Collection Of POS Taggers, consists of four taggers of different frameworks: Maximum Entropy Tagger (MET), Trigram Tagger (T3), Error-driven Transformation-Based Tagger (TBT) and Example-based tagger (ET). The SNoW-based Part of Speech Tagger[3] and LBJ Part of Speech Tagger[4] make use of the Sequential Model. NLTK,[5] the Natural Language Toolkit, includes Python re-implementations of several POS taggers such as; Regexp Tagger, N-Gram Tagger, Brill Tagger and HMM Tagger; in addition NLTK includes tutorials and documentation on tagging. RelEx[6] provides English-language part-of-speech tagging, entity tagging, as well as other types of tags (gender, date, money, etc.). Spejd[7] – Shallow Parsing and Disambiguation Engine is a tool for simultaneous rule-based morphosyntactic disambiguation and partial parsing. VISL Constraint Grammar[8] is an example of rule based disambiguation.

Enriching the source text samples of corpora with part-of-speech information for each word, as a first level of linguistic enrichment, results in more useful research resources. English corpora have been developed for a long time and for a variety of formats, types and genres. Several English corpora have been enriched with Part-of-Speech tagging, and a variety of different English corpus part-of-speech tag sets have been developed, including: the Brown corpus (BROWN), the

Lancaster/ Oslo-Bergen corpus (LOB), the Spoken English Corpus (SEC), the Polytechnic of Wales corpus (PoW), the University of Pennsylvania corpus (UPenn), the London-Lund Corpus (LLC), the International Corpus of English (ICE), the British National Corpus (BNC), the Spoken Corpus Recordings In British English (SCRIBE), etc (Atwell 2008). The AMALGAM[9] multi-tagged corpus amalgamates all these tagging schemes in a common collection of English texts: in the AMALGAM corpus, the different part-of-speech tag sets used in these English general-purpose corpora are applied to illustrate the range of rival English corpus tagging schemes, and the texts are also parsed according to a range of rival parsing schemes, so each sentence has more than one parse-tree, called 'a forest' (Atwell, Demetriou, Hughes, Schiffrin, Souter & Wilcock 2000). Part-of-speech tag sets and taggers have also been developed for other European languages. The EAGLES, European Advisory Group on Language Engineering Standards project, drew up standards for tag sets, morphological classes and codes for (western) European languages, including EAGLES Recommendations for the morphosyntactic annotation of corpora (Leech & Wilson 1999); a synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora: a common proposal and applications to European languages (Monachini & Calzolari 1996); and an EAGLES study of the relation between tag sets and taggers (Teufel, Schmid, Heid & Schiller 1996).

The potential uses of a part-of-speech tagged corpus are key factors in deciding the range and number of part-of-speech tags. Many linguistic analyses use part-of-speech tagged corpora to analyse text and extract information, where part-of-speech tags play an essential role in classifying text and direct search to the actions, events, places, etc described in the text. The most obvious applications are in lexicography and natural language processing (NLP) computational linguistics. Further applications include using the tags in data compression (Teahan 1998); and as a possible guide in the search for extra-terrestrial intelligence (Elliott & Atwell 2000). Other generic applications that make use of part-of-speech tag information are: searching and concordancing, grammatical error detection in Word Processing, training Neural Networks for grammatical analysis of text, or training statistical language processing models (Atwell 2008). Part-of-Speech tagging is a key technology in discovering suspicious events from text (Zolfagharifard 2009), and processing Arabic is a key task in discovering these suspicious events.

1.1 Arabic language part-of-speech taggers and corpora

Arabic part-of-speech tagging development started more recently. A range of different techniques have been used to solve the problem of part-of-speech tagging of Arabic. The APT tagger uses a combination of both statistical Viterbi algorithm, and rule-based techniques (Khoja 2001). Brill's 'transformation-based' or 'rule-based' part-of-speech tagger has been applied for Arabic (Freeman 2001). Harmain (2004) developed a web-based Arabic tagger. Diab, Hacioglu & Jurafsky (2004) used Support Vector Machines (SVM), a supervised learning algorithm, to achieve an accuracy of 95%. Habash & Rambow (2005) developed another part-of-speech tagger that uses SVM and

Viterbi decoding. HMM has been widely used in part-of-speech tagging for Arabic, with reported accuracy of 97% on LDC's Arabic Treebank of Modern Standard Arabic (Al-Shamsi & Guessoum 2006) and 70% when tested on CallHome Egyptian Colloquial Arabic (ECA) and the LDC Levantine Arabic (LDC) (Duh & Kirchhoff 2005). Applications of Memory-based learning to morphological analysis and part-of-speech tagging of written Arabic have been explored (Marsi, Bosch & Soudi 2005). Also, combinations of rule based and machine learning methods for tagging Arabic words (Tlili-Guiassa 2006). A multi-agent architecture was developed to address the problem of part-of-speech tagging of Arabic text with vowel marks (Zibri, Torjmen & Ahmad 2006). A rule-based PoS tagging system, Arabic Morphosyntactic Tagger AMT (Alqrainy 2008), uses two different techniques: the pattern-based technique, which is based on using Pattern-Matching Algorithm (PMA), and lexical and contextual technique. The AMT tagger makes use of the last diacritic mark of Arabic words to reduce the tagging ambiguity. The accuracy of the AMT tagger reported was 91%.

Nearly all these Arabic part-of-speech taggers were developed by NLP research groups for their own internal use, and are not freely downloadable by other researchers. The taggers use different tag sets, and accuracies are reported on different test corpora.

Arabic corpora[10] started to appear in the late 1980s; the following list of Arabic corpora developed outlines their size, type, purpose of development and the materials of construction (Al-Sulaiti & Atwell 2006):

- **Buckwalter Arabic Corpus** (1986–2003) consists of about 3 million words of public resources in the web to be used in lexicography.
- **Leuven Corpus** (1990–2004) developed at Catholic University Leuven, Belgium, consists of about 3 million words of written and spoken text from internet sources, radio and TV and primary school books, to be used in the development of Arabic-Dutch/Dutch-Arabic learner's dictionaries.
- **Arabic Newswire Corpus** (1994) developed at the University of Pennsylvania LDC, consists of 80 million words of written text collected from Agence France Presse (AFP), Xinhua News Agency, and Umma Press, to be used in education and the development of technology.
- **CALLFRIEND Corpus** (1995) developed at the University of Pennsylvania LDC consists of 60 telephone conversations of Egyptian native speakers, to be used in the development of language identification technology.
- **Nijmegen Corpus** (1996) developed at Nijmegen University consists of over 2 million written words collected from magazines and fiction, to be used in Arabic-Dutch/Dutch-Arabic dictionaries.
- **CALLHOME Corpus** (1997) developed at the University of Pennsylvania LDC, consists of 120 telephone conversations of Egyptian native speakers, to be used in telephony and speech recognition.
- **CLARA** (1997) developed at Charles University, Prague, consists of 50 million words collected from periodicals, books, internet sources from 1975-present, to be used for lexicography.

- **Egypt** (1999) developed at Johns Hopkins University, a parallel corpus of the Qur'an in English and Arabic to be used in machine translation.
- **Broadcast News Speech** (2000) developed at University of Pennsylvania LDC, consists of more than 110 news broadcasts from the Voice of America radio station, to be used in speech recognition.
- **DINAR Corpus** (2000) developed at Nijmegen University and SOTETEL-IT, in co-ordination with Lyon2 University, consists of 10 million words, to be used in lexicography, general research, and NLP.
- **An-Nahar Corpus** (2001) developed by ELRA, consists of 140 million words of written text collected from An-Nahar newspaper (Lebanon), to be used in general text research.
- **Al-Hayat Corpus** (2002) developed by ELRA consists of 18.6 million of written text collected from Al-Hayat newspaper (Lebanon), to be used for language engineering and information retrieval applications.
- **Arabic Gigaword** (2002) developed at the University of Pennsylvania LDC, consists of around 400 million words collected from Agence France Presse (AFP), Al-Hayat news agency, An-Nahar news agency and Xinhua news agency, to be used in natural language processing, information retrieval and language modelling.
- **E-A Parallel Corpus** (2003) developed at the University of Kuwait, consists of 3 million words of written text collected from publications from Kuwait National Council, to be used in teaching, translation and lexicography.
- **General Scientific Arabic Corpus** (2004) developed at UMIST, UK, consists of 1.6 words of written text, to be used in investigating Arabic compounds.
- **Classical Arabic Corpus** (CAC) (2004) developed at UMIST, UK, consists of 5 million words of written text, to be used in lexical analysis.
- **Multilingual Corpus** (2004) developed at UMIST, UK, consists of 11.5 million words of written text including 2.5 million words in Arabic, collected from IT-specialized websites to be used in translation studies.
- **SOTETEL Corpus** developed at SOTETEL-IT, Tunisia, consists of 8 million words of written text collected from literature, academic and journalistic materials, to be used in lexicography.
- **Corpus of Contemporary Arabic (CCA)** (2004) developed at the University of Leeds, consists of 1 million words of written and spoken data, collected from websites and online magazines, to be used in language teaching and language technology.
- **DARPA Babylon Levantine Arabic Speech and Transcripts** (2005) developed at the University of Pennsylvania LDC, consists of about 2,000 telephone calls collected from Fisher style telephone speech collection, to be used in machine translation, speech recognition and spoken dialogue systems.
- **The Penn Arabic Treebank** (2001) Part 1 consists of 166,000 words of written Modern Standard Arabic newswire from the Agence France Presse

corpus; and Part 2 consists of 144,000 words from Al-Hayat distributed by Ummah Arabic News Text, to be used in computational linguistics. New features of annotation in the UMAAH (UMmah Arabic Al-Hayat) corpus include complete vocalization (including case endings), lemma IDs, and more specific part-of-speech tags for verbs and particles. The Arabic Treebank corpora are annotated for morphological information, part-of-speech, English gloss (all in the "part-of-speech" phase of annotation), and for syntactic structure (Maamouri & Bies, 2004).

- **The Quranic Arabic Corpus (QAC)** (2009) contains the classical Arabic source text of the Quran, the holy book of Islam. The text consists of nearly 80,000 words, divided into numbered chapters and verses. The text is being enriched with morphological analysis, Part-of-Speech tagging, dependency parsing, coreference resolution, and other linguistic markup, via a collaborative web-based project. The annotated corpus is online, used by Quranic scholars, linguists, and the general public with an interest in Islam.

Nearly all these corpora have been collected by Arabic corpus linguistics research groups for their own purposes, and are not freely downloadable. The Corpus of Contemporary Arabic (CCA) (Al-Sulaiti & Atwell 2004; Al-Sulaiti & Atwell 2005; Al-Sulaiti & Atwell 2006), and the Quranic Arabic Corpus (QAC) (Dukes, Atwell & Sharaf 2010), both developed at the University of Leeds, are the only freely available Arabic corpora on the web which have been widely reused for linguistic research. The CCA is not annotated with part-of-speech tags, but the QAC is annotated with morphological segmentation and morpho-syntactic tags. In computational linguistics research, the most widely used annotated corpus of Arabic is the Penn Arabic Treebank (Maamouri & Bies 2004) developed at the University of Pennsylvania and distributed (at cost) by the LDC Linguistic Data Consortium.

1.2 Traditional Arabic part-of-speech classification

Arabic, unlike English and modern European languages, has a long tradition of scholarly research into its grammatical description, spanning over a millennium. Most traditional Arabic grammar studies follow the order established by سِيبَوَيْه *Sībawayh*, about fourteen hundred years ago. It starts with syntax نحو *naḥw*, followed by morphology تصريف *taṣrīf*, and phonology علم الأصوات *'ilm al-'aṣwāt*. The grammarian's main preoccupation was the explanation of the case ending of the words in the sentence, called إعراب *'i'rāb*. The term originally meant the correct use of Arabic according to the language of the Bedouins but came to mean 'declension'. Classical Arabic linguists classify words into three main parts of speech: Noun, name of a person, place, or object which does not have any tense; Verb, a word which indicates an action and has tense; and Particle, a word which cannot be understood without being connected to a noun or a verb or both. However, there are also morphological criteria for this classification: a verb can be defined as a word derived from a specified morphological pattern, and has morphological features such as person and mood; while

a noun can be definite or indefinite and has number and gender features. Derived nouns, which are derived from verbs, may have the same pattern as verbs. Particles are considered the most idiosyncratic words in Arabic, as these particles might span several grammatical categories. For example the particle *wa* وَ can indicate a conjunction between two adjectives قَضيتُ وقْتاً سعيداً وَ مُمتعاً في الحَفْلَةِ *qaḍaytu waqt^{an} sa'īd^{an} wa mumti'^{an} fī al-ḥaflati* 'I spent an interesting and happy time at the party', while, in another case, the same particle *wa* وَ functions as locative preposition in the sentence مَشَيتُ وَ النَّهرَ *mašaytu wa an-nahra* 'I walked along the river' (Al-Ghalayyni 2005).

Arabic is a highly inflectional language, and the traditional classification into nouns, verbs and particles does not say much about word structure. Arabic has many morphological and grammatical features, including sub-categories, person, number, gender, case, mood, etc. (Atwell 2008). A fine-grained tag set is appropriate for morphology research. The additional information may also help to disambiguate the base grammatical class (Schmid & Laws 2008). We aim to develop a part-of-speech tagger for annotating general-purpose Arabic corpus resources, in a wide range of text formats, domains and genres, including both vowelized and non-vowelized text; enriching the text with linguistic analysis will maximize the potential for corpus re-use in a wide range of applications. We foresee an advantage in enriching the text with part-of-speech tags showing very fine-grained grammatical distinctions, which reflect expert interest in syntax and morphology, rather than specific needs of end-users, because end-user applications are not known in advance.

Very fine-grain distinctions may cause problems for automatic tagging if some words can change grammatical tag depending on function and context (Atwell 2008); on the other hand, fine-grained distinctions may actually help to disambiguate other words in the local context. Practical experiments using a fine-grain morphological tag set were reported by (Schmid & Laws 2008). Their experiments were carried out using German and Czech as examples of highly inflectional languages. Their HMM part-of-speech tagger makes good use of the fine-grain tag set; it splits the part-of-speech into attribute vectors and estimates the conditional probabilities of the attribute with decision trees. This method achieved a higher tagging accuracy than two state-of-the-art general-purpose part-of-speech taggers (TnT and SVMTool). We believe that this kind of approach may yield better results for an Arabic part-of-speech tag set including fine-grained morphological features.

1.3 Existing Arabic part-of-speech tag sets

This section covers the most important Arabic tag sets and tag set design methodologies. These tag sets are; (1) Khoja's Arabic tag set, (2) Penn Arabic Treebank tag set, (3) ARBTAGS, (4) The Quranic Arabic Corpus morphological tag set, (5) The MorphoChallenge 2009 Qur'an Gold Standard tag set and (6) CATiB part-of-speech tag set. The section describes each tag set and their characteristics, and a comparison table illustrates the differences between the different Arabic tag sets. The tag sets range from a small set of short tags analogous to BNC or LOB tag sets for English on one hand,. to On the other hand, longer more detailed morphological tag

Word			Khoja's part-of-speech tag
تنفيذاً	*tanfīḏ*^{an}	Implementation	**NCSgMI**
لتوجيهات	*li-tawǧīhāt*	directives	**PPr'NCSgMI**
خادم	*ḫādim*	Custodian	**NCSgMI**
الحرمين	*al-ḥaramayn*	Two Mosques	**NCDuMD**
الشريفين	*aš-šarīfayn*	Holy	**NCDuMD**

Figure 1. Example of tagged sentence using Khoja's tag set.

sets (*e.g.* Penn Arabic Treebank (FULL) tag set) which are analogous to the ICE tag set for English.

1.3.1 Khoja's Arabic tag set

During early research on developing a part-of-speech tagger for Arabic text, Khoja (Khoja, Garside & Knowles 2001; Khoja 2003) developed a tag set for Arabic which is based on traditional Arabic grammar categories rather than on modern European EAGLES standards. The reasons for not following EAGLES morphosyntactic guidelines were: Arabic belongs to the Semitic language family while EAGLES guidelines were designed for European languages; and following EAGLES guidelines would not capture some Arabic morphosyntactic information such as imperative or jussive mood, dual number and inheritance. Inheritance is an important aspect of Arabic, where all subclasses of words inherit properties from the classes from which they are derived. Khoja's tag set contains 177 tags; 103 types of noun, 57 verbs, 9 particles, 7 residuals and 1 punctuation. Khoja's tag set includes the morphological features of gender, number, person, case, definiteness and mood. Figure 1 shows an example of a part-of-speech annotated sentence تنفيذاً لتوجيهات خادم الحرمين الشريفين *tanfīḏ*^{an} *li-tawǧīhāt ḫādim al-ḥaramayn aš-šarīfayn* 'Implementation of the directives of the Custodian of the Two Holy Mosques', taken from the training corpus of the APT tagger (Khoja 2003).

1.3.2 Penn Arabic Treebank (PATB) part-of-speech tag set

The most widely used tag set for Arabic is the Penn Arabic Treebank tag set used to annotate the Penn Arabic Treebank (PATB) with part-of-speech tags. Tim Buckwalter's morphological analyser was used to compute a set of candidate solutions or analyses for each word, and then Arabic linguists selected the solution which best fitted the context. The Penn Arabic Treebank model postulates a FULL tag set which compromises over 2,000 tag types (Diab 2007). This includes combinations of 114 basic tags listed in the Linguistic Data Consortium (LDC) Arabic part-of-speech/morphological tagging documentation.[11] Figure 2 shows these basic tags.

ABBREV	IVSUFF_SUBJ:2FS_MOOD:SJ	POSS_PRON_3MP
ADJ	IVSUFF_SUBJ:D_MOOD:I	POSS_PRON_3MS
ADV	IVSUFF_SUBJ:D_MOOD:SJ	PREP
CONJ	IVSUFF_SUBJ:FP	PRON_1P
DEM_PRON_F	IVSUFF_SUBJ:MP_MOOD:I	PRON_1S
DEM_PRON_FD	IVSUFF_SUBJ:MP_MOOD:SJ	PRON_2FS
DEM_PRON_FS	NEG_PART	PRON_2MP
DEM_PRON_MD	NO_FUNC	PRON_2MS
DEM_PRON_MP	NON_ALPHABETIC	PRON_3D
DEM_PRON_MS	NON_ARABIC	PRON_3FP
DET	NOUN	PRON_3FS
EMPHATIC_PARTICLE	NOUN_PROP	PRON_3MP
EXCEPT_PART	NSUFF_FEM_DU_ACCGEN	PRON_3MS
FUNC_WORD	NSUFF_FEM_DU_ACCGEN_POSS	PUNC
FUT	NSUFF_FEM_DU_NOM	PVSUFF_DO:1P
INTERJ	NSUFF_FEM_DU_NOM_POSS	PVSUFF_DO:1S
INTERROG_PART	NSUFF_FEM_PL	PVSUFF_DO:3D
IV1P	NSUFF_FEM_SG	PVSUFF_DO:3FS
IV1S	NSUFF_MASC_DU_ACCGEN	PVSUFF_DO:3MP
IV2D	NSUFF_MASC_DU_ACCGEN_POSS	PVSUFF_DO:3MS
IV2FS	NSUFF_MASC_DU_NOM	PVSUFF_SUBJ:1P
IV2MP	NSUFF_MASC_DU_NOM_POSS	PVSUFF_SUBJ:1S
IV2MS	NSUFF_MASC_PL_ACCGEN	PVSUFF_SUBJ:2FS
IV3FD	NSUFF_MASC_PL_ACCGEN_POSS	PVSUFF_SUBJ:2MP
IV3FP	NSUFF_MASC_PL_NOM	PVSUFF_SUBJ:3FD
IV3FS	NSUFF_MASC_PL_NOM_POSS	PVSUFF_SUBJ:3FP
IV3MD	NSUFF_MASC_SG_ACC_INDEF	PVSUFF_SUBJ:3FS
IV3MP	NUM	PVSUFF_SUBJ:3MD
IV3MS	NUMERIC_COMMA	PVSUFF_SUBJ:3MP
IVSUFF_DO:1P	PART	PVSUFF_SUBJ:3MS
IVSUFF_DO:1S	POSS_PRON_1P	REL_PRON
IVSUFF_DO:2MP	POSS_PRON_1S	REL_ADV
IVSUFF_DO:2MS	POSS_PRON_2FS	RESULT_CLAUSE_PART
IVSUFF_DO:3D	POSS_PRON_2MP	ICLE
IVSUFF_DO:3FS	POSS_PRON_2MS	SUBJUNC
IVSUFF_DO:3MP	POSS_PRON_3D	VERB_IMPERFECT
IVSUFF_DO:3MS	POSS_PRON_3FP	VERB_PERFECT
	POSS_PRON_3FS	VERB_PASSIVE

Figure 2. The Penn Arabic Treebank Tag Set; basic tags, which can be combined.

The FULL tag set exhibits a wide range of morphological features: case, gender, number, definiteness, mood, person, voice, tense, aspect, etc. The LDC also introduced the reduced tag set (RTS) of 25 tags which is designed to maximize the performance of Arabic syntactic parsing. The RTS follows the tag set designed for the English Wall Street Journal. The morphological features marked by the RTS tag set are case, mood, gender, person and definiteness (Diab 2007).

Figures 3–6 show examples of two sentences tagged by the FULL tag set. The first sentence is a newspaper text taken from the Arabic Treebank: تم أعداد الوثائق المتوفرة بكثرة حول أول رحلة طيران عثمانية فوق البلاد العربية *tamma ’i‘dād al-waṯā’iqa al-mutawaffira^{ti} bikaṯra^{in} ḥawla ’awwali riḥla^{ti}ṭayyarān^{in} ‘uṯmāniyya^{tin} fawqa al-bilādi al-‘arabiyya^{ti}* ‘Many available documents relate to the first Ottoman’s flight over the Arab countries’. The second sentence is taken from the Qur’an (chapter 29): وَوَصَّيْنَا الْإِنسَانَ بِوَالِدَيْهِ حُسْنًا *wa waṣṣaynā al-’insāna biwālidayhi ḥusn^{an}* ‘We have enjoined

```
INPUT STRING: تم
LOOK-UP WORD: tm
     Comment:
* SOLUTION 1: (tam~) tam~/VERB_PERFECT
     (GLOSS):  + conclude/take place +
INPUT STRING: اعداد
LOOK-UP WORD: AEdAd
     Comment:
  SOLUTION 1: (>aEodAd) >aEodAd/NOUN
     (GLOSS):  + numbers/issues +
* SOLUTION 2: (<iEodAd) <iEodAd/NOUN
     (GLOSS):  + preparation +
INPUT STRING: الوثائق
LOOK-UP WORD: AlwvA}q
     Comment:
* SOLUTION 1: (AlwavA}iq) Al/DET+wavA}iq/NOUN
     (GLOSS): the + documents/charters +
INPUT STRING: المتوفرة
LOOK-UP WORD: Almtwfrp
     Comment:
* SOLUTION 1: (Almutawaf~irap) Al/DET+mutawaf~ir/ADJ+ap/NSUFF_FEM_SG
     (GLOSS): the + available/abundant + [fem.sg.]
INPUT STRING: ب
LOOK-UP WORD: b
     Comment: Separated
* SOLUTION 1: (bi-) bi-/PREP
     (GLOSS): by/with
INPUT STRING: كثرة
LOOK-UP WORD: kvrp
     Comment:
* SOLUTION 1: (-kavorap) -kavor/NOUN+ap/NSUFF_FEM_SG
     (GLOSS): abundance/frequency + [fem.sg.]
INPUT STRING: حول
LOOK-UP WORD: Hwl
     Comment:
* SOLUTION 1: (Hawola) Hawola/PREP
     (GLOSS):  + about/around +
  SOLUTION 2: (Haw~al) Haw~al/VERB_PERFECT
     (GLOSS):  + change/convert/switch +
  SOLUTION 3: (Hawol) Hawol/NOUN
     (GLOSS):  + power +
INPUT STRING: أول
LOOK-UP WORD: >wl
     Comment:
  SOLUTION 1: (>aw~al) >aw~al/VERB_PERFECT
     (GLOSS):  + explain/interpret +
* SOLUTION 2: (>aw~al) >aw~al/ADJ
     (GLOSS):  + first +
  SOLUTION 3: (>uwal) >uwal/ADJ
     (GLOSS):  + first +
INPUT STRING: رحلة
LOOK-UP WORD: rHlp
     Comment:
* SOLUTION 1: (riHolap) riHol/NOUN+ap/NSUFF_FEM_SG
     (GLOSS):  + journey/career + [fem.sg.]
INPUT STRING: طيران
LOOK-UP WORD: TyrAn
     Comment:
* SOLUTION 1: (TayarAn) TayarAn/NOUN
     (GLOSS):  + airline/aviation +
INPUT STRING: عثمانية
LOOK-UP WORD: EvmAnyp
     Comment:
  SOLUTION 1: (EuvomAniy~ap) EuvomAniy~/NOUN+ap/NSUFF_FEM_SG
     (GLOSS):  + Ottoman + [fem.sg.]
* SOLUTION 2: (EuvomAniy~ap) EuvomAniy~/ADJ+ap/NSUFF_FEM_SG
     (GLOSS):  + Ottoman + [fem.sg.]
INPUT STRING: فوق
LOOK-UP WORD: fwq
     Comment:
* SOLUTION 1: (fawoq) fawoq/PREP
     (GLOSS):  + above/over +
  SOLUTION 2: (fawoq) fawoq/NOUN
     (GLOSS):  + top/upper part +
INPUT STRING: البلاد
LOOK-UP WORD: AlblAd
     Comment:
* SOLUTION 1: (AlbilAd) Al/DET+bilAd/NOUN
     (GLOSS): the + (native) country/countries +
INPUT STRING: العربية
LOOK-UP WORD: AlErbyp
     Comment:
  SOLUTION 1: (AlEarabiy~ap) Al/DET+Earabiy~/ADJ+ap/NSUFF_FEM_SG
     (GLOSS): the + Arab/Arabic + [fem.sg.]
* SOLUTION 2: (AlEarabiy~ap) Al/DET+Earabiy~/ADJ+ap/NSUFF_FEM_SG
     (GLOSS): the + Arab + [fem.sg.]
```

Figure 3. Buckwalter's morphological analysis of a sentence from the Arabic Treebank.

```
تم (tam~)                  tam~/VERB_PERFECT
اعداد(<iEodAd)             <iEodAd/NOUN
الوثائق(AlwavA}iq)         Al/DET+wavA}iq/NOUN
المتوفرة(Almutawaf~irap)   Al/DET+mutawaf~ir/ADJ+ap/NSUFF_FEM_SG
ب (bi-)                    bi-/PREP
كثرة (-kavorap)            -kavor/NOUN+ap/NSUFF_FEM_SG
حول (Hawola)               Hawola/PREP
أول(>aw~al)                >aw~al/ADJ
رحلة (riHolap)             riHol/NOUN+ap/NSUFF_FEM_SG
طيران (TayarAn)            TayarAn/NOUN
عثمانية(EuvomAniy~ap)      EuvomAniy~/ADJ+ap/NSUFF_FEM_SG
فوق (fawoq)                fawoq/PREP
البلاد(AlbilAd)            Al/DET+bilAd/NOUN
العربية(AlEarabiy~ap)      Al/DET+Earabiy~/ADJ+ap/NSUFF_FEM_SG
```

Figure 4. Disambiguated sentence from the Arabic Treebank using the FULL tag set.

```
INPUT STRING: وَوَصَّيْنَا
LOOK-UP WORD: wwSynA
* SOLUTION 1: (wawaS~ayonA) [waS~aY_1] wa/CONJ+waS~ay/VERB_PERFECT+nA/PVSUFF_SUBJ:1P
     (GLOSS): and + recommend/advise + we <verb>
  SOLUTION 2: (wawaSiy~nA) [waSiy~_1] wa/CONJ+waSiy~/NOUN+nA/POSS_PRON_1P
     (GLOSS): and + authorized agent/trustee + our

INPUT STRING: الإِنسَانَ
LOOK-UP WORD: Al<nsAn
* SOLUTION 1: (Al<inosAn) [<inosAn_1] Al/DET+<inosAn/NOUN
     (GLOSS): the + human being +

INPUT STRING: بِوَالِدَيْهِ
LOOK-UP WORD: bwAldyh
  SOLUTION 1: (biwAlidiy~h) [wAlidiy~_1] bi/PREP+wAlidiy~/ADJ+hu/POSS_PRON_3MS
     (GLOSS): by/with + parental + its/his
* SOLUTION 2: (biwAlidayohi) [wAlid_1] bi/PREP+wAlid/NOUN+ayo/NSUFF_MASC_DU_ACCGEN+hu/POSS_PRON_3MS
     (GLOSS): by/with + parents/father and mother + his/its two

INPUT STRING: حُسْنًا
LOOK-UP WORD: HsnA
  SOLUTION 1: (Hasun~A) [Hasun-u_1] Hasun/VERB_PERFECT+nA/PVSUFF_SUBJ:1P
     (GLOSS):  + be beautiful/be good + we <verb>
  SOLUTION 2: (HasunA) [Hasun-u_1] Hasun/VERB_PERFECT+A/PVSUFF_SUBJ:3MD
     (GLOSS):  + be beautiful/be good + they (both) <verb>
  SOLUTION 3: (Has~an~A) [Has~an_1] Has~an/VERB_PERFECT+nA/PVSUFF_SUBJ:1P
     (GLOSS):  + improve/decorate + we <verb>
  SOLUTION 4: (Has~anA) [Has~an_1] Has~an/VERB_PERFECT+A/PVSUFF_SUBJ:3MD
     (GLOSS):  + improve/decorate + they (both) <verb>
* SOLUTION 5: (HusonAF) [Huson_1] Huson/NOUN+AF/NSUFF_MASC_SG_ACC_INDEF
     (GLOSS):  + good/beauty + [acc.indef.]
  SOLUTION 6: (HasanAF) [Hasan_2] Hasan/NOUN+AF/NSUFF_MASC_SG_ACC_INDEF
     (GLOSS):  + good + [acc.indef.]
  SOLUTION 7: (HasanA) [Hasan_2] Hasan/NOUN+A/NSUFF_MASC_DU_NOM_POSS
     (GLOSS):  + good + two
  SOLUTION 8: (HasanAF) [Hasan_2] Hasan/ADV+AF/NSUFF_MASC_SG_ACC_INDEF
     (GLOSS):  + well + [acc.indef.]
  SOLUTION 9: (Has~anA) [Has~-i_1] Has~/VERB_PERFECT+a/PVSUFF_SUBJ:3MS+nA/PVSUFF_DO:1P
     (GLOSS):  + feel + he/it <verb> us
  SOLUTION 10: (Has~nA) [Has~_1] Has~/NOUN+nA/POSS_PRON_1P
     (GLOSS):  + perception/feeling + our
  SOLUTION 11: (His~nA) [His~_1] His~/NOUN+nA/POSS_PRON_1P
     (GLOSS):  + sensation/perception + our
```

Figure 5. Buckwalter's morphological analysis of a sentence from the Quran.

on man kindness to parents'. Figures 3 and 5 show the full outputs of Buckwalter's morphological analyser including several possible solutions for some words. Figures 4 and 6 show the correct disambiguated solution for each word in context.

```
وَوَصَّيْنَـا (wawaS~ayonA) wa/CONJ+waS~ay/VERB_PERFECT+nA/PVSUFF_SUBJ:1P
الْإِنسَـانَ (Al<inosAn)  Al/DET+<inosAn/NOUN
بِـوَالِـدَيْـه(biwAlidayohi)bi/PREP
                          +wAlid/NOUN
                          +ayo/NSUFF_MASC_DU_ACCGEN+hu/POSS_PRON_3MS
حُسْنًـا (HusonAF)        Huson/NOUN+AF/NSUFF_MASC_SG_ACC_INDEF
```

Figure 6. Disambiguated sentence from the Quran using the FULL tag set.

			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

Figure 7. A sample of tagged sentence using the FULL, RTS and ERTS tag sets.

Diab (2007) compared the FULL and RTS tag sets introduced by the LDC to PoS-tag the Arabic Treebank. The study is about designing the optimal part-of-speech tag set for Arabic. By analyzing the Arabic Treebank data, the RTS tag set is extended from 25 tags to 75 tags. Only morphological features, which are explicitly marked on the words, are added to the RTS. The new tag set is called the ERTS (extended reduced tag set). The ERST has only the explicit or marked morphological features of gender, number and definiteness on nominals while maintaining the existing features from RTS. Figure 7 illustrates some differences between the three tag sets: FULL, RTS and ERTS from (Diab 2007).

1.3.3 ARBTAGS tag set

Alqrainy (2008) developed a new part-of-speech tag set called ARBTAGS to be used in the development of a part-of-speech tagger. The tag set design followed the criteria proposed by Atwell (2008). Like Khoja, Alqrainy built on traditional Arabic grammar books to design the tag set. Six morphological features of Arabic words were included: gender, number, case, mood, person and state. ARBTAGS contains 161 detailed tags and 28 general tags to cover the main part-of-speech classes and sub-classes. The 161 detailed tags are divided into 101 nouns, 50 verbs, 9 particles and 1 punctuation mark. Figure 8 shows the 28 general tags of the ARBTAGS tag set.

1.3.4 MorphoChallenge 2009 Qur'an gold standard part-of-speech tag set

MorphoChallenge 2009[12] Qur'an gold standard is developed using the data of Morphological Tagging of the Qur'an database (Talmon & Wintner 2003; Dror,

Tag	Description	Tag	Description
VePe	*Perfect verb*	NuCd	*Conditional noun*
VePi	*Imperfect verb*	NuDe	*Demonstrative noun*
VePm	*Imperative verb*	NuIn	*Interrogative noun*
NuPo	*Proper noun*	NuAd	*Adverb*
NuCn	*Common noun*	NuNn	*Numeral noun*
NuAj	*Adjective noun*	Fw	*Foreign noun*
NuIf	*Infinitive noun*	Pun	*Punctuation mark*
NuRe	*Relative noun*	PrPp	*Preposition*
NuDm	*Diminutive noun*	PrVo	*Vocative Particle*
NuIs	*Instrument noun*	PrCo	*Conjunction Particle*
NuPn	*Noun of Place*	PrEx	*Exception Particle*
NuTn	*Noun of Time*	PrAn	*Annulment Particle*
NuPs	*Pronoun*	PrSb	*Subjunctive Particle*
NuCv	*Conjunctive noun*	PrJs	*Jussive Particle*

Figure 8. The 28 general tags of the ARBTAGS tag set.

وَوَصَّيْنَا	وصي	يُفَعِّلُ	وَ +**Particle** +**Conjunction** وَصصَينَا +**Verb** +**Perf** +**Act** +**1P** +**Pl** +**Masc/Fem**
الْإِنسَانَ	ءنس	فِعلَان	ءِنسَان +**Noun** +**Triptotic** +**Sg** +**Masc** +**Acc** +**Def**
بِوَالِدَيْهِ	ولد	فَاعِل	ب +**Prep** وَالِد +**Noun** +**Triptotic** +**Dual** +**Masc** +**Obliquus** +**Pron** +**Dependent** +**3P** +**Sg** +**Masc**
حُسْنًا	حسن	فُعل	حُسن +**Noun** +**Triptotic** +**Sg** +**Masc** +**Acc** +**Tanwiin**

wawaS~ayonaA	wSy	yufaE~ilu	wa +**Particle** +**Conjunction** waSSaynaA +**Verb** +**Perf** +**Act** +**1P** +**Pl** +**Masc/Fem**
Alo<insaAna	'ns	fiElaAn	'insaAn +**Noun** +**Triptotic** +**Sg** +**Masc** +**Acc** +**Def**
biwaAlidayohi	wld	faAEil	b +**Prep** waAlid +**Noun** +**Triptotic** +**Dual** +**Masc** +**Obliquus** +**Pron** +**Dependent** +**3P** +**Sg** +**Masc**
HusonFA	Hsn	fuEl	Husn +**Noun** +**Triptotic** +**Sg** +**Masc** +**Acc** +**Tanwiin**

Figure 9. A sample of tagged sentence taken from the MorphoChallenge 2009 Qur'an Gold Standard, the first part uses Arabic script and the second one uses romanized letters using Tim Buckwalter's; transliteration scheme.

Shaharabani, Talmon & Wintner 2004). It was developed to be used to evaluate morphological analyzers in the Morphochallenge 2009 competition, which aims to develop an unsupervised morphological analyzer to be used for different languages including Arabic. It contains the full morphological analysis for each word, according to the Tagged database of the Qur'an but reformatted to match other Morphochallenge test sets in other languages. The word's morphological analysis is shown after each word where the morphological features are separated by space and "+" sign. These features include the part-of-speech of the word, number, gender,

person, case, definiteness, voice and others. Figure 9 shows a sample of the Qur'an gold standard.

1.3.5 The Quranic Arabic Corpus part-of-speech tag set

The Quranic Arabic Corpus is a newly available resource enriched with multiple layers of annotation including morphological segmentation and part-of-speech tagging. The motivation behind this work is to produce a resource that enables further analysis of the Qur'an; a genre difficult to compare with other forms of Arabic, since the vocabulary and the spelling differ from modern standard Arabic (Dukes & Habash 2010).

Buckwalter's Arabic Morphological Analyzer (BAMA) was used to generate the initial tagging. The analyzer was adapted to work with the Quranic Arabic text. After that, the annotated corpus was put online to allow for collaborative annotation (Dukes & Habash 2010; Dukes, Atwell & Habash 2011).

A mapping was required to convert from the BAMA tag set to the Quranic Arabic Corpus tag set. Manual disambiguation was required for a few cases, where one-to-one mapping was not applicable such as particles. In order to adapt BAMA to process the Quranic Arabic Corpus text three modifications were made. First, spelling in the Qur'an differs from MSA. The differences involve orthographic variations of *hamza*^{h}, *'alif* and the long vowel *ā*. Second, the multiple diacritized analyses produced by BAMA for the processed words were ranked in terms of their edit-distance from the Qur'anic diacritization, with closer match ranked higher. Finally, filtering was done by choosing the highest rank analysis's part of speech as a solution (Dukes & Habash 2010).

The Quranic Arabic Corpus tag set adapts historical traditional Arabic grammar, which leads to morphological annotation that uses terminology familiar to many readers of the Qur'an. This terminology enables people with Qur'anic syntax experience to participate in the online annotation to be verified against existing authenticated books on Quranic Grammar (Dukes & Habash 2010). Figure 10 shows a sample of the morphological and part-of-speech tags of the Quranic Arabic Corpus.

1.3.6 Columbia Arabic Treebank CATiB part-of-speech tag set

Another tag set was designed for the part-of-speech and syntactic annotation in the Columbia Arabic Treebank (CATiB). A part-of-speech tag set consisting of only six tags is used for the part-of-speech annotation of CATiB. The main reason for using such a small tag set is a tradeoff between linguistic richness and Treebank size. The researchers' assumption for morpho-syntactically rich languages such as Arabic, is that the cost of fine-grain annotation is a slower annotation process, a smaller Treebank and less data to train tools. CATiB is inspired by two ideas. First, it avoids annotation of redundant linguistic information. Second, it uses linguistic representation and terminology from traditional Arabic syntactic studies

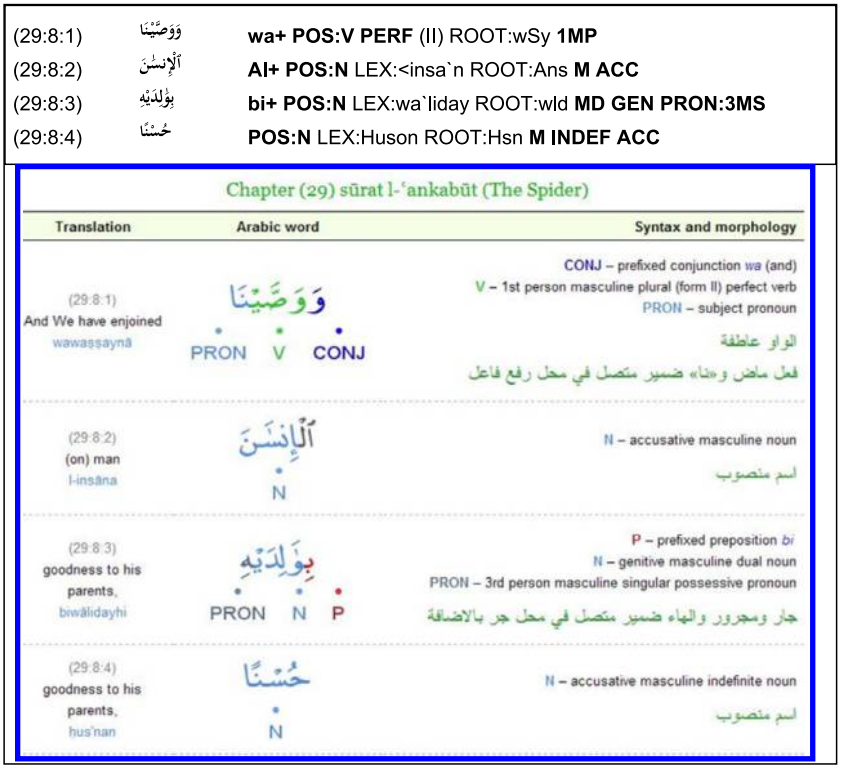

Figure 10. A sample of a tagged sentence taken from the Quranic Arabic Corpus.

(Habash, Faraj & Roth 2009). The tag set is much smaller than the FULL tag set used by the Penn Arabic Treebank:

> [. . .] CATiB uses the same tokenization scheme used by PATB and PADT. However, unlike these resources, the CATiB POS tag set is much smaller. Whereas PATB uses 2,200 tags specifying every aspect of Arabic word morphology such as definiteness, gender, number, person, mood, voice and case; CATiB uses six POS tags: NOM (nominals such as nouns, pronouns, adjectives and adverbs), PROP (proper noun), VRB (verb), VRB-PASS (passive verb), PRT (particles such as prepositions or conjunctions) and PNX (punctuation). (Habash & Roth 2009: 2)

Figure 11 shows an example of the sentence, خمسون ألف سائح زاروا لبنان وسوريا في أيلول الماضي *ẖamsūn 'alf sā'iḥ zārū lubnān wa sūriyyā fī 'aylūl al-māḍī* '50 thousand tourists visited Lebanon and Syria last September', tagged using part-of-speech tags used in the CATiB (Habash & Roth 2009).

Word			CATiB part-of-speech tag	CATiB annotation
خمسون	*ẖamsūn*	Fifty	**NOM**	
ألف	*'alf*	Thousand	**NOM**	
سائح	*sā'iḥ*	Tourist	**NOM**	
زاروا	*zārū*	Visited	**VRB**	
لبنان	*lubnān*	Lebanon	**PROP**	
و	*wa*	And	**PRT**	
سوريا	*sūriyyā*	Syria	**PROP**	
في	*fī*	In	**PRT**	
أيلول	*'aylūl*	September	**NOM**	
الماضي	*al-māḍī*	Past	**NOM**	

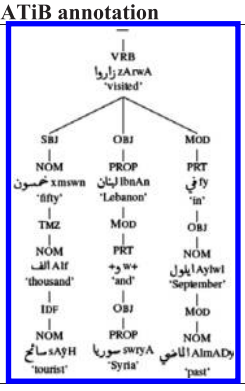

Figure 11. Example of part-of-speech tagged sentence using CATiB tag set.

1.3.7 Comparison of Arabic part-of-speech tag sets

Table 1 shows a comparison of the seven Arabic tag sets studied in this section. The comparison summarizes the characteristics of each tag set and helps to show the differences between them clearly. The drawbacks of the existing tag sets for Arabic were found to be:

- Existing Arabic tag sets vary in size from 6 tags to 2,000 or more tags.
- Some of these tag sets follow standards for tag set design for English such as the PATB tag sets, and these may not always be appropriate for Arabic.
- The tag sets share common morphological features such as gender, number, person, case, mood and definiteness, but the attributes of the morphological feature categories are not standardized.
- These tag sets lack standardization in defining a suitable scheme for tokenizing Arabic words into their morphemes and they mix morpheme tagging with whole word tagging.
- They also lack suitable documentation that illustrates the decision made for each design dimension of the tag set.
- The tags assigned to words in a corpus are not consistent in either presentation of the tag itself or the morphological features which are encoded within the tag.

Moreover, the most widely used and important morphosyntactic annotation standards and guidelines, namely EAGLES (see section 2), are designed for Indo-European languages. These guidelines are not entirely suitable for Arabic. These drawbacks of existing tag sets are the motivation behind the SALMA (Sawalha Atwell Leeds Morphological Analysis) Tag Set for Arabic.

Table 1. Comparison of Arabic part-of-speech tag sets.

Khoja's tag set	
Purpose of design	Compiling a tag set as a standard tag set.
Main characteristics	Based on traditional Arabic grammar rather than being based on an Indo-European one. Only the main classes and subclasses have been chosen.
Tag set size	177 tags (103 types of noun, 57 verbs, 9 particles, 7 residuals, 1 punctuation mark)
Morphological features	Gender, Number, Case, Definiteness, Person, Mood
Applications	Used in the design of the APT tagger, and in the annotation of the training data of the APT tagger.
Penn Arabic Treebank (PATB) Part-of-Speech Tag Set (FULL)	
Purpose of design	Annotating the Arabic Treebank with part-of-speech tags.
Main characteristics	Aims to cover detailed grammar features.
Tag set size	The FULL tag set comprises over 2,000 tag types. This includes combinations of 114 basic tags.
Morphological features	Case, Gender, Number, Definiteness, Mood, Person, Voice, Tense, Aspect
Applications	Used in Tim Buckwalter's morphological analyser to annotate the Penn Arabic Treebank with part-of-speech tags.
Penn Arabic Treebank (PATB) Reduced Part-of-Speech Tag Set (RTS)	
Purpose of design	Maximizing the performance of Arabic syntactic parsing.
Main characteristics	Follows the tag set designed for the English Wall Street Journal.
Tag set size	25 tags
Morphological features	Case, Mood, Gender, Person, Definiteness
Applications	Used in the syntactic annotation of the Penn Arabic Treebank
Penn Arabic Treebank (PATB) Extended Reduced Part-of-Speech Tag Set (ERTS)	
Purpose of design	To be used for higher order processing of the language
Main characteristics	Is an extension of the RTS tag set, which has only the explicit or marked morphological features of gender, number and definiteness on nominals.
Tag set size	75 tags
Morphological features	Gender, Number, Definiteness on nominals
Applications	To be used for parsing.
ARABTAGS	
Purpose of design	Standardizing and building a comprehensive Arabic tag set.
Main characteristics	The tag set hierarchy follows the tradition of Arabic grammar.

Tag set size	161 detailed tags (101 nouns, 50 verbs, 9 particles, 1 punctuation mark including 28 different POS general tags to cover the main part-of-speech classes and sub-classes.
Morphological features	Gender, Number, Case, Mood, Person, State
Applications	Used in the Arabic Morphosyntactic Tagger AMT
MorphoChallenge 2009 Qur'an gold standard tag set	
Purpose of design	To annotate the Qur'an gold standard to be used to evaluate morphological analyzers in the MorphoChallenge 2009 competition.
Main characteristics	It was developed using the data for Morphological Tagging of the Qur'an database.
Tag set size	The tag set involves combinations of the POS main and sub-classes and the morphological features of the analysed words.
Morphological features	Gender, Number, Person, Case, Mood, Aspect, Voice, Definiteness, Diptotic
Applications	Used to construct the Qur'an gold standard for evaluating morphological analyzers in the MorphoChallenge 2009 competition.
Quranic Arabic Corpus POS tag set	
Purpose of design	To annotate the Qur'an by morphological and part-of-speech tagging information.
Main characteristics	Used Tim Buckwalter's morphological analyzer as initial tagging, then a mapping from Buckwalter's tag set to the Quranic Arabic Corpus tag set. It adapts traditional Arabic grammar.
Tag set size	The tag set involves combinations of the POS main and sub classes and the morphological features of the analysed words.
Morphological features	Person, Gender, Number, Aspect, Mood, Voice, Verb form, Derivation, State
Applications	Used in the morphological and part-of-speech annotation of the Quranic Arabic Corpus.
Columbia Arabic Treebank POS tag set	
Purpose of design	To be used for the part-of-speech annotation of Columbia Arabic Treebank CATiB.
Main characteristics	CATiB avoids the annotation of redundant linguistic information that is determinable automatically from syntax and morphological analysis, e.g., nominal case. CATiB uses linguistic representation and terminology inspired by the long tradition of Arabic syntactic studies.
Tag set size	6 part-of-speech tags (**VRB** – all verbs, **VRB-PASS** – passive-voice verbs, **NOM** – all nominals, **PROP** – proper nouns, **PRT** – particles, **PNX** – all punctuation marks)
Morphological features	No morphological features are encoded in the part-of-speech tag set of Columbia Arabic Treebank CATiB.
Applications	Used in the part-of-speech annotation of Columbia Arabic Treebank CATiB.

2. Morphological features in tag set design criteria

EAGLES[13] proposed recommendations (guidelines) for morphosyntactic categories for European languages. The aim of the EAGLES guidelines is to propose standards in developing tag sets for morphosyntactic tagging, in the interest of comparability, interchangeability and reusability of annotated corpora. In addition to preferred standards, EAGLES guidelines also cater for extensibility, allowing specifications to extend to language-specific phenomena. The guidelines proposed standardisation in three important areas:

1- Representation/Encoding: transparency, processability, brevity and unambiguity.
2- Identifying categories/subcategories/structure: agreement on common categories and allowance for variation (obligatory, recommended and optional specification).
3- Annotation schemes and their application to text: detailed annotation schemes should be made available to end-users and to annotators.

EAGLES recognizes four degrees of constraint in the description of word categories for morphosyntactic tags. First, *obligatory*: attributes have to be included in any morphosyntactic tag set (main categories of part-of-speech Noun, Verb, Adjective, Pronoun/Determiner, Article, Adverb, Adposition, Conjunction, Interjection, Unique/Unassigned, Residual, Punctuation). Second, *recommended*: attributes and values of widely-recognized grammatical categories which occur in conventional grammatical description (*e.g.* Gender, Number, Person, etc.). Third, *generic special extensions*: attributes and values which are not usually encoded, but might be included for particular purposes, for example semantic classes such as temporal nouns, manner adverbs, place names, etc. Finally, *language-specific special extensions*: additional attributes or values which may be important for a particular language.

Khoja et al. (2001) compared their Arabic tag set against the EAGLES guidelines. The comparison showed: first, EAGLES tag set guidelines are based on Latin as a common ancestor, while Arabic has some novel features not found in Latin, for example certain categories and subcategories that inherit properties from the parent categories. Second, a classical Arabic tag set has three main categories (nouns, verbs and particles), while EAGLES has eleven major part-of-speech categories. Third, apart from nouns and verbs, other major categories in EAGLES such as pronouns, numerals and adjectives are described as subcategories of major categories in a Classical Arabic tag set. Fourth, Arabic, not only has singular and plural numbers, but it also has dual number. Moreover, Arabic verbs are classified as being perfect, imperfect and imperative, which differs from EAGLES classification of past, present and future tenses. Finally, the mood morphological feature is not covered by the EAGLES guidelines.

Atwell (2008) proposed criteria for tag set development, and stated that there are dimensions (choices) to be made by developers of a new part-of-speech tag set.

Developers must decide on the set of grammatical tags or categories, and their definitions and boundaries. These criteria were applied to Arabic when the ARABTAGS tag set (Alqrainy 2008) was designed. We followed the same criteria as Atwell (2008) in designing the general-purpose morphological features tag set. Sections 2.1–2.12 explain the criteria and how they are applied in the SALMA – Tag set.

2.1 Mnemonic tag names

Generally, tag names for English PoS tag sets are chosen to help linguists to remember the grammatical categories such as **CC** for ***C****oordinating* ***C****onjunction* and **VB** for ***VerB***. The SALMA Tag Set for Arabic has to encode much richer morphology: the tag is represented by a string of 22 characters. Each character represents a value or attribute which belongs to a morphological feature category. The position of the character in the tag string is important as it identifies the morphological feature category. The value of the feature is represented by one lower case character, which is intended to remain readable, such as: **v** in the first position to indicate *verb*, **n** in the second position to indicate ***name***, gender category values in the seventh position where ***m****asculine* is represented by **m**, *feminine* is represented by **f** and *common gender* is represented by **x**. If the value of a certain feature is not applicable for the tagged word then dash '-' is used to indicate this. A question mark '?' indicates 'unknown': a certain feature normally belongs to the word but at the moment is not available or the automatic tagger could not guess it.

The interpretation of the tag is handled by referring to the attribute value and its position in the tag string. The position of the attribute in the tag string identifies the morphological feature category, while the attribute value is identified by searching the morphological feature category for the specified symbol. Then, all these single interpretations of attributes are grouped together to represent the full tag of the word. The tag is intended to remain readable by linguists. Moreover, the tag is straightforwardly readable by software, for example by a search tool matching specified feature-value(s).

2.2 Underlying linguistic theory

Linguists who develop new tag sets will inevitably be swayed by the linguistic theories they espouse. In the case of English, there is disagreement between grammar theories on the range of grammatical categories and features to be tagged, and more complicated structural issues. It is difficult to have theory-neutral annotation, because every tagging scheme makes some theoretical assumptions (Atwell 2008).

Khoja's mophosyntactic tag set was derived from classical Arabic grammar (Khoja et al. 2001; Khoja 2003). ARBTAGS also tried to follow the Arabic grammatical system, which is based upon main three part-of-speech classes: verbs, nouns and particles, and enriched with inflectional features (Alqrainy 2008). The Arabic Penn Treebank tag set follows the same criteria used to develop the English Treebank

(Maamouri & Bies 2004). ERTS (extended reduced tag set) extends the LDC reduced tag set (RTS) by adding morphological features namely (case, mood, definiteness, gender, number and person). This extends the 25 RTS tag set to 75 tag set of ERTS (Diab 2007).

The proposed SALMA Tag Set adds more fine-grained details to the existing tag sets. The tag set follows traditional Arabic grammar theory (Dahdah 1987; Dahdah 1993; Wright 1996; Al-Ghalayyni 2005; Ryding 2005) in specifying 22 morphological features categories and their attributes or values. Section 4 justifies the SALMA Tags in terms of this underlying theory.

2.3 Classification by form or function

For English an ambiguous word like '*open*' is tagged according to its function, and only its inflected forms are tagged by their form. Arabic words are highly inflected and hence word classification tends to be dependent on form. Classification by form is dependent on the word, while classification by function is dependent on the function of the word in context. For Arabic, the word class is heavily constrained by form, but if there is only one analysis, then it is determined by function. If there are two analyses, one needs to take context into account which means it is partially determined by function. In this case the function has to be taken into account for classification.

Arabic word-class is dependent on form. Traditional Arabic grammar groups words according to their inflexional behaviour. A challenging characteristic of Arabic is the treatment of short vowels, which are normally omitted in written Arabic. These short vowels can help in specifying some morphological feature information of grammatical categories. The Qur'an is fully vowelized to ensure it is pronounced correctly. This makes the Qur'an a potential 'Gold Standard' corpus for Arabic tagging and NLP research (Atwell 2008).

Another challenge of Arabic words can appear when classifying words according to certain morphological features such as gender. Classifying nouns into masculine or feminine can be viewed from two perspectives. First, according to the word's structure or morphologically; masculine singular nouns are not normally marked by any suffix, while feminine nouns have a suffix – normally $-a^h$ – added at the end of the noun. Second, semantically; nouns are arbitrary classified into masculine or feminine, except when a noun refers to a human being or other creature having natural gender (sex), when it is normally conforms to natural gender (Ryding 2005). On rare occasions a noun has the 'morphological' feminine suffix $-a^h$, but indicates a male and is therefore masculine in gender, for example حَمْزَة *ḥamza*^{h} 'Hamza (male proper name)'. Conversely, a few nouns which are feminine in gender do not have the 'morphological' feminine suffix $-a^h$, an example being مَرْيَم *maryam* 'Mary (female proper name).

2.4 Idiosyncratic words

Arabic has some words with special, idiosyncratic behaviour, such as particles which cannot be analysed morphologically according to a root and a pattern. Khoja, Garside

et al (2001) includes examples of this type in an 'Exception' category, which covers group of particles that are equivalent to the English word 'except' and the prefixes *non-*, *un-* , and *im-*.

2.5 Categorization problems

A detailed categorisation scheme requires each tag to be defined clearly and unambiguously, by giving examples in a 'case-law' document. This definition should include how to decide difficult, borderline cases, so that all examples in the corpus can be tagged consistently. Many words can belong to more than one grammatical category, depending on context of use. Tagging schemes should specify how to choose one tag as appropriate, if a word can have different part-of-speech tags in different contexts (Atwell 2008).

Vowelized Arabic text has less ambiguity than non-vowelized Arabic text. Short vowels and some affixes add linguistic information, which reduces the ambiguity. In the SALMA Tag Set, each feature category is described, clearly documented and examples are provided. Moreover, tagging guidelines define the appropriate attribute for the morphological feature category.

2.6 Tokenisation: what counts as a word?

Arabic text tokenisation is not an easy task. Simple tokenisation of text can be carried out by dividing text into words by spaces, or punctuation. This tokenisation process is primitive and the first step in tokenising Arabic text. The majority of Arabic words are complex words; one or more clitics can be attached to the beginning and the end of the word [clitic(s) + word + clitic(s)]. These clitics are particles, pronouns or the definite article. A tag is provided for each clitic attached to a word along with the tag of the word. For instance, the word وَبِحَسَنَاتِهِم *wabiḥasanātihim* 'and with their good deeds', consists of four parts, the conjunction وَ *wa* 'and', the preposition بِ *bi* 'with', the word حَسَنَاتِ *ḥasanāti* 'good deeds' and the pronoun هِم *him* ' their'. The tag of this word will be the tags of the four elements and the whole word tag which is a combination of the morpheme tags. The clitics will help the tagging scheme in identifying some of the morphological attributes; the preposition بِ *bi* governs the genitive case of the noun.

2.7 Multi-word lexical items

Multi-words lexical items are rare in Arabic (Alqrainy 2008). Such items might consist of two words; noun followed by adjective describing the preceding noun, some compound proper names such as عَبْدُ الله *'abdu allāh* 'Abdullah', or compound particles such as فِيْمَا *fīmā* which consists of the preposition فِي *fī* and the non-human relative noun مَا *mā*. In the case of proper names a single tag might be more appropriate, while, for the other cases, a separate tag for each part of the lexical item will give more morphological detail about the multi-word lexical items.

The Penn Arabic Treebank guidelines ignore multi-word lexical items and tag each word of a compound word separately:

> [. . .] Divided/compound proper names in Arabic (Abdul Ahmed, e.g.): Label all parts of the name with the 'Is a name' button.
>
> Idioms: (for example, in what in them = 'included'): Label each word independently for its own part of speech (ignore the idiomatic meaning).[14]

2.8 Target users and/or applications

Fitness for purpose and customer satisfaction are the most important practical criteria for a new tag set. One common use of part-of-speech tagged corpora is language teaching and research. A detailed tag set is required in teaching and learning to reflect fine distinctions of grammar, even though Machine Learning systems could cope better with a smaller tag set. General-purpose tag set developers should be more aware of potential re-use: detailed and more sophisticated part-of-speech tag schemes allow wider re-use of the corpus in future research (Atwell 2008).

The SALMA Tag Set is a general-purpose tag set. It encodes detailed information of morphological features embedded in any word. This morphological features information enables the tag set to be widely re-used.

2.9 Availability and/or adaptability of tagger software

If a part-of-speech tag set is implemented in automatic tagger software, this has a clear advantage over a purely theoretical tag set (Atwell 2008). HMM taggers can be re-used for any language including Arabic. Experiments on highly inflectional languages such as German and Czech using an HMM tagger with a fine-grain tag set achieved higher tagging accuracy than two state-of-the-art general purpose part-of-speech taggers, The TnT tagger and SVMTool (Schmid & Laws 2008). Another experiment that uses a fine-grain tag set was done for Latin. Latin words require morphological analysis of nine features: part-of-speech, person, number, tense, mood, voice, gender, case and degree. The experiment used the TreeTagger analyzer, which achieved an accuracy of 83% in correctly disambiguating the full morphological analysis (Bamman & Crane 2008).

2.10 Adherence to standards

The EAGLES guidelines are designed for European languages. However, the Arabic language is different from Indo-European languages and has its own structure and morphological features. Instead, the standard adhered to in the SALMA Tag Set is that of traditional Arabic grammar books e.g. (Dahdah 1987; Dahdah 1993; Wright 1996; Al-Ghalayyni 2005; Ryding 2005).

2.11 Genre, register or type of language

The SALMA Tag Set is intended to be general-purpose and to be used in part-of-speech tagging of different text types, formats and genres, of both vowelized and non-vowelized text. We plan to evaluate the tagging schemes and the tag set on variety of text types, formats and genres. Corpora can include text in classical Arabic such as the Qur'an, Classical Arabic dictionaries and poems from ancient Arabic literature, as well as Modern Standard Arabic text from newspapers, magazines, web pages, blogs, children's books, school text books, etc.

2.12 Degree of delicacy of the tag set

The total number of tags is an indicator of the level of fine-grainedness of analysis. Existing Arabic corpus tag sets have a degree of delicacy ranging from 25 for the RTS tag set of the Penn Arabic Treebank, to 75 tags for ERTS, 161 tags for ARABTAGS, and 177 tags for Khoja's tag set. The SALMA Tag Set is a fine-grain tag set. It is unfeasible to enumerate all possible tags that can be generated from valid combinations of the 22 morphological feature categories; however, we can count the attributes of each feature category, and use these to estimate an upper bound or limit on the degree of delicacy of the SALMA Tag Set. Section 4 discusses four selected examples of the 22 morphological features of the SALMA Tag Set and their attributes.

An upper limit of possible feature combinations is 4.07E+16, the total number of possible combinations of features in the SALMA Tag Set of Arabic, calculated by multiplying together the number of attributes of each of the 22 morphological features. But, of course, this includes many invalid tags that will never be used. A more realistic upper bound is given by counting the possible feature combinations for each major part of speech, and summing these. Table 2 shows the absolute upper limit of possible feature combinations for each major part of speech (Noun, Verb, Particle, Other (Residual), and Punctuation); this gives an upper limit of 101,945,168 possible morphological feature combinations: about one hundred million possible SALMA tags.

3. The Complex morphology of Arabic

Most Arabic words are derived from their roots following certain templates called patterns. The derivation process adds prefixes, suffixes and infixes to the root letters to generate a new word, which has a new function or meaning but preserves the main concept or meaning carried by the root. Moreover, using the derived word in a certain context will require clitics to be added to the beginning and the end of the word. Proclitics include prepositions, conjunctions and definite articles, and enclitics include pronouns. In addition, one or more affixes or clitics can be added to the derived word. In conclusion, most Arabic words are complex words consisting of multiple morphemes.

Table 2. The upper limit of possible combinations of SALMA features.

Feature		Number of attributes	Part of speech									
			Noun		Verb		Particle		Other		Punctuation	
			Template	Combinations	Template	Combinations	Template	Combinations	Template	Combinations	Template	Combinations
1	Main Part-of-Speech	5	n	1	v	1	p	1	r	1	u	1
2	Part-of-Speech: Noun	34	?	34	-	1	-	1	-	1	-	1
3	Part-of-Speech: Verb	3	-	1	?	3	-	1	-	1	-	1
4	Part-of-Speech: Particle	22	-	1	-	1	?	22	-	1	-	1
5	Part-of-Speech: Other (Residual)	15	-	1	-	1	-	1	?	15	-	1
6	Punctuation marks	12	-	1	-	1	-	1	-	1	?	12
7	Gender	3	?	3	-	1	-	1	?	3	-	1
8	Number	9	?	9	-	1	-	1	?	3	-	1
9	Person	3	-	1	?	3	-	1	?	3	-	1
10	Inflectional morphology	4	?	3	?	2	?	1	?	1	-	1
11	Case or Mood	4	?	3	?	3	-	1	-	1	-	1
12	Case and Mood marks	10	?	7	?	6	?	4	?	4	-	1
13	Definiteness	2	?	2	-	1	-	1	-	1	-	1
14	Voice	2	-	1	?	2	-	1	-	1	-	1
15	Emphasized and non-emphasized	2	-	1	?	2	-	1	-	1	-	1
16	Transitivity	4	-	1	?	4	-	1	-	1	-	1
17	Rational	2	?	2	?	2	?	2	-	1	-	1
18	Declension and Conjugation	9	?	4	?	6	?	1	-	1	-	1
19	Unaugmented and Augmented	5	?	5	?	5	-	1	-	1	-	1
20	Number of root letters	3	?	3	?	2	-	1	-	1	-	1
21	Verb root	30	-	1	?	30	-	1	-	1	-	1
22	Noun finals	6	?	6	-	1	-	1	-	1	-	1
Total		**4.1E+16**	**83,280,960**		**18,662,400**		**176**		**1620**		**12**	
Upper limit of possible morphological feature combinations											**101,945,168**	

To specify a word's morphemes, tokenization is needed to analyse the word morphemes as clitics, affixes or stem. For example the tokenizer will specify the morphemes of the word وسيكتبونها *wasayaktubūnahā* 'and they will write it' as follows: preclitic و *wa* 'and' (conjunction), prefixes س *sa* 'will' and ي *ya* (imperfect prefix), the

Analyzed sentence: أقمت بمدينتي الجديدة لمدة عامين *'aqamtu bimadīnatī al-ğadīdat limuddat 'āmayn* "I have stayed in my new city for two years"

Analyzed word: بمدينتي *bimadīnatī* in my city

Step 1 : Tokenization of words into morphemes

Word	Proclitics	prefixes	Stem	Suffixes	enclitics
بمدينتي	ب ***bi*** in	-------	مدين ***madīna*** city	ت (ة) *t* feminine *tā'*	ي *ī* my

Step 2 : Assign morpheme tags

Morpheme	*Tag*	*Description*
ب *bi* in	`p--p------------------`	Particle; Preposition
مدين *madīna* city	`nl-------vg?i----tat-s`	Noun; Noun of place; Varied; Genitive; Indefinite; Primitive/ Concrete noun; Augmented by one letter; Triliteral root; Sound noun.
ت *t* feminine *tā'*	`r---f-fs-s-k----------`	Other (Residual); *tā'* of femininization; feminine; Singular; Invariable; *kasra*^{h};
ي *ī* my	`r---r-msfsgs----------`	Other (Residual); Connected pronoun; Common gender; Singular; First person; Invariable; Genitive; *sukūn* (Silence)

Step 3: Assign word tag

Word	*Tag*	*Description*
بمدينتي ***bimadīnatī***	`nl----fs-vgki----tat-s`	Noun; Noun of place; feminine; Singular; Declined; Genitive; *kasra*^{h}; Indefinite; Primitive/ Concrete noun; Augmented by one letter; Triliteral root; Sound noun.

Figure 12. Example of tokenization, the SALMA tag assignment for separate morphemes and the combination of the morpheme tags into the word tag.

stem (*i.e.* lemma) كتب *kataba* 'write' , the suffix ون *ūn* 'they' and the enclitic ها *hā* 'it' (object suffixed pronoun). The word consists of 6 morphemes. Each morpheme carries morphological features and belongs to a specific part of speech category. Our SALMA Tag Set assigns a tag to each morpheme of the word. Then the morphemes' tags are combined into one word tag. The word tag inherits its morphological feature attributes using an algorithm that establishes agreements on morphological feature attributes. The description of the algorithm is beyond the scope of this paper. This paper is about the output of the tagger rather than describing the algorithm of tagging and combining morpheme tags into word tags. The following example in figure 12 shows the tokenization of the word into morphemes, the assignment of the part of speech tag for each morpheme and the result of combining the morpheme tags into one whole word tag. Tokenization is a well-known problem even for English corpus tagging.

The tagged LOB corpus defines the word or graphic word as a sequence of characters surrounded by spaces (or punctuation marks). Each word is assigned a tag. Differences in tagging occur due to: 1. variation in segmentation of compound terms, as in: *fancy free* given the tags **NN** (noun, singular, common) **JJ** (adjective), and *fancy-free* given the tag **JJ** (adjective); 2. hyphenated sequences, as in: *an above-the-rooftops position* given the tag **JJB** (adjective, attributive-only); 3. syntactic boundaries, as in: *Henry* **NP** (noun, singular, proper) *8's* **CD$** (numeral, cardinal, genitive) *hall*. In some cases, the LOB Corpus tagging guidelines have changed from 'one-word-one-tag-approach' to idiom tagging to handle the cases of recurrent multiword sequences functioning as units (Johansson, Atwell, Garside & Leech 1986).

On the other hand, contractions forming regular patterns such as, *I'll, she's, John's, let's, d'you*, etc. are split up in the tagged LOB corpus as the following: *I' ll*, *she' s, John' s, let' s, d' you*. Each part is treated as a separate word and assigned a single tag. Except where *'s* is possessive suffix, then the word gets a single tag entry **$** *e.g. John's* gets the tag **NP$** (Johansson et al. 1986).

4. The standard tag set expounding morphological features

The SALMA tag set is a general-purpose fine-grained tag set. It is intended that this tag-set will be used by part-of-speech tagging software to annotate corpora with detailed morphological information for each word, and to enable direct comparisons between tagging algorithms and taggers using the same tag set. The tag set has been designed by grouping 22 morphological feature categories in one tag. Most of these morphological categories are described in any traditional Arabic language grammar book. In our study, all the morphological features are attested in five well-known traditional Arabic grammar books (Dahdah 1987; Dahdah 1993; Wright 1996; Al-Ghalayyni 2005; Ryding 2005). Table 3 shows the 22 morphological feature categories.

The tag string consists of 22 characters. Each character represents a value or attribute which belongs to a morphological feature category. The position of the character in the tag string is important to identify the morphological feature category. Each morphological feature category attribute is represented by one lower case letter, which is still human-readable, such as **v** in the first position to indicate *verb*, **n** in the second position to indicate *name*, gender category values in the seventh position: *masculine* represented by **m**, *feminine* represented by **f** and *common gender* represented by **x**. If the value of a certain feature is not applicable for the word, then a dash '-' is used to indicate this; *e.g.* the mood morphological feature is not a noun feature. In contrast, a question mark '**?**' means a certain feature belongs to a word but, at the moment, the feature value is not available or the automatic tagger could not guess it.

The tag is intended to remain readable by linguists. Moreover, it can be rendered more readable if the interpretation of the tag string features is generated automatically: software can convert each position + letter to a human-readable English and/or Arabic

Table 3. Arabic Morphological Feature Categories.

Position	Morphological Features Categories		
1	Main Part-of-Speech	أقسام الكلام الرئيسيَّة	*'aqsām al-kalām ar-ra'īsiyya^{t}*
2	Part-of-Speech: Noun	أقسام الكلام الفرعيَّة (الاسم)	*'aqsām al-kalām al-far'iyya^{t} (al-'ism)*
3	Part-of-Speech: Verb	أقسام الكلام الفرعيَّة (الفعل)	*'aqsām al-kalām al-far'iyya^{t} (al-fi'l)*
4	Part-of-Speech: Particle	أقسام الكلام الفرعيَّة (الحرف)	*'aqsām al-kalām al-far'iyya^{t} (al-ḥarf)*
5	Part-of-Speech: Other (Residual)	أقسام الكلام الفرعيَّة (أخرى)	*'aqsām al-kalām al-far'iyya^{t} ('uḫrā)*
6	Punctuation marks	أقسام الكلام الفرعيَّة (علامات الترقيم)	*'aqsām al-kalām al-far'iyya^{t} ('alāmāt at-tarqīm)*
7	Gender	المُذَكَّر والمُؤَنَّث	*al-muḏakkar wa al-mu'annaṯ*
8	Number	العدد	*al-'adad*
9	Person	الاسناد	*al-'isnād*
10	Inflectional morphology	الصَّرف	*aṣ-ṣarf*
11	Case or Mood	الحالة الإعرابية للاسم أو الفعل	*al-ḥāla^{tu} al-'i'rābiyya^{tu} lil-'ism 'aw al-fi'l*
12	Case and Mood marks	علامة الإعراب أو البناء	*'alāmāt al-'i'rāb wa al-binā'*
13	Definiteness	المَعْرِفَة والنَّكِرَة	*al-ma'rifa^{ti} wa an-nakira^{ti}*
14	Voice	المَبْني لِلمَعْلُوم و المَبْني لِلمَجْهُول	*al-mabnī lil-ma'lūm wa al-mabnī lil-maǧhūl*
15	Emphasized and non-emphasized	المُؤَكَّد وغيرُ المُؤَكَّد	*al-mu'akkad wa ḡayr al-mu'akkad*
16	Transitivity	اللازم والمتعدي	*al-lāzim wa al-muta'addi*
17	Rational	العاقل وغير العاقل	*al-'āqil wa ḡayr al-'āqil*
18	Declension and Conjugation	التَّصريف	*at-taṣrīf*
19	Unaugmented and Augmented	المجرَّد والمزيد	*al-muǧarrad wa al-mazīd*
20	Number of root letters	عَدَد أحْرُف الجَذْر	*'adad 'aḥruf al-ǧaḏr*
21	Verb root	بُنية الفعل	*bunya^{tu} al-fi'l*
22	Noun finals	أقسام الأسم تبعاً للفظ آخره	*'aqsām al-'ismi tib'^{an} li-lafẓi 'āḫirhi*

grammar term. Figures 13 and 14 show samples from the Penn Arabic Treebank and the Qur'an (the same sentences from section 1.4), tagged using the SALMA Tag Set.

The categories and features are drawn from traditional Arabic grammar books (Dahdah 1987; Dahdah 1993; Wright 1996; Al-Ghalayyni 2005; Ryding 2005). In most cases there is agreement among them, but in some cases there are discrepancies. When there is agreement, the approach taken is simply a matter of presenting the agreed features. When there is a discrepancy in most cases the difference is that one text has more fine-grained subcategories which are merged in other texts; so the more fine-grained wider sub-classification is adopted. The only significant disagreement is in the

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----------

Figure 13. Sample of tagged vowelized Qur'an text using the SALMA Tag Set.

number of nouns; see section 4.2, and in that case we adopted the widest most fine-grained sub-classification system.

Arabic grammar terms used to describe the attributes of the morphological feature categories in the SALMA – Tag Set are the same terms used by traditional Arabic grammar. The equivalent English translations of these grammar terms were extracted from 4 well-known traditional Arabic grammar reference books written in English. These books are: (Wright 1996), (Ryding 2005), (Dahdah 1993) and (Cachia 1973). These reference books agree on translating general Arabic grammar terms such as, noun, verb, adjective, person, number, case and mood. However, these reference books do not agree on translating some fine-grained attribute names such as الفعل السالم *al-fi'l as-sālim*, which is translated into 'the strong verb' by Wright (1996), 'regular (sound) root' by Ryding (2005), 'intact verb' by Dahdah (1993), and 'sound verb; strong verb; verbum firmum' by Cachia (1973). The agreed English translations of the grammar terms were directly used. For the non-agreed English translation, Professor James Dickins (head of Arabic and Middle Eastern Studies, University of Leeds, UK) was consulted to give advice on those English translations of Arabic grammar terms that would be clearest to English speaking linguists.

Appendix A lists the morphological features categories and their attribute values at each position of the 22 positions of the tag string.

The following sections 4.1 to 4.4 describe four morphological categories selected to show examples of the detailed descriptions of the morphological categories and their attributes. The first selected category is the main part-of-speech. The second category is the part-of-speech subcategories of Noun; representing a detailed example of the subcategories of the main part-of-speech. Gender is selected to show an example of the morphological features of Arabic words. Finally, the morphological feature of Augmented and Unaugmented is selected to as an example of the word's internal structure features. The complete description of the 22 morphological features can be found in the annotation manual[15] of the morphological features tag set of Arabic. The complete description also appears in Sawalha (2011).

Word		Morpheme		Tag
tamma Accomplished	تم	تم	*tamma* Accomplished	v-p---msts-f-amihdstb-
'i'dādu Preparing	اعداد	اعداد	*'i'dādu* Preparing	ng----ms-vndi---?db3-s
al-waṯā'iqa Documents	الوثائق	ال	*al* The	r---d----------
		وثائق	*waṯā'iqa* Documents	nq----fb-vafd---ndbt-s
al-mutawaffira^{ti} Available	المتوفرة	ال	*al* The	r---d----------
		متوفر	*mutawaffira* Available	nj----fs-vafd---ndtt-s
		ة	^{ti}	r---t-fs----------
bi kaṯra^{tin} In Many	بكثرة	ب	*bi* In	p--p----------
		كثر	*kaṯra* Many	nj----fb-vgki----dat-s
		ة	^{tin}	r --t-fs----------
ḥawla About	حول	حول	*ḥawla* About	nv----m--s-fi----nst-s
'awwali First	أول	أول	*'awwali* First	n+----ms-vgki----dst-s
riḥla^{ti} Trip	رحلة	رحل	*riḥla* Trip	no----fs-vgki----dat-s
		ة	^{ti}	r---t-fs----------
tayyarānin Flight	طيران	طيران	*tayyarānin* Flight	ng----ms-vgki----dbt-s
uṯmāniyya^{t} Ottomani	عثمانية	عثمان	*uṯmān* Ottoman	n*----fs-pgki----daq-s
		ي	*iyya*	r---y----------
		ة	^{t} tā' marbūṭa^{h}	r---t-fs----------
fawqa Over	فوق	فوق	*fawqa* Over	nv----m--s-fi----nst-s
al-bilādi Countries	البلاد	ال	*al* The	r---d----------
		بلاد	*bilād* Countries	nl----mb-vgkd---ndat-s
al-'arabiyyati Arabian	العربية	ال	*al* The	r---d----------
		عرب	*'arab* Arab	n*----fb-vgkd---hdst-s
		ي	*iyya*	r---y----------
		ة	^{ti} tā' marbūṭa^{h}	r---t-fs----------

Figure 14. Sample of tagged non-vowelized newspaper text using the SALMA Tag Set.

4.1 Main part of speech categories

Generally, there is agreement among existing Arabic tag sets on the classification of main part-of-speech categories in traditional Arabic grammar books (*e.g.* Dahdah 1987; Dahdah 1993; Wright 1996; Al-Ghalayyni 2005; Ryding 2005; ALECSO 2008). Arabic language scholars classify Arabic words into three main part-of-speech categories: namely nouns, verbs and particles. Khoja's tag set added categories of

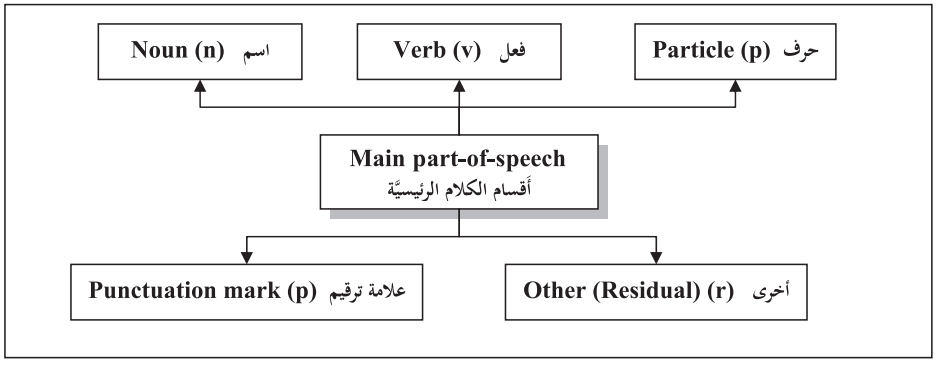

Figure 15. Main part-of-speech category attributes and letters used to represent them at position 1.

punctuation marks and residuals. The punctuation marks used in Arabic are (! ؛ : ؟ – . ،). Others (residuals) include other non-Arabic words appearing in the text such as currency, numbers or words in other languages. Figure 15 lists the attributes of the main part-of-speech category, which occupies the first character in the tag string.

4.2 Part-of-speech subcategories of noun

A noun is defined as a word that has complete meaning and no tense associated with it. The Arabic concept of complete meaning corresponds approximately to content words except that it also includes pronouns. Traditional Arabic grammar uses the concept of meaning to separate nouns and verbs from particles. This is roughly equivalent to content vs. function or lexical vs. grammatical in contemporary lexical terminology. This is not an exact correspondence since pronouns – a grammatical category – are a sub class of nouns. Arabic linguists distinguish many kinds of nouns. According to Dahdah (1987) nouns are classified into 21 kinds. Other classifications overlap. We classified nouns into 34 different types. Table 4 shows the 34 different types of nouns and examples of each type. Figure 16 shows the classification attributes of the noun part-of-speech category, which occupies the second character in the tag string.

4.3 Morphological feature of gender

Arabic classifies nouns according to gender into three classes:[16] nouns which are only masculine (مُذَكَّر) *mudakkar*, nouns which are only feminine (مُؤَنَّث) *mu'annat*, and nouns which are both masculine and feminine (common gender or neuter gender) (مُذَكَّر أَوْ مُؤَنَّث) *mudakkar 'aw mu'annat* such as ملح *milḥ* 'salt', and روح *rūḥ* 'spirit' (Wright 1996). Figure 17 shows the morphological feature of gender subcategories. Table 5 lists the 3 subcategories, with examples of masculine, feminine and common

Table 4. Noun types as classified by Arabic grammar scholars.

	Noun types	T	Meaning and Examples
1	Gerund / verbal noun المصدر *al-maṣdar*	**g**	A noun which indicates a case or an action that is not related to time or tense. E.g. فَرَحٌ *faraḥ*^{un} 'happiness'.
2	Gerund / verbal noun with initial *mīm* المصدر الميمي *al-maṣdar al-mīmī*	**m**	A noun which indicates a case or an action that is not related to time or tense. It has certain patterns which have the augmented letter **(م)** *mīm* at the beginning of the word. E.g. مُنْقَلِب ***m***unqalib 'turned over', مَوْعِد ***m***aw'id 'date'.
3	Gerund of instance مصدر المرَّة *maṣdar al-marra*^{h}	**o**	A noun that describes an action that has taken place once. It is formed by adding the feminine termination (ة) to the verbal noun. E.g. وَقْفَة *waqfa*^{h} 'one stop', زِيَارَة *ziyāra*^{h} 'a visit'.
4	Noun of state مصدر الهيئة/ مصدر النوع *maṣdar al-hay'a*^{h} */ maṣdar al-naw'*	**s**	A noun that describes an action. It indicates the manner (state, character and representation) of the action expressed by the verb. It always has the form فِعْلَةٌ *fi'la*^{tun}. E.g. مَشى الأسَد مِشْيَةَ *mašā mišya*^{ta} *al-'asad* 'he walked like a lion'.
5	Gerund of emphasis مصدر التوكيد *maṣdar al-tawkīd*	**e**	A noun that emphasizes an action. E.g. صَوَّرَ اللهُ الخَلْقَ تَصْوِيراً *ṣawwara allāhu al-ḫalqa taṣwīr*^{an} 'God does shape the creatures'.
6	Gerund of profession المصدر الصناعي *al-maṣdar al-ṣinā'ī*	**i**	A noun which indicates an industry or profession. The gerund of industry ends with doubled *ya'* followed by feminine *tā' marbūṭa*^{h} **(ة)**. E.g. إنسانيّة *'insāniyya*^{h} 'humanity', وطنيّة *waṭaniyya*^{h} 'nationality' and عالَمية *'ālamiyya*^{h} 'internationality'.
7	Pronoun الضمير *al-ḍamīr*	**p**	Pronouns that belong to this category are the disconnected pronouns. A sentence can start with a pronoun. Pronouns can follow the word **(إلّا)** *'illā* 'except'. E.g أنا مُجْتَهِدٌ *'anā muǧtahid*^{un} 'I am a hard worker', and ما اجْتَهَدَ إلّا أنا *mā 'iǧtahada 'illā 'anā* 'no one worked hard except me'. There are 24 pronouns classified into 12 nominative pronouns and 12 accusative pronouns. The nominative pronouns are: أنا *'anā* 'I', نَحْنُ *naḥnu* 'We', أنْتَ *'anta* 'You', أنْتِ *'anti* 'You', أنْتُما *'antumā* 'You', أنْتُم *'antum* 'You', أنْتُنَّ *'antunna* 'You', هُوَ *huwa* 'He', هِيَ *hiya* 'She', هُما *humā* 'They', هُم *hum* 'They', and هُنَّ *hunna* 'They'. The accusative pronouns are: إيَّايَ *'iyyāya* 'Me', إيَّانَا *'iyyānā* 'us', إيَّاكَ *'iyyāka* 'your', إيَّاكِ *'iyyāki* 'your', إيَّاكما *'iyyākumā* 'your', إيَّاكُم *'iyyākum* 'your', إيَّاكُنَّ *'iyyākunna* 'your', إيَّاهُ *'iyyāhu* 'his', إيَّاهَا *'iyyāhā* 'her', إيَّاهما *'iyyāhumā* 'they', إيَّاهم *'iyyāhum* 'they', إيَّاهنَّ *'iyyāhunna* 'they'.

	Noun types	T	Meaning and Examples
8	Demonstrative pronoun اسم الإشارة *'ism al-'išāra*^{h}	**d**	A noun that indicates by a tangible sign a person, an animal, a thing or a place such as جَاءَ هذا الرجل *ğā' hādā ar-raǧul* ' this man came', and رَأَيتُ تَينَ الفتاتين *ra'aytu tayna al-fatātayn* ' I saw these two girls'.
9	Specific relative pronoun اسم الموصول الخاص *'ism al-mawṣūl al-ḫāṣ*	**r**	A group of nouns that connect two sentences to give a full meaning. The special relative pronouns are affected by three morphological feature categories, number, gender and humanness. E.g. الَّذي *al-laḏī* 'who' is a singular masculine human pronoun; التي *al-latī* 'who' is a singular feminine human pronoun; اللواتي *al-lawātī* 'who' is a plural feminine human pronoun.
10	Non-specific relative pronoun اسم الموصول المشترك *'ism al-mawṣūl al-muštarak*	**c**	A group of nouns that connect two sentences to give a full meaning. The common relative pronouns are not affected by gender and number, so they have invariable form. They are affected by the morphological feature of humanness. E.g. مَنْ *man* 'who' is used for human nouns, ما *mā* 'who' is used for non-human nouns, and ذا *ḏā* 'what' and أيّ *'ayyu* 'which' are used for non-human nouns.
11	Interrogative pronoun اسم الاستفهام *'ism al-'istfhām*	**b**	A pronoun used to make a query or question about a thing or an action, e.g. مَنْ هذا؟ *man haḏā?* 'who is this?' ما العمل؟ *mā al- 'amal?* 'what shall we do?'. The nouns مَنْ *man 'who'* and ما *mā* 'what' are interrogative nouns.
12	Conditional noun اسم الشرط *'ism al-šarṭ*	**h**	A noun which connects two sentences. It indicates that the action in the second sentence does not occur unless the action of the first sentence has occurred, e.g. أَيُّ تِلْميذٍ يَجْتَهِدْ يَنْجَحْ *'ayyu tilmīḏ^{in} yaǧtahid yanǧaḥ* 'if any student studies hard, then he will succeed'. The noun أَيُّ *'ayyu* 'if any', is a conditional noun.
13	Allusive noun الكناية *al-kināya*^{h}	**a**	A noun which indicates a specific intention by means of unclear terms. These nouns are: كَأَيِّ *ka'ayyi* 'Any', كَذا *kaḏā* 'So and so', كم *kam* 'How …', كَيْتَ *kayta* 'So and so', ذَيْتَ *ḏayta* 'So and so', بِضْعْ *biḍ'u* 'few', فُلانُ *fulān* 'someone', e.g. عصفوراً اصطدتَ كَأَيِّ *ka'ayyi 'usfūr^{an} 'iṣṭadta* 'Like any bird you have hunted'. The word كَأَيِّ *ka'ayyi* 'As any' is a generalization
14	Adverb الظَّرف *aẓ-ẓarf*	**v**	A noun which indicates the time or place of the action. It incorporates into its overall meaning a sence of relative locality on time or place, *e.g.* حينَ *ḥīna* 'when', مُدَّة *mudda^{tu}* 'at a period of', and أمام *'amām* 'straight forward (direction)'

	Noun types	T	Meaning and Examples
15	Active participle اسم الفاعل *'ism al-fā'il*	**u**	A form that describes the doer of the action. This noun is derived from the action or the verb itself. E.g. كاتِبٌ *kātib*^{un} 'writer'. This noun is derived from the action of *writing* or the verb *write* كَتَبَ *kataba.*
16	Intensive active participle مُبَالَغة اسم الفاعِل *mubālaġa*^{t} *'ism al-fā'il*	**w**	A noun which has the same basic meaning as the present participle اسم الفاعل *'ism al-fā'il* but indicates an augmentation of the meaning of the present participle. E.g. كَتَّابٌ *kattāb*^{un} 'writer', which indicates that the *writer* writes a lot. *kattāb*^{un} is derived from the verb 'write' كَتَبَ *kataba.*
17	Passive participle اسم المفعول *'ism al-maf'ūl*	**k**	A derived noun which indicates an abstract meaning that describes something or someone affected by an action. E.g. مَكْسورٌ *maksūr*^{un} 'broken'. This noun is derived from the verb break كَسَرَ *kasara.*
18	Adjective الصِّفة المشبَّهة *aṣ-ṣifa*^{h} *al-mušabbaha*^{h}	**j**	A derived noun which indicates a meaning of firmness, *i.e.* the absolute existence of the quality in its possessor. *E.g.* الجُنْدِيُّ شُجَاعٌ *al-ǧundiyyu šuǧā'*^{un} 'brave soldier'. The word شُجَاعٌ *šuǧā'*^{un} 'brave' describes the soldier. This word is an adjective.
19	Noun of place اسم المكان *'ism al-mkān*	**l**	A derived noun which indicates the place of an action. E.g. مَطْبَخٌ *maṭbaḫ*^{un} 'kitchen' indicates the place of cooking.
20	Noun of time اسم زمان *'ism zamān*	**t**	A derived noun which indicates the time of the action or a verb. E.g. مَغْرِبٌ *maġrib*^{un} 'sunset'.
21	Instrumental noun اسم الآلة *'ism al-'āla*^{h}	**z**	A derived noun which indicates a tool used to some work. E.g. مِفْتاحٌ *miftāḥ*^{un} 'key', مِنشار *minšār* 'saw', and مِصباح *miṣbāḥ* 'light'.
22	Proper noun اسم العلم *'ism al-'alam*	**n**	The name of a dedicated or specific instance in a group or type. E.g. خالِدٌ *ḫālid*^{un} 'Khalid', عَبْدُاللهِ *'abdu allāhi* 'Abdullah', بَيْروتُ *bayrūt* 'Beirut (the capital city of Lebanon)'.
23	Generic noun اسم الجنس *'ism al-ǧins*	**q**	Indicates what is common to every element of the genus without being specific to any one of them. E.g. كِتابٌ *kitāb*^{un} 'book', رَجل *raǧul* 'man', and بيت *bayt* 'home'.
24	Numeral اسم العدد *'ism al-'adad*	+	A noun that indicates the quantity and order of countable nouns by transferring the numbers into the correct form of Arabic words. E.g. رَجُلٌ واحِدٌ *raǧul*^{un} ***wāḥid***^{un} 'one man'. رَجُلانِ إثنانِ *raǧulāni* ***'iṯnāni*** 'two men'. رِجالٍ ثَلاثَةُ *ṯalāṯatu riǧāl*^{in} 'three men'. The words واحد ، اثنان و ثلاثة *wāḥid, 'iṯnāni and ṯalāṯa*^{h} 'one', 'two' and 'three', are ordinal numeral nouns.

	Noun types	T	Meaning and Examples
25	Verb-like noun اسم الفعل *'ism al-fi'il*	&	A noun which acts as a verb in its meaning. It indicates time of action, e.g. شَتَّانَ *šattāna* 'how different they are!', هَيهَات *hayhāt* 'but oh! far from the mark!' and بَعُدَ *ba'uda* 'far away'.
26	The five nouns الأسماء الخمسة *al-'asmā' al-ḫamsa^{h}*	f	The five nouns are a group of five nouns belonging to the category of noun of genus. However, unlike standard nouns, which have three root letters, each of these nouns has only two root letters the third root letter being deemed to have been deleted. The five nouns are أَبٌ *'ab^{un}* 'father', أَخٌ *'aḫ^{un}* 'brother', حَمٌ *ḥam^{un}* 'father in law', فو *fū* (فَم *fam*)'mouth', and ذو *ḏū* 'owner'.
27	Relative noun اسم منسوب *'ism mansūb*	*	A declinable noun which has the suffix *-iyy*. It indicates affiliation of something to this noun. E.g. أُرْدُنِيٌّ *'urduniyy^{un}* 'Jordanian' (*i.e.* affiliated to Jordan).
28	Diminutive اسم تصغير *'ism taṣḡīr*	y	A declinable noun which has the sound *-ai-* after its second root letter. It indicates paucity, contempt or affection. E.g. دُرَيْهِمات *duraihimāt* 'a few dirhams', شُوَيْعِر *šuway'ir* 'poetaster', and بُنَيَّ *bunayya* 'my (little) son'.
29	Form of exaggeration صيغة مبالغة *ṣīḡaʿ al-mubālaḡa^{h}*	x	It indicates exaggeration of the quality of the qualified noun and occurs as a derived noun with the basic meaning of the present participle. E.g. زَرَّاع *zarrā'* 'a very good cultivator'.
30	Collective noun اسم جمع *'ism ǧam'*	$	A noun which indicates two or more. A singular form cannot be derived from this kind of noun. E.g. جَيْش *ǧayš* 'army', the corresponding singular being جندي *ǧundī* 'a soldier', or خَيْل *ḫayl* 'horses' the corresponding singular being فَرَس *faras* 'a horse'.
31	Plural collective noun اسم جنس جمعي *'ism ǧins ǧam'ī*	#	A noun of genus where the singular and plural share the same basic form in meaning and pronunciation. The singular form is distinguished by adding the feminine *tā' marbūtah* or the relative suffix *–ī*. E.g. زهر (زهرة) *zahr (zahra^{h})* 'flowers' ('a flower'), and عرب (عربي) *'arab ('arabī)* 'Arabs' ('an Arab').
32	Elative noun اسم تفضيل *'ism tafḍīl*	@	A derived noun used for the comparative and superlative when comparing persons or things. E.g. الأَسَدُ أقوى مِنَ الرَّجُلِ *al-'asadu 'aqwā mina ar-raǧuli* 'The lion is stronger than the man'. The noun أقوى *'aqwā* 'stronger' is used for comparing the strength of the lion and the man.

	Noun types	T	Meaning and Examples
33	Blend noun اسم منحوت *'ism manḥūt*	%	This consists in composing a single word by the fusion of two or more words, so that some letters are dropped from each word on condition that the resultive form has an authentically acceptable pronunciation and meaning. E.g. جَعْفَلَ *ğa'falu* 'Could I but sacrifice myself for you' composed from the words جَعِلْتُ فِدَاكَ *ğa'altu fidāka* (same meaning).
34	Ideophonic interjection اسم صوت *'ism ṣawt*	!	A noun improvised by human spontaneity and used initially as a verbal noun to talk to animals and small children, e.g. آه *āh* "Oh", هَال *hāl* used for horses.

gender words. The morphological feature of gender is represented at position 7 in the tag string.

Morphologically the masculine form shows the simplest and most basic shape (word structure), whereas feminine nouns usually have a suffix that marks their gender. On the other hand, semantically, nouns are arbitrarily classified into masculine or feminine, except where a noun refers to a human being or other creature, when it normally conforms to natural gender (Ryding 2005). Therefore, we can distinguish between two types of the morphological feature of gender that nouns can indicate: semantic gender and morphological gender. Semantic gender occurs where nouns indicate the natural gender of a human being or animal (male or female) or figurative gender for things that do not have natural gender. Morphological gender is defined by the noun being in its simplest form or by containing a feminine suffix attached to it. Discussion of the detailed classification of the morphological feature of gender into morphological gender and semantic gender is beyond the scope of this paper; we hope to present this in a later paper.

4.4 The morphological feature of unaugmented and augmented

Arabic verbs have roots consisting of three or four letters. From these roots many verbs can be derived by following certain patterns. There are many patterns for Arabic verbs. The standard way of determining the pattern of a verb is to refer to an Arabic lexicon or dictionary. Nonetheless, Arabic linguists have constructed general rules to extract these patterns. Verbs have two basic patterns consisting of three or four letters فَعَلَ *fa'ala* and فَعْلَلَ *fa'lala* respectively. Any verb derived following these two patterns is called an unaugmented verb (فعل مُجَرَّد) *fi'l muğarrad*. From فَعَلَ *fa'ala*, the basic triliteral pattern, 10 more patterns can be derived, and from فَعْلَلَ *fa'lala*, the basic quadriliteral pattern, 3 more patterns can be derived. These new patterns are derived by adding one, two or three letters to the basic patterns or by duplicating the second letter ع *'ayn* of the basic pattern. The group of letters that are added to the basic patterns to produce the other 13 patterns are ا , أ , ت , س , ل , م , ن , هـ , و , ي (*ā, ' , t, s, l, m, n , h, w, y*)

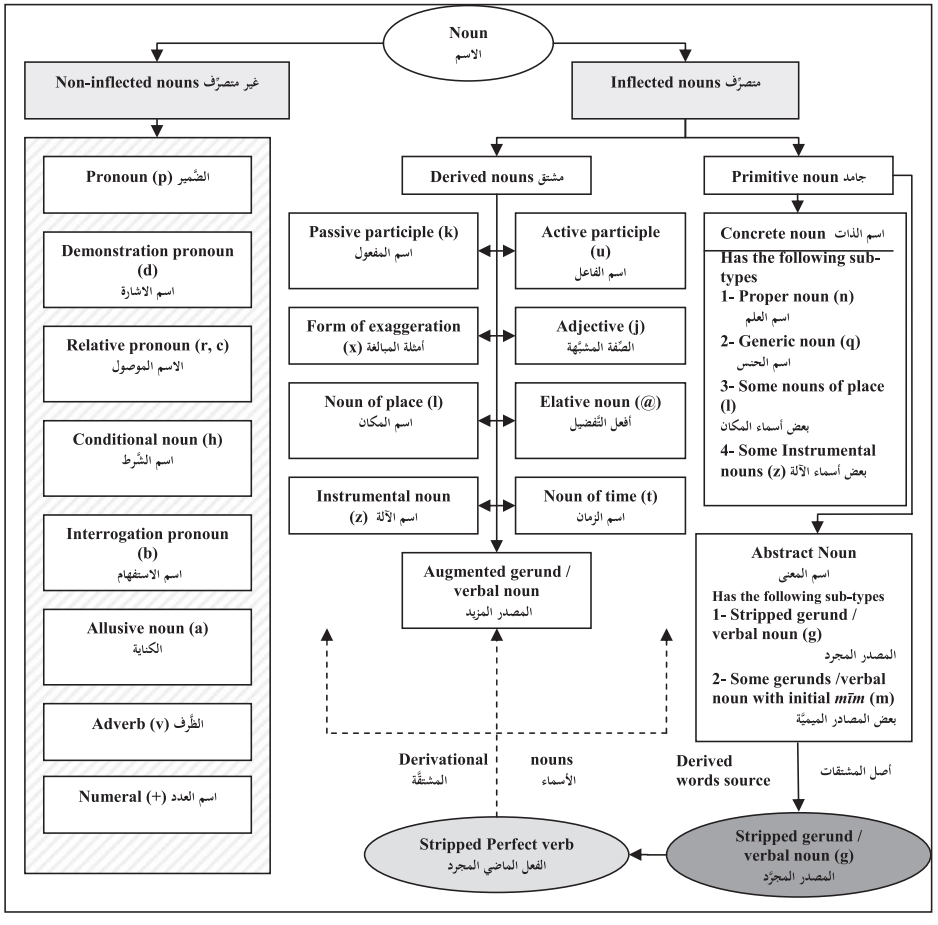

Figure 16. The classification attributes of noun part-of-speech subcategories with letter at position 2.

that combine with the word سألتمونيها *sa'altumūnīhā* 'you (second person, plural) asked me it (feminine, singular)' (Dahdah 1987; Dahdah 1993; Al-Ghalayyni 2005).

Unagmented declinable nouns are either triliteral ثُلاثي *ṯulāṯī* such as حجر *ḥağr* 'stone', quadriliteral رُباعي *rubā'ī* such as جعفر *ğa'far* 'male proper name', or quinquiliteral خُماسي *ḫumāsī* such as سَفرجل *safarğal* 'quince [kind of fruit]'. A noun which consists of more than five letters is an augmented noun. A noun can be augmented by one letter مزيد بحرف *mazīd bi ḥarf* such as حصان *ḥiṣān* 'horse' (augmented by *ā* ا) and قنديل *qindīl* 'light' (augmented by *ī* ي), augmented by two letters مزيد بحرفين *mazīd bi ḥarfayn* such as مصباح *miṣbāḥ* 'lamp' (augmented by *m* م and *ā* ا), augmented by three letters مزيد بثلاثة أحرف *mazīd bi ṯalāṯa*^{ti} *'aḥruf* such as انطلاق *'inṭilāq* 'starting' (augmented by *ā* ا, *n* ن and *ā* ا) and احرنجام *'iḥranğām* 'crowded' (augmented by *ā* ا, *n* ن and *ā* ا),

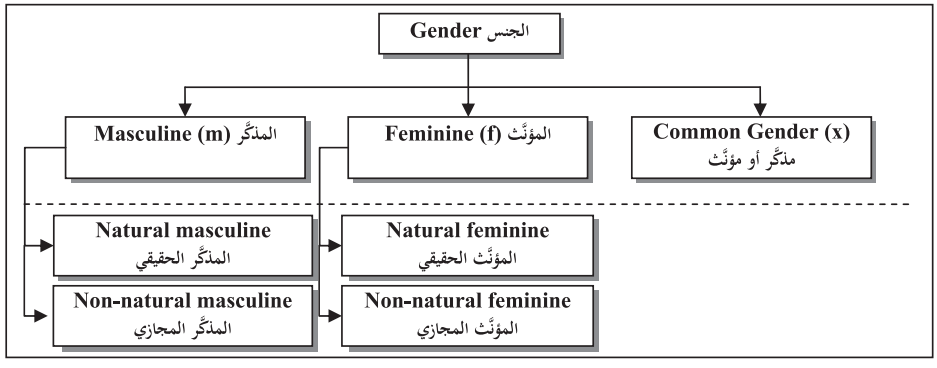

Figure 17. Arabic classification of nouns according to gender, with letter at position 7.

Table 5. Examples of gender category attributes for nouns, verbs, adjectives and pronouns.

#	Subcategories of gender	T	Examples			
			Noun	Verb	Adjective	Pronoun
1	Masculine مذكر *muḏakkar*	**m**	كتاب *kitāb* book	يكتبون *yaktubūn* they are writing (pl. / masc.)	كاتب *kātib* writer (sing. / masc.)	هو *huwa* he
2	Feminine مؤنث *mu'annaṯ*	**f**	مكتبة *maktaba*^{h} library	تكتبين *taktubīn* you are writing (sing. / fem.)	كاتبة *kātiba*^{h} writer (sing. / fem.)	هي *hiya* she
3	Common gender مذكر أو مؤنث *muḏakkar 'aw mu'annaṯ*	**x**	ملح *milḥ* salt	نكتب *naktubu* we are writing (pl. / masc. or fem.)	نائب[17] *nā'ib* member of parliament (sing./ masc. or fem.)	هما *humā* they (dual)

or augmented by four letters مزيد بأربعة أحرف *mazīd bi 'arba'a*^{ti} *'aḥruf* such as استغفار *'istiḡfār* 'asking for forgiveness' (augmented by *ā* ا, *s* س, *t* ت, and *ā* ا).

Table 6 shows examples of the 5 Unaugmented and Augmented category attributes. Figure 18 shows the 5 attributes of the Unaugmented and Augmented category, represented at position 19 in the tag string.

5. Evaluation

Two ways to validate the SALMA Tag Set of Arabic are: one, to propose it as a standard to the Arabic language computing community and have the standard adopted by others; two, to see how readily it can be applied to a sample of Arabic text, for example by mapping from an existing tagged corpus to the SALMA tag set.

Table 6. Examples of Unaugmented and Augmented category attributes.

Unaugmented and Augmented	T	Example		
		Triliteral verbs	Quadriliteral verbs	Nouns
Unaugmented المُجَرَّد *al-muğarrad*	**s**	فَتَحَ *fataḥa* 'he opened'.	دَحْرَجَ *daḥrağa* 'rolled'.	حجر*ḥağr* 'stone'. جعفر *ğa'far* 'male proper name'. سَفرجل *safarğal* 'quince, [kind of fruits]'
Augmented by one letter مَزِيْد بِحَرف *mazīd bi ḥarf*	**a**	يَفْتَحُ *yaftaḥu* 'he is opening'. The letter *yā'* يَ is added to the beginning of the verb stem فَتَحَ *fataḥa*	يُدَحْرِجُ *yudaḥriğu* 'he is rolling'. The letter *yā'* يَ is added to the beginning of the verb stem دَحْرَجَ *daḥrağa.*	حصان *ḥiṣān* 'horse'. قنديل *qindīl* 'light'.
Augmented by two letters مَزِيْد بِحَرفَيْن *mazīd bi ḥarfayn*	**b**	انْكَسَرَ *'inkasara* ' has broken'. The letters *'alif* ا and *nūn* نْ are added to the beginning of the verb stem كَسَرَ *kasara* 'broke'.	يَتَدَحْرَجُ *yatadaḥrağu* ' is rolling'. The letters *yā'* يَ and *tā'* تَ are added to the verb stem دَحْرَجَ *daḥrağa* 'rolled'.	مصباح *miṣbāḥ* 'lamp'.
Augmented by three letters مَزِيْد بِثَلاثَةِ حُرُوف *mazīd bi ṯalāṯa^{ti} 'aḥuf*	**t**	اسْتَخْرَجَ *'istaḫrağa* has extracted. The letters *'alif* ا, *sīn* س *and tā'* تَ are added to the beginning of the verb stem خَرَجَ *ḫarağa* 'extracted'.	----------	انطلاق *'inṭilāq* 'starting' احرنجام *'iḥranğām* 'crowded'
Augmented by four letters مزيد بأربعة أحرف *mazīd bi 'arba'a^{ti} 'aḥruf*	**q**	----------	----------	استغفار *'istiğfār* 'asking for forgiveness'

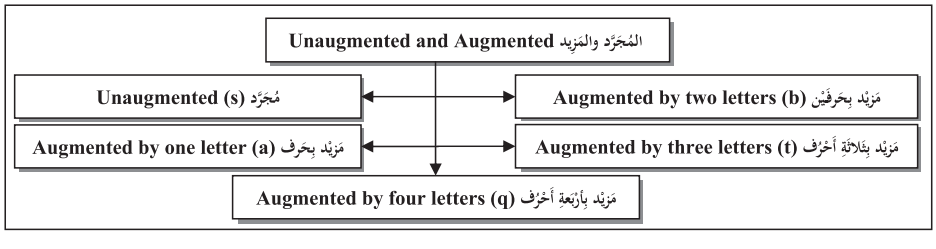

Figure 18. The Unaugmented and Augmented category attributes, with letter at position 19.

The SALMA – Tag Set has been used in the SALMA – Tagger (***S***awalha ***A***twell ***L***eeds ***M***orphological ***A***nalysis – Tagger). It is used as the standard for specifying the word's morphemes and for encoding the morphological features of each morpheme (Sawalha & Atwell 2009b; Sawalha & Atwell 2009a). The SALMA – Tag Set has been published online (http://www.comp.leeds.ac.uk/sawalha/tagset.html) and has been

adopted as a standard by other Arabic language computing researchers. For instance, part of the tag set is also used in the Arabic morphological analyzer and part-of-speech tagger Qutuf (Altabbaa, Al-Zaraee & Shukairy 2010). Qutuf uses the main part-of-speech, the subcategories of nouns, the subcategories of verbs named as verb aspects, the subcategories of particles and the morphological features of gender, number, person, case or mood, definiteness, voice, transitivity, and part of the declension and conjugation category named as perfectness. Qutuf does not use the SALMA – Tag format. Rather it uses a tag consisting of slots for each feature separated by a comma. Another re-use of the SALMA – Tag Set has been reported as a standard for evaluating Arabic morphological analyzers, and for building a Gold Standard for evaluating Arabic morphological analyzers and part-of-speech taggers (Hamada 2010).

Our second method for evaluating the SALMA – Tag Set is to apply it to a sample of Arabic text, by mapping from an existing broad tag set to the more fine-grained SALMA – Tag Set. We used the Quranic Arabic Corpus morphological annotation of a sample text, chapter 29, consisting of about 1,000 words. We developed an automated mapping algorithm to map the Quranic Arabic Corpus morphological tags to our SALMA – Tags. After that, the automatically mapped morphological features tags were manually verified and corrected to provide a new fine-grain Gold Standard for evaluating Arabic morphological analyzers and part-of-speech taggers.

The mapping from the Quranic Arabic Corpus morphological tag set to the SALMA – Tag Set was done by the following five-step procedure. First, mapping classical to modern character-set: the Quranic Arabic Corpus uses the classical Othmani script of the Qur'an (77,430 words) which was mapped to the Modern Standard Arabic (MSA) script (77,797 words). This was achieved by applying one-to-one mapping except for some cases where one word in Othmani script is mapped to two words in MSA such as the word يٰمُوسَىٰٓ *yāmūsā* 'O Musa "Moses"!'. In Othmani script this is one word but it is written as two words in MSA script: يَا مُوسَى *yā mūsā*. Second, splitting whole-word tags into morpheme-tags: the morphological tag in the Quranic Arabic Corpus is a whole-word tag, composed by combining the prefix with the stem and suffix morphological tags, separated by (+) signs. The words and their morphological tags were automatically divided into morphemes and morphemes tags. Third, mapping of feature-labels: the mnemonics of the Quranic Arabic Corpus tags were mapped to their equivalent in the SALMA – Tag Set. Then, SALMA – Tag Set templates were applied to specify the applicable and non-applicable morphological features of the analyzed morpheme. Fourth, adjustments to morpheme tokenization: due the differences between the underlying word tokenization model used in the Quranic Arabic Corpus and the one required for the SALMA – Tag Set, we replaced the mapped tags of the prefixes and suffixes with SALMA tags by matching them to the clitics and affixes lists used by the SALMA – Tagger (Sawalha & Atwell 2009a; Sawalha & Atwell 2010). Fifth, extrapolation of missing fine-grain features: for these morphological features which are not included in the Quranic Arabic Corpus tag set, automatic 'feature-prediction' procedures applied linguistic knowledge extracted from traditional Arabic grammar textbooks, encoded as a computational

	QAC morpheme tag	SALMA tags: steps 1-4	SALMA tags: step 5	SALMA tags: corrected
الم	POS:INL	p--?-----????—?-----	p--?-----s-s----------	p--b-----s-s----------
أَ	A:INTG+	p--i-----s----------	p--i-----s----------	p--i-----s----------
حَسِبَ	POS:V PERF 3MS	v-p---mst--?-?-??????-	v-p---msts-f-ambhvsta-	v-p---msts-f-amohvsta-
ال	Al+	r---d----------	r---d----------	r---d----------
نَاسُ	POS:N MP NOM	n?—--mp-?n??---????-?	n?—--mp-vndd---ndst-s	n#----mj-vndd---hdst-s
أَنْ	POS:SUB	p--g-------?----------	p--g-----s-s----------	p--g-----s-s----------
يُ	NULL	r---a----------	r---a----------	r---a----------
تُرَكُ	POS:V IMPF PASS 3MP MOOD:SUBJ	v-c---mptda?-p???????-	v-c---mptdao-pmbhvtta-	v-c---mptdao-pmohvtta-
وا	PRON:3MP	r---r-mptsnw----------	r---r-mptsnw----------	r---r-mpts-s----------
أَنْ	POS:SUB	p--g-------?----------	p--g-----s-s----------	p--g-----s-s----------
يَ	NULL	r---a----------	r---a----------	r---a----------
قُولُ	POS:V IMPF 3MP MOOD:SUBJ	v-c---mptda?-????????-	v-c---mptdao-amohvtto-	v-c---mptdao-amohvtto-
وا	PRON:3MP	r---r-mptsnw----------	r---r-mptsnw----------	r---r-mpts-s----------
آمَنَ	POS:V PERF (IV) 1MP	v-p---mpf--?-?-??????-	v-p---mpfs-s-amohvttc-	v-p---mpfs-s-amohvttc-
نَا	PRON:1MP	r---r-xpfs??----------	r---r-xpfs??----------	r---r-xpfs-s----------
وَ	wa+	p--c----------	p--c----------	p--c-----s-f-------- -
هُمْ	POS:PRON 3MP	np----mpt--??---?-----	np----mpts-si---hn---?	np----mpts-si---h n----
لَا	POS:NEG	p--n-------?----------	p--n-----s-s----------	p--n-----s-s----------
يُ	NULL	r---a----------	r---a----------	r---a----------
قُتَّلُ	POS:V IMPF PASS 3MP	v-c---mpt-??-p???????-	v-c---mptdnn-pmohvtta-	v-c---mptdnn-pmohvtta-
ونَ	PRON:3MP	r---r-mp?snn----------	r---r-mp?snn----------	r---r-mpts-f----------

Figure 19. A sample of the Quranic Arabic Corpus tags and their mapped SALMA tags after applying the mapping procedure steps 1–4, step 5 and manually correcting the tags.

rule-based system, to automatically predict the values of the missing morphological features of the word. Finally, the mapped SALMA tags were manually proofread and corrected by an Arabic language expert. The result is a sample Gold Standard annotated corpus for evaluating morphological analyzers and part-of-speech taggers for Arabic text.

Figure 19 shows examples of mapping from the Quranic Arabic Corpus tags to SALMA – tags, at various stages of processing: results after applying steps 1 to 4, the results after applying step 5, and the results after manually proofreading and correcting the tags. Figure 20 shows the percentage of cases mapped correctly for each morphological feature after applying steps 1 to 4, step 5, and the percentage of cases corrected manually for each category. Individual features required varying amounts of manual correction, ranging from Punctuation and Verb Root features which were predicted with 0% error rate, to 37.26% error rate in predicting Case and Mood Marks. Overall, 53.5% of whole tags needed some correction in the final proofreading

Category		Mapping steps 1-4	mapping step 5	Manual correction
1	Main Part-of-Speech	99.17%	100.00%	2.91%
2	Part-of-Speech: Noun	50.60%	83.05%	20.04%
3	Part-of-Speech: Verb	100.00%	100.00%	0.07%
4	Part-of-Speech: Particle	93.93%	97.79%	5.44%
5	Part-of-Speech: Other (Residual)	100.00%	100.00%	7.38%
6	Punctuation marks	100.00%	100.00%	0.00%
7	Gender	89.13%	92.83%	15.87%
8	Number	77.57%	82.24%	30.25%
9	Person	84.58%	100.00%	8.27%
10	Inflectional Morphology	89.29%	100.00%	16.69%
11	Case or Mood	68.61%	85.24%	29.36%
12	Case and Mood Marks	17.28%	86.46%	37.26%
13	Definiteness	14.10%	100.00%	5.22%
14	Voice	7.17%	100.00%	2.01%
15	Emphasized and Non-emphasized	12.31%	100.00%	0.07%
16	Transitivity	0.00%	100.00%	0.45%
17	Rational	0.00%	100.00%	8.20%
18	Declension and Conjugation	4.93%	100.00%	14.31%
19	Unaugmented and Augmented	0.00%	100.00%	6.93%
20	Number of Root Letters	0.00%	100.00%	0.37%
21	Verb Root	0.00%	100.00%	0.00%
22	Noun Finals	0.00%	91.70%	9.69%

Figure 20. The percentage of each morphological feature mapped after applying steps 1 to 4, step 5, and the percentage of errors corrected in final proofreading for each category.

stage; however, many of these corrections were very minor such as replacing '?' (unknown) with '-' (not applicable). The use of 22 morphological feature categories for each morpheme is bound to increase the potential for making annotation mistakes; however, this result demonstrates that the SALMA – Tag Set can feasibly be used to annotate Arabic text corpora with rich morphological information, appropriate to the rich morphology of Arabic.

6. Conclusions

A range of Arabic Part-of-Speech taggers exist, each with a different tag set; we have illustrated and compared some of these, and this suggests the need for a common standard to simplify and promote comparisons and sharing of resources. We review generic design criteria for corpus tag sets, and see that some of these principles have been applied in existing tag sets; but there is still room for improvement, in the design of a standard tag set for Arabic Part-of-Speech taggers and tagged corpora. The SALMA – Tag Set captures long-established traditional morphological features of Arabic, in a compact yet transparent notation. A tag consists of 22 characters; each position represents a feature and the letter at that location represents a value or attribute of the morphological feature; the dash '-' represents

a feature not relevant to a given word. The SALMA – Tag Set is not tied to a specific tagging algorithm or theory, and other tag sets could be mapped onto this standard, to simplify and promote comparisons between and reuse of Arabic taggers and tagged corpora.

The SALMA – Tag Set has been validated in two ways. First, it was validated by proposing it as a standard to Arabic language computing community, and has been adopted in Arabic language processing systems. The SALMA – Tag Set has been used in the SALMA – Tagger to encode the morphological features of each morpheme (Sawalha & Atwell 2009a; Sawalha & Atwell 2010). Parts of The SALMA – Tag Set were also used in the Arabic morphological analyzer and part-of-speech tagger Qutuf (Altabbaa et al. 2010). Moreover, the SALMA – Tag Set has been reported as a standard for evaluating morphological analyzers for Arabic text and for building a gold standard for evaluating morphological analyzers and part-of-speech taggers for Arabic text (Hamada 2010).

Second, we presented an empirical approach to evaluating the SALMA – Tag Set of Arabic, showing that it can be applied to an Arabic text corpus, by mapping from an existing tag set to the SALMA – Tag Set. The morphological tags of a 1,000-word test text, chapter 29 of the Quranic Arabic Corpus, were automatically mapped to SALMA tags. Then, the mapped tags were proofread and corrected. The result of mapping and correction of the SALMA – tagging of this corpus is a new Gold Standard for evaluating Arabic morphological analyzers and part-of-speech taggers with a detailed fine-grain description of the morphological features of each morpheme, encoded using SALMA tags.

We invite other Arabic language computing researchers to take up the SALMA – Tag Set and Gold Standard tagged corpus, to promote comparability and interoperability of Arabic morphological analysers and Part-of-Speech taggers.

Appendix – A The SALMA Tag Set for Arabic text

Table A.1. SALMA Tag Set categories.

Position	Morphological Features Categories		
1	Main Part-of-Speech	أقسام الكلام الرئيسيَّة	*'aqsām al-kalām ar-r'īsiyya^{t}*
2	Part-of-Speech: Noun	أقسام الكلام الفرعيَّة (الاسم)	*'aqsām al-kalām al-far'iyya^{t} (al-'ism)*
3	Part-of-Speech: Verb	أقسام الكلام الفرعيَّة (الفعل)	*'aqsām al-kalām al-far'iyya^{t} (al-fi'l)*
4	Part-of-Speech: Particle	أقسام الكلام الفرعيَّة (الحرف)	*'aqsām al-kalām al-far'iyya^{t} (al-ḥarf)*
5	Part-of-Speech: Other (Residual)	أقسام الكلام الفرعيَّة (أخرى)	*'aqsām al-kalām al-far'iyya^{t} ('uẖrā)*
6	Punctuation marks	أقسام الكلام الفرعيَّة (علامات الترقيم)	*'aqsām al-kalām al-far'iyya^{t} ('alāmāt at-tarqīm)*
7	Gender	المُذَكَّر والمُؤَنَّث	*al-muḏakkar wa al-mu'annaṯ*
8	Number	العدد	*al-'adad*
9	Person	الاسناد	*al-'isnād*

Position	Morphological Features Categories		
10	Inflectional Morphology	الصَّرف	*aṣ-ṣarf*
11	Case or Mood	الحالة الإعرابية للاسم أو الفعل	*al-ḥāla^{tu} al-'i'rābiyya^{tu} lil-'ism 'aw al-fi'l*
12	Case and Mood Marks	علامة الإعراب أو البناء	*'alāmāt al-'i'rāb wa al-binā'*
13	Definiteness	المَعْرِفَة والنَّكِرَة	*al-ma'rifa^{ti} wa an-nakira^{ti}*
14	Voice	المَبْنِي لِلمَعْلُوم و المَبْنِي لِلمَجْهُول	*al-mabnī lil-ma'lūm wa al-mabnī lil-maǧhūl*
15	Emphasized and Non-emphasized	المُؤَكَّد وغيرُ المُؤَكَّد	*al-mu'akkad wa ǧayir al-mu'akkad*
16	Transitivity	اللازم والمتعدي	*Al-lāzim wa al-muta'adi*
17	Rational	العاقل وغير العاقل	*al-'āqil wa ǧayir al-'āqil*
18	Declension and Conjugation	التَّصريف	*at-taṣrīf*
19	Unaugmented and Augmented	المجرَّد والمزيد	*al-muǧarrad wa al-mazīd*
20	Number of Root Letters	عَدَد أَحْرُف الجَذْر	*'adad 'aḥruf al-ǧaḏr*
21	Verb Root	بُنية الفعل	*bunya^{tu} al-fi'l*
22	Noun Finals	أقسام الأسم تبعاً للفظ آخره	*'aqsām al-'ismi tib'^{an} li-lafẓi 'āḫirhi*

Table A.2. Main part-of-speech category attributes and tags at position 1.

Position	Feature Name				Tag
1	**Main Part-of-Speech** أَقسام الكلام الرئيسيَّة *'aqsām al-kalām ar-r'īsiyya^{t}*				
	Noun	اسم	*'ism*	كِتَاب *kitāb* 'book'	**n**
	Verb	فعل	*fi'l*	كَتَبَ *katab* 'wrote'	**v**
	Particle	حرف	*ḥarf*	عَلَى *'alā* 'on'	**p**
	Other (Residual)	أخرى	*'uḫrā*	كاتِبَةٌ *kātiba^{tun}* 'writer. FEM	**r**
	Punctuation	علامة ترقيم	*'alāmat tarqīm*	قالَ: أنا ذاهِبٌ *qāla : 'anā ḏāhib^{un}* 'he said: I am leaving'	**u**

Table A.3. Part-of-Speech subcategories of Noun attributes and their tags at position 2.

Position	Feature Name				Tag
2	**Part-of-Speech: Noun** أَقسام الكلام الفرعيَّة (الاسم) *'aqsām al-kalām al-far'iyya^{t} (al-'ism)*				
	Gerund / Verbal noun	المصدر	*al-maṣdar*	ضَرْب *ḍarb* 'hitting'	**g**
	Gerund/ verbal noun with initial *mīm*	المصدر الميمي	*al-maṣdar al-mīmī*	مَوعِد *maw'id* 'date'	**m**
	Gerund of instance	مصدر المرَّة	*maṣdar al-marra^{h}*	نَظْرَة *naẓra^{h}* 'one look'	**o**
	Gerund of state	مصدر الهيئة/ مصدر النوع	*maṣdar al-hay'a^{h} / maṣdar al-naw'*	جِلْسَة *ǧilsa^{h}* 'sitting position'	**s**

Position	Feature Name				Tag
2	**Part-of-Speech: Noun** (الاسم) أقسام الكلام الفرعيّة *'aqsām al-kalām al-far'iyya^{t} (al-'ism)*				
	Gerund of emphasis	مصدر التوكيد	*maṣdar al-tawkīd*	حطّمتُ الخزانةَ تحطيماً *ḥaṭṭamtu al-ḫizāna^{ta} taḥṭīm^{an}* 'I completely destroyed the wardrobe'	**e**
	Gerund of profession	المصدر الصناعي	*al-maṣdar al-ṣinā'ī*	فُروسيَّة *furūsiyya^{h}* 'Horsemanship'	**i**
	Pronoun	الضمير	*al-ḍamīr*	هو *huwa* 'He'	**p**
	Demonstrative pronoun	اسم الإشارة	*'ism al-'šāra^{h}*	هذا *hāḏā* 'This'	**d**
	Specific relative pronoun	اسم الموصول الخاص	*'ism al-mawṣūl al-ḫāṣ*	الذي *al-laḏī* 'Who'	**r**
	Non-specific relative pronoun	اسم الموصول المشترك	*'ism al-mawṣūl al-muštarak*	مَنْ *man* 'Who'	**c**
	Interrogative pronoun	اسم الاستفهام	*'ism al-'istfhām*	مَنْ *man* 'Who?'	**b**
	Conditional noun	اسم الشرط	*'ism al-šarṭ*	أينما *aynamā* 'Where ever'	**h**
	Allusive noun	الكناية	*al-kināya^{h}*	كذا *kaḏā* 'As well as'	**a**
	Adverb	الظَّرف	*aẓ-ẓarf*	يوم *yawm* 'Day'	**v**
	Active participle	اسم الفاعل	*'ism al-fā'il*	ضارِب *ḍārib* 'Hitter'	**u**
	Intensive Active participle	مبالغة اسم الفاعل	*mubālagā^{t} 'ism al-fā'il*	جَرَّاح *ǧarraḥ* 'Surgeon'	**w**
	Passive participle	اسم المفعول	*'ism al-mf'ūl*	مَضْرُوب *maḍrūb* 'Struck'	**k**
	Adjective	الصّفة المشبَّهة	*aṣ-ṣifa^{h} al-mušabbaha^{h}*	طويل *ṭawīl* 'Tall'	**j**
	Noun of place	اسم المكان	*'ism al-mkān*	مَكْتَب *maktab* 'Office'	**l**
	Noun of time	اسم زمان	*'ism zamān*	مَطْلِع *maṭli'* 'Start time'	**t**
	Instrumental noun	اسم الآلة	*'ism al-'āla^{h}*	مِنْشار *minšār* 'Saw'	**z**
	Proper noun	اسم العلم	*'ism al-'alam*	فاطِمَة *fāṭima^{h}* 'Fatima'	**n**
	Generic noun	اسم الجنس	*'ism al-ǧins*	حِصان *hiṣān* 'Horse'	**q**
	Numeral	اسم العدد	*'ism al-'adad*	ثلاثة *ṯalāṯa^{h}* 'Three'	**+**
	Verb-like noun	اسم الفعل	*'ism al-fi'l*	هيهات *hayhāt* Wishing	**&**
	Five nouns	الأسماء الخمسة	*al-'asmā' al-ḫamsa^{h}*	أبٌ *'ab^{un}* 'Father'	**f**
	Relative noun	اسم منسوب	*'ism mansūb*	عِلْمِيّ *'ilmiyy^{un}* Scientific	*****
	Diminutive	اسم تصغير	*'ism taṣḡīr*	شُجَيْرَة *šuǧayra^{h}* 'Bush'	**y**
	Form of exaggeration	صيغة مبالغة	*ṣīgā^{t} al-mubālagā^{h}*	جَبَّار *ǧabbār* 'Tremendous'	**x**
	Collective noun	اسم جمع	*'ism ǧam'*	قوم *qawm* 'Folk'	**$**
	Plural generic noun	اسم جنس جمعي	*'ism ǧins ǧam'ī*	تفاح *tuffāḥ* 'Apple'	**#**
	Elative noun	اسم تفضيل	*'ism tafḍīl*	أفضل *'afḍal* 'Better'	**@**
	Blend noun	اسم منحوت	*'ism manḥūt*	بسملة *basmala^{h}* 'Bismallah'	**%**
	Ideophonic interjection	اسم صوت	*'ism ṣawt*	آه *'āh* 'Ah'	**!**

Table A.4. Part-of-Speech subcategory of verb attributes and their tags at position 3.

Position	Feature Name				Tag
3	**Part-of-Speech: Verb** (الفعل) أقسام الكلام الفرعيّة *'aqsām al-kalām al-far'iyya[t] (al-fi'l)*				
	Perfect verb	فعل ماضٍ	*fi'l māḍi[in]*	كَتَبَ *kataba* 'He wrote'	**p**
	Imperfect verb	فعل مضارع	*fi'l muḍāri'*	يَكْتُبُ *yaktubu* 'He is writing'	**c**
	Imperative verb	فعل الأمر	*fi'l al-'amr*	اُكْتُبْ *'uktub* 'write'	**i**

Table A.5. Part-of-speech subcategories of Particles attributes and their tags at position 4.

Position	Feature Name				Tag
4	**Part-of-Speech: Particle** (الحروف) أقسام الكلام الفرعيّة *'aqsām al-kalām al-far'iyya[t] (al-ḥarf)*				
	Jussive-governing particle	حرف جزم	*ḥarf ğazim*	لَمْ *lam* 'No'	**j**
	Subjunctive-governing particle	حرف نصب	*ḥarf naṣib*	كَيْ *kay* 'So that'	**o**
	Partially subjunctive-governing particle	حرف النصب الفرعي	*ḥarf naṣib far'ī*	حتى *ḥattā* 'Till'	**u**
	Preposition	حرف جر	*ḥarf ğarr*	إلى *'ilā* 'To'	**p**
	Annulling particle	حرف ناسخ	*ḥarf nāsiẖ*	ما *mā* 'No'	**a**
	Conjunction	حرف عطف	*ḥarf 'aṭif*	و *wa* 'And'	**c**
	Vocative particle	حرف نداء	*ḥarf nidā'*	يا *yā* 'Oh'	**v**
	Exceptive particle	حرف استثناء	*ḥarf 'stiṯnā'*	إلاَّ *'illā* 'Except'	**x**
	Interrogative particle	حرف استفهام	*ḥarf 'stifhām*	هل *hal* 'Is?'	**i**
	Particle of futurity	حرف استقبال	*ḥarf 'stiqbāl*	سوف *sawfa* 'Will'	**f**
	Causative particle	حرف تعليل	*ḥarf ta'līl*	كي *kay* 'To'	**s**
	Negative particle	حرف نفي	*ḥarf nafī*	لَمْ *lam* 'No'	**n**
	Jurative particle	حرف قسم	*ḥarf qasam*	بِ *bi* 'Swear'	**q**
	Yes/No response particle	حرف الجواب	*ḥarf ğawāb*	نعم *na'am* 'Yes'	**w**
	Jussive-governing conditional particle	حرف شرط جازم	*ḥarf šart ğāzim*	إنْ *'in* 'If'	**k**
	Particle of incitement	حرف تحضيض	*ḥarf taḥḍīḍ*	هلَّا *hallā* 'Would'	**m**
	Gerund-equivalent particle	حرف مصدري	*ḥarf maṣdarī*	أنْ *'an* 'To'	**g**
	Particle of attention	حرف تنبيه	*ḥarf tanbī[h]*	ألا *'alā* 'Careful'	**t**
	Emphatic particle	حرف توكيد	*ḥarf tawkīd*	إنَّ *'inna* 'Emphasis'	**z**
	Explanatory particle	حرف تفسير	*ḥarf tafsīr*	أي *'ay* 'i.e'	**d**
	Particle of comparison	حرف تشبيه	*ḥarf tašbī[h]*	كَأنَّ *ka'anna* 'Similar'	**l**
	Non-governing particles	حرف غير عامل	*ḥarf ğayr 'āmil*	قَدْ *qad* 'Already or perhaps'	**b**

Table A.6. Part-of-speech subcategories of Other (Residuals) attributes and their tags at position 5.

Position	Feature Name				Tag
5	**Part-of-Speech: Other** أقسام الكلام الفرعيّة (أُخرَى) *'aqsām al-kalām al-far'iyya^{t} ('uẖrā)*				
	Prefix	زيادة في أول الكلمة	*ziyāda^{h} fī 'awal al-kalima^{h}*	استكتبني *'istaktabanī* 'He employed me as a writer'	**p**
	Suffix	زيادة في آخر الكلمة	*ziyāda^{h} fī 'āẖir al-kalima^{h}*	أصدقاء *'aṣdiqā'* 'Friends'	**s**
	Suffixed pronoun	ضمير متصل	*ḍamīr mutaṣil*	كِتابُهُ *kitaba^{hu}* 'His book'	**r**
	tā' marbūṭa^{h}	تاء مربوطة	*tā' marbūṭa^{h}*	كاتبة *kātiba^{tun}* 'She-writer'	**t**
	Relative *yā'*	ياء النسبة	*yā' an-nisba^{h}*	عَرَبِيّ *'arabiyy* 'Arabian'	**y**
	tanwīn	تنوين	*tanwīn*	كِتابٌ *kitāb^{un}* 'A book'	**k**
	tā' of femininization	تاء التأنيث	*tā' al-ta'nīṯ*	كَتَبَتْ *katabaṯ* 'She wrote'	**f**
	nūn of protection	نون الوقاية	*nūn al-wiqāya^{h}*	سَأَلَنِي *sa'alanī* 'He asked me'	**n**
	Emphatic *nūn*	نون التوكيد	*nūn al-tawkīd*	يَضْرِبَنَّ *yaḍribanna* 'They are hitting'	**z**
	Imperfect prefix	حرف مضارعة	*ḥarf muḍāra'a^{h}*	يَسْأَلُ *yas'alu* 'He is asking'	**a**
	Definite article	أداة تعريف	*'adā^{t} ta'rīf*	الكتاب *al-kitāb* 'The book'	**d**
	Masculine sound plural letters	حروف جمع المذكر السالم	*ḥurūf ǧam' al-muḏakkar as-sālim*	الكاتبون *al-kātibūn* 'The writers (MAS)'	**m**
	Feminine sound plural letters	حروف جمع المؤنث السالم	*ḥurūf ǧam' al-mu'nnaṯ as-sālim*	الكاتبات *al-kātibāt* 'The writers (FEM)'	**l**
	Dual letters	حروف المثنى	*ḥurūf al-muṯannā*	الكاتبان *al-kātibān* 'The two writers'	**u**
	Imperative prefix	حروف الأمر	*ḥurūf al-'amr*	اكتب *'uktub* 'Write'	**I**
	Number (digits)	رَقْم	*raqam*	(+325461) (-897,653) (0.986)	**g**
	Currency	عُمْلَة	*'umla^{t}*	(1,500د.أ) (2,927ر.س) ($250)	**c**
	Date	تَارِيخ	*tārīẖ*	(27/09/2011) (27 أيلول 2011)	**e**
	Non-Arabic word	كَلِمَة غَيْر عَرَبِيّة	*kalima^{t} ḡayr 'arabiyya^{h}*	windows, photoshop, games, download	**w**
	Borrowed (foreign) word	كَلِمَة مُعَرَّبَة	*kalima^{t} mu'arraba^{h}*	كُوزْمُوبُولِيتَان *kuzmūbūlītān* 'cosmopolitan'	**x**

Table A.7. Part-of-speech subcategories of Punctuation Marks attributes and their tags at position 6.

Position	Feature Name				Tag
6	**Punctuation Marks** أقسام الكلام الفرعية (علامات الترقيم) *'aqsām al-kalām al-far'iyya^{t} ('alāmāt at-tarqīm)*				
	Full stop	نقطة	*nuqṭa^{h}*	(.)	**s**
	Comma	فاصلة	*fāṣila^{h}*	(،)	**c**
	Colon	نقطتان	*nuqṭatān*	(:)	**n**
	Semi colon	فاصلة منقوطة	*fāṣila^{h} manqūṭa^{h}*	(؛)	**l**
	Parentheses	قوسان	*qawsān*	(( ))	**p**
	Square brackets	قوسان حاصرتان	*qawsān ḥāṣiratān*	([ ])	**b**
	Quotation mark	علامة اقتباس	*'alāma^{tu} 'iqtibās*	(" ")	**t**
	Dash	شرطة معترضة	*šarṭa^{h} mu'tariḍa^{h}*	(–)	**d**
	Question mark	علامة استفهام	*'alāma^{tu} 'istifhām*	(؟)	**q**
	Exclamation mark	علامة تعجب	*'alāma^{tu} ta'aǧǧub*	(!)	**e**
	Ellipsis mark	علامة حذف	*'alāma^{tu} ḥaḏf*	(…)	**i**
	Continuation mark	علامة التَّابعية	*'alāma^{tu} at-tabi'yya^{h}*	(=)	**f**

Table A.8. Morphological feature of Gender attributes and their tags at position 7.

Position	Feature Name				Tag
7	**Morphological Gender** المُذَكَّر والمُؤَنَّث *al-muḏakkar wa al-mu'annaṯ*				
	Masculine	مذكر	*muḏakkar*	رجل *raǧul* 'man'	**m**
	Feminine	مؤنث	*mu'annaṯ*	امرأة *'imra'a*^{h} Woman	**f**
	Common gender	مذكر أو مؤنث	*muḏakkar 'aw mu'annaṯ*	ملح *milḥ* 'Salt' روح *rūḥ* 'Soul'	**x**

Table A.9. Morphological feature of Number attributes and their tags at position 8.

Position	Feature Name				Tag
8	**Number** العدد *al-'adad*				
	Singular	مفرد	*mufrad*	قلم *qalam* 'A pen' فلّاح *fallāḥ* 'Farmer' منارة *manāra*^{h} 'A minaret'	**s**
	Dual	مثنى	*muṯannā*	(قلم: قلمان، قلمين) (qalam: qalamān, qalamayn) '(A pen: two pens)' (منارة: منارتان، منارتين) (*manāra*^{h}*: manāratān, manāratayn)*(A minaret: two minarets)	**d**
	Sound plural	جمع سالم	*ǧami' sālim*	(فلّاح: فلّاحون، فلّاحين) (*fallāḥ: fallāḥūn, fallāḥīn)* (A farmer: Farmers)' (منارة: منارات) (*manāra*^{h}*: manārāt)* (A minaret: minarets)	**p**
	Broken plural	جمع تكسير	*ǧami' taksīr*	(قلم: أقلام) (*qalam: 'aqlām)* '(A pen: pens)'	**b**
	Plural of paucity	جمع قلة	*ǧami' qilla*^{h}	(حرف: أحرف) *(ḥarf: 'aḥruf)* (A letter: letters)	**m**
	Plural of multitude	جمع كثرة	*ǧami' kaṯra*^{h}	(حرف: حروف) *(ḥarf: ḥurūf)* (A letter: letters)	**j**
	Ultimate plural	منتهى الجموع	*munthā al-ǧumū'*	(مسجد: مساجد) *(masǧid: masāǧid)* (A mosque: mosques)	**u**
	Plural of plural	جمع الجمع	*ǧami' al-ǧami'*	(بيت: بُيوت، بُيوتات) *(bayt: buyūt, buyūtāt)* '(A home: homes)	**l**
	Undefined	غير مُعَرَّف	*ǧayr mu'arraf*	كَتَبَ الطَّالِبُ الدَّرْسَ *katab aṭ-ṭālibu ad-darasa* 'the student wrote the lesson'; كَتَبَ الطَّالِبَانِ الدَّرْسَ *katab aṭ-ṭāliban ad-darsa* 'the two students wrote the lesson'; كَتَبَ الطَّلَابُ الدَّرْسَ *kataba aṭ-ṭulābu ad-darsa* 'the students wrote the lesson'	**x**

Table A.10. Morphological feature of Person category attributes and their tags at position 9.

Position	Feature Name				Tag
9	**Person** الاسناد *al-'isnād*				
	First Person	المُتَكَلِّم	*al-mutakallim*	كَتَبتُ *katabtu*'I wrote'	**f**
	Second Person	المُخاطَب	*al-muḫāṭab*	كَتَبتُما *katabtumā* 'You wrote'	**s**
	Third Person	الغَائِب	*al-ǧā'ib*	كَتَبنَ *katabna*'They Wrote'	**t**

Table A.11. The morphological feature category of Inflectional Morphology attributes and their tags at position 10.

Position	Feature Name				Tag
10	**Inflectional Morphology** الصَّرف *aṣ-ṣarf*				
	Declined (noun) Conjugated (verb)	مُعرب	*mu'rab*	يَغيبُ *yaġību* 'Miss'	**d**
	Triptote / fully declined	مُعرب – منصرف	*mu'rab - munṣarif*	غائبٌ *ġā'ib* 'Absent'	**v**
	Non-declinable	مُعرب – ممنوع من الصرف	*mu'rab - mamnū' mina aṣ-ṣarf*	عُثمانُ *'uṯmānu* 'Othman'	**p**
	Invariable (v, n)	مبني	*mabnī*	هؤلاءِ *hā'ulā'i* 'Those' فَعَلَ *fa'ala* 'Did' لَيْتَ *layta* 'Wish'	**s**

Table A.12. The morphological feature of Case or Mood category attributes and their tags at position 11.

Position	Feature Name						Tag
11	**Case or Mood** الحالة الإعرابية للاسم أو الفعل *al-ḥālatu al-'i'rābiyyatu lil-'ism 'aw al-fi'l*						
	Nominative	Indicative	مرفوع	*marfū'*	يَكْتُبُ *yaktubu* 'He is writing'	الكتابُ *al-kitābu* 'The Book'	**n**
	Accusative	Subjunctive	منصوب	*manṣūb*	لن يَكْتُبَ *lan yaktuba* 'He will not write'	الكتابَ *al-kitāba* 'The Book'	**a**
	Genitive	--------	مجرور	*maǧrūr*	———	الكتابِ *al-kitābi* 'The Book'	**g**
	-------	Imperative or jussive	مجزوم	*maǧzūm*	لَمْ يَكْتُبْ *lam yaktub* 'He did not write'	———	**j**

Table A.13. The morphological feature category of Case and Mood Marks attributes and tags at position 12.

Position	Feature Name				Tag
12	**Case and Mood Marks** علامة الإعراب أو البناء *'alāmāt al-'i'rāb wa al-binā'*				
	ḍamma^h	الضمة / الضم	*al-ḍamma*^{h} / *al-ḍamm*	قدمَ الوزيرُ *qadima al-wazīru* 'The minister arrived' يَصومُ أحمد *yaṣūmu aḥmad* 'Ahmad fasts'	**d**
	fatḥa^h	الفتحة / الفتح	*al-fatḥa*^{h} / *al-fatḥ*	أكرمَ صالحٌ الوزيرَ *'akrama ṣāliḥun al-wazīra* 'Salih honored the minister' لنْ نَصبرَ على الذُّلِّ *lan naṣbira 'alā aḍ-ḍulli* 'We are not standing the humiliation'	**f**
	kasra^h	الكسرة	*al-kasra*^{h} / *al-kasr*	خلق الله السماواتِ والأرضَ *ḫalaqa allahu as-samāwāti wa al-'arḍa* 'God created the skys and the earth'	**k**

Position	Feature Name				Tag
12	**Case and Mood Marks** علامة الإعراب أو البناء *'alāmāt al-'i'rāb wa al-binā'*				
	sukūn (Silence)	السكون	*as-sukūn*	لَمْ أُسافِرْ إلى المدينة *lam 'usāfir 'ilā al-madīnati* 'I did not travel to the city'	**s**
	wāw	الواو	*al-wāw*	إذا جاءَكَ المنافقونَ *'iḏā ǧā'aka al-munāfiqūn* 'If the Hypocrites come to thee'	**w**
	'alif	الألف	*al-'alif*	التقى الفريقانِ *'iltaqā al-farīqān* 'The two teams have met'	**a**
	yā'	الياء	*al-yā'*	ذهبتُ إلى أخيكَ *ḏahbtu 'ilā 'aḫīka* 'I went to your brother'	**y**
	Inflectional *nūn*	ثبوت النون	*ṯubūt an-nūn*	المرشحان يتقدمانِ الإنتخابات *al-muraššḥāni yataqddamāni al-'intiḫābāt* ' Both candidates are ahead of elections'	**n**
	Deletion of *nūn*	حذف النون	*ḥaḏf an-nūn*	المسلمون لنْ يَصبروا على الذُّلِّ *al-muslimūn lan yasbirū 'alā aḏ-ḏulli* 'Muslims will not stand to the humiliation'	**o**
	Deletion of vowel letter	حذف حرف العلَّة	*ḥaḏf ḥarf al-'illa*^{h}	لَمْ يَخشَ صالحٌ إلا الله *lam yaḫša ṣāliḥ 'illā allāha* 'Salih does not afraid except of God'	**v**

Table A.14. The morphological feature of Definiteness category attributes and their tags at position 13.

Position	Feature Name				Tag
13	**Definiteness** المَعْرِفَة والنَّكِرَة *al-ma'rifa*^{ti} *wa an-nakira*^{ti}				
	Definiteness	مَعْرِفَة	*ma'rifa*^{h}	الكتاب *al-kitāb* 'The book'	**d**
	Indefiniteness	نَكِرَة	*nakira*^{h}	كتاب *kitāb* 'A book'	**i**

Table A.15. The morphological feature of Voice category attributes and their tags at position 14.

Position	Feature Name				Tag
14	**Voice** المَبْني لِلمَعْلُوم و المَبْني لِلمَجْهُول *al-mabnī lil-ma'lūm wa al-mabnī lil-maǧhūl*				
	Active voice	مَبْني للمَعْلُوم	*mabnī lil-ma'lūm*	كَتَبَ *kataba* 'He wrote'	**a**
	Passive voice	مَبْني للمَجْهُول	*mabnī lil-maǧhūl*	كُتِبَ *kutiba* 'It was written'	**p**

Table A.16. The morphological feature of Emphasized and Non-emphasized category attributes and their tags at position 15.

Position	Feature Name				Tag
15	**Emphasized and Non-emphasized** المُؤَكَّد وغيرُ المُؤَكَّد *al-mu'akkad wa ǧayir al-mu'akkad*				
	Emphatic verb	فعل مُؤَكَّد	*fi'l mu'akkad*	لأكتُبَنَّ *la'aktubanna* 'I will write'	**n**
	Non-emphatic verb	فعل غَيْر مُؤَكَّد	*fi'l ǧayr mu'akkad*	أكْتُبُ *'aktubu* 'I am writing'	**m**

Table A.17. The morphological feature of Transitivity category attributes and their tags at position 16.

Position	Feature Name				Tag
16	**Transitivity** اللازم والمتعدي *al-lāzim wa al-muta'adi*				
	Intransitive	لازِم	*lāzim*	نامَ الولدُ *nāma al-waladu* 'The boy slept'	**i**
	Singly transitive	مُتعدٍّ إلى مَفعُول واحِد	*muta'addī 'ilā maf'ūlin wāḥid*	فَتَحَ الرجلُ البابَ *fataḥa ar-raǧulu al-bāba* 'The man opened the door'	**o**
	Doubly transitive	مُتعدٍّ إلى مَفْعُولَين	*muta'addī 'ilā maf'ūlayn*	أعطاه دينارًا *'a'ṭāhu dīnār^{an}* 'He gave him a dinar'	**b**
	Triply transitive	مُتعدٍّ إلى ثَلاثَة مَفاعِيل	*muta'addī 'ilā ṯalāṯati mafā'īl*	أنبأتُهُ الخبرَ صحيحًا *'anb'tuhu al-ḫabara ṣaḥīḥ^{an}* 'I announced him the correct news'	**t**

Table A.18. Morphological feature category of Rational attributes and their tags at position 17.

Position	Feature Name				Tag
17	**Rational** العاقل وغير العاقل *al-'āqil wa ǧayir al-'āqil*				
	Rational	عاقِل	*'āqil*	قَرَأَ *qara'a* 'Read'	**h**
	Irrational	غَيْر عَاقِل	*ǧayr 'āqil*	نَبَحَ *nabaḥa* 'Bark'	**n**

Table A.19. The morphological feature of Declension and Conjugation category attributes and their tags at position 18.

Position	Feature Name				Tag
18	**Declention and Conjugation** التَّصريف *at-taṣrīf*				
	Non-Inflected (n, v)	غير مُتصَرِّف	*ǧayr mutaṣarrif*	هُوَ *huwa* 'Him'	**n**
	Primitive / Concrete noun	مُتَصَرِّف – جامِد– اسم ذات	*mutaṣarrif – ǧāmid – 'ism ḏāt*	شَجرةٌ *šaǧara^{h}* 'A tree'	**t**
	Primitive / Abstract noun	مُتَصَرِّف – جامِد– اسم معنى	*mutaṣarrif – ǧāmid – 'ism ma'nā*	ذكاءٌ *ḏakā^{un}* 'Intelligence'	**a**
	Inflected / Derived noun	مُتَصَرِّف – اسم مُشْتَقٌّ	*mutaṣarrif – 'ism mušṭaqq*	كتابٌ *kitāb^{un}* 'A book' مكتبةٌ *maktaba^{tun}* 'A library'	**d**
	Non-conjugated / restricted to the perfect	فعل جامِد– ملازم للماضي	*fi'l ǧāmid- mulāzim lil-maḍī*	نَعِمَ *na'ima* 'Be happy'	**p**
	Non-conjugated / restricted to the imperfect	فعل جامِد– ملازم للمضارع	*fi'l ǧāmid- mulāzim lil-muḍāri'*	يَهِيطُ *yahīṭu* 'Scream'	**c**
	Non-conjugated / restricted to the imperative	فعل جامِد– ملازم للأمر	*fi'l ǧāmid- mulāzim lil-'amr*	هَبْ *hab* 'Suppose'	**i**
	Conjugated / fully conjugated verb	مُتَصَرِّف – فعل تام التَّصريف	*mutaṣarrif – fi'l tām at-taṣarīf*	يكتبُ *yaktubu* 'He is' writing'	**v**
	Conjugated / partially conjugated verb	مُتَصَرِّف – فعل ناقص التَّصريف	*mutaṣarrif –fi'l nāqiṣ at-taṣarīf*	كادَ *kāda* 'Close; near or almost'	**m**

Table A.20. The morphological feature of Unaugmented and Augmented category attributes and their tags at position 19.

Position	Feature Name				Tag
19	**Unaugmented and Augmented** المُجَرَّد والمَزِيد *al-muǧarrad wa al-mazīd*				
	Unaugmented	مُجَرَّد	*al-muǧarrad*	كَتَبَ *kataba* 'Wrote'	**s**
	Augmented by one letter	مَزِيد بِحَرف	*mazīd bi ḥarf*	كَاتَبَ *kātaba* 'Wrote'	**a**
	Augmented by two letters	مَزِيد بِحَرفَيْن	*mazīd bi ḥarfayn*	اكْتَتَبَ *'iktataba* 'Subscribed'	**b**
	Augmented by three letters	مَزِيد بِثَلاثَةِ أحرف	*mazīd bi ṯalāṯat 'aḥruf*	اسْتَكْتَبَ *'istaktaba* 'Registered'	**t**
	Augmented by four letters	مَزِيد بأربعة أحرف	*mazīd bi 'arba'ati 'aḥruf*	استقبال *'istiqbāl* 'Reception'	**q**

Table A.21. The morphological feature of Number of Root Letters category attributes and their tags at position 20.

Position	Feature Name				Tag
20	**Number of Root Letters** عَدَد أحْرُف الجَذْر *adad 'aḥruf al-ǧaḏr*				
	Triliteral	ثُلاثِي	*ṯulāṯī*	ك ت ب *k t b* 'Wrote'	**t**
	Quadriliteral	رُباعِي	*rubā'ī*	د ح ر ج *d ḥ r ǧ* 'Rolled'	**q**
	Quinqueliteral	خُماسِي	*ḫumāsī*	ز ب ر ج د *z b r ǧ d* 'Chrysolite'	**f**

Table A.22. The morphological feature of Verb Root category attributes and their tags at position 21.

Position	Feature Name			Tag
21	**Verb Root** بُنية الفعل *bunya^{tu} al-fi'l*			
	Intact verb	صحيح	*saḥīḥ*	**a**
	Doubled verb	مضعف	*muḍa''af*	**b**
	Initially-hamzated verb	مهموز الفاء	*mahmūz al-fā'*	**c**
	Initially-hamzated and doubled verb	مهموز الفاء مضعّف	*mahmūz al-fā' muḍa''af*	**d**
	Initially and finally hamzated verb	مهموز الفاء ومهموز اللام	*mahmūz al-fā' wa mahmūz al-lām*	**e**
	Medially-hamzated verb	مهموز العين	*mahmūz al-'ayn*	**f**
	Finally-hamzated verb	مهموز اللام	*mahmūz al-lām*	**g**
	wāw-initial verb	مثال واوي	*miṯāl wāwī*	**h**
	wāw-initial and doubled verb	مثال واوي مضعف	*miṯāl wāwī muḍa''af*	**i**
	wāw- initial and medially-hamzated verb	مثال واوي مهموز العين	*miṯāl wāwī mahmūz al-'ayn*	**j**
	wāw-initial and finally-hamzated verb	مثال واوي مهموز اللام	*miṯāl wāwī mahmūz al-lām*	**k**
	yā'-initial verb	مثال يائي	*miṯāl yā'ī*	**l**
	yā'-initial and doubled verb	مثال يائي مضعف	*miṯāl yā'ī muḍa''af*	**m**
	yā'- initial and medially-hamzated verb	مثال يائي مهموز العين	*miṯāl yā'ī mahmūz al-'ayn*	**n**
	Hollow with *wāw*	أجوف واوي	*'aǧwaf wāwī*	**o**
	Hollow with *wāw* and initially-hamzated verb	أجوف واوي مهموز الفاء	*'aǧwaf wāwī mahmūz al-fā'*	**p**

Position	Feature Name			Tag
21	**Verb Root** بُنية الفعل *bunya^{tu} al-fi'l*			
	Hollow with *wāw* and finally-hamzated verb	أجوف واوي مهموز اللام	*'aǧwaf wāwī mahmūz al-lām*	**q**
	Hollow with *yā'*	أجوف يائي	*'aǧwaf yā'ī*	**r**
	Hollow with *yā'* and initially-hamzated verb	أجوف يائي مهموز الفاء	*'aǧwaf yā'ī mahmūz al-fā'*	**s**
	Hollow with *yā'* and finally-hamzated verb	أجوف يائي مهموز اللام	*'aǧwaf yā'ī mahmūz al-lām*	**t**
	Defective with *wāw* verb	ناقص واوي	*nāqiṣ wāwī*	**u**
	Defective with *wāw* and initially-hamzated verb	ناقص واوي مهموز الفاء	*nāqiṣ wāwī mahmūz al-fā'*	**v**
	Defective with *wāw* and medially-hamzated verb	ناقص واوي مهموز العين	*nāqiṣ wāwī mahmūz al-'ayn*	**w**
	Defective with *yā'* verb	ناقص يائي	*nāqiṣ yā'ī*	**x**
	Defective with *yā'* and initially-hamzated verb	ناقص يائي مهموز الفاء	*nāqiṣ yā'ī mahmūz al-fā'*	**y**
	Defective with *yā'* and medially-hamzated verb	ناقص يائي مهموز العين	*nāqiṣ yā'ī mahmūz al-'ayn*	**z**
	Adjacent doubly-weak verb	لفيف مقرون	*lafīf maqrūn*	*****
	Adjacent doubly-weak and initially-hamzated verb	لفيف مقرون مهموز الفاء	*lafīf maqrūn mahmūz al-fā'*	**$**
	Separated doubly-weak verb	لفيف مفروق	*lafīf mafrūq*	**&**
	Separated doubly-weak and medially-hamzated verb	لفيف مفروق مهموز العين	*lafīf mafrūq mahmūz al-'ayn*	**@**

Table A.23. The morphological feature of Noun Finals category attributes and their tags at position 22.

Position	Feature Name				Tag
22	**Noun Finals** أقسام الأسم تبعاً للفظ آخره *'aqsām al-'ismi tib^{an} li-lafẓi 'āẖirhi*				
	Sound noun	الاسم صحيح الآخر	*al-'ism ṣaḥīḥ al-'āir*	جبل *ǧabal* 'Mountain' نهر *nahr* 'River' درهم *dirham* 'Dirham (currency)'	**s**
	Semi-sound noun	الاسم شبه الصحيح	*al-'ism šibh aṣ-ṣaḥīḥ*	دَلْو *dalw* 'Bucket' بهو *bahw* 'Hall'	**i**
	Noun with shortened ending	الاسم المقصور	*al-'ism al-maqṣūr*	بُشْرَى *bušrā* 'Glad tidings'	**t**
	Noun with extended ending	الاسم الممدود	*al-'ism al-mamdūd*	سَمَاء *samā'* 'Sky'	**e**
	Noun with curtailed ending	الاسم المنقوص	*al-'ism al-manqūṣ*	القَاضِي *al-qāḍī* 'The' judge'	**c**
	Noun with deleted ending	الاسم محذوف الآخر	*al-'ism maḥḏūf al-'āẖir*	يَدْ *yad* 'Hand', سَنَة *sana^{h}* 'Year', and لُغَة *luġa^{h}* Language'.	**d**

Notes

1. We would like to thank all the participants of the workshop of morphological analyzer experts for Arabic language, organized by the Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul Aziz City for Science and Technology (KACST) and the Arabic Language Academy, Damascus, Syria, 26–28 April 2009, for their suggestions and agreement on the classification of morphological features of Arabic words. We want to thank Mr. Marwan Al-Bawab (Member of the Arabic

Language Academy in Damascus, Syria), for his valuable advice and comments on designing the SALMA – Tag Set of Arabic to ensure that it adheres to traditional Arabic grammar.

We would like to thank Professor James Dickins, Head of Arabic and Middle Eastern Studies, University of Leeds, Leeds, UK, for standardizing the English translations of Arabic grammar terms in this paper, and for his efforts in reviewing the paper.

2. http://acopost.sourceforge.net/
3. http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=POS
4. http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=FLBJPOS
5. http://www.nltk.org/
6. http://opencog.org/wiki/RelEx
7. http://nlp.ipipan.waw.pl/Spejd/
8. http://beta.visl.sdu.dk/cg3.html
9. Automatic Mapping Among Lexico-Grammatical Annotation Models (AMALGAM) http://www.comp.leeds.ac.uk/amalgam/amalgam/amalghome.htm
10. http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpora.htm
11. LDC Arabic POS tagging documentation http://www.ircs.upenn.edu/arabic/Jan03release/POS-info.txt
12. MorphoChallenge 2009 Qur'an Gold Standard http://www.cis.hut.fi/morphochallenge2009/datasets.shtml
13. EAGLES Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES document EAG-TCWG-MAC/R. http://www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/annotate.ps.gz
14. http://www.ircs.upenn.edu/arabic/pos.html
15. The Annotation Manual of the SALMA Tag Set for Arabic http://www.comp.leeds.ac.uk/sawalha/tagset.html
16. According to Wright's (1986) classifications. Ryding (2005) classifies nouns according to gender into two classes: *masculine* and *feminine*, and the '*dual gender noun*' is mentioned in a footnote on page 119.
17. Recently the word نائب *nā'ib* is being used for both masculine and feminine as the regular feminine form of this word نائبة *nā'iba*^{h} means 'disaster' which is not suitable to indicate *feminine parliament member*.

References

Al-Ghalayyni 2005. جامع الدروس العربية "*Jami' Al-Duroos Al-Arabia*". Saida – Lebanon: Al-Maktaba Al-Asriyiah "المكتبة العصرية".

Al-Shamsi, Fatima and Guessoum, Ahmad 2006. A Hidden Markov Model-Based POS Tagger for Arabic. *8es Journées internationales d'Analyse statistique des Données Textuelles*.

Al-Sulaiti, Latifa and Atwell, Eric 2004. Designing and developing a corpus of contemporary Arabic *TALC 2004: Proceedings of the sixth Teaching And Language Corpora conference 92–93*.

Al-Sulaiti, Latifa and Atwell, Eric 2005. Extending the Corpus of Contemporary Arabic. *Corpus Linguistics conference 2005* University of Birmingham, UK.

Al-Sulaiti, Latifa and Atwell, Eric 2006. The design of a corpus of contemporary Arabic? *International Journal of Corpus Linguistics* 11: 135–171.

ALECSO 2008. Sarf - Arabic Morphology System. The Arab League Educational, Cultural and Scientific Organization (ALECSO).

Alqrainy, Shihadeh 2008. A Morphological-Syntactical Analysis Approach For Arabic Textual Tagging. Leicester, UK: De Montfort University.

Altabbaa, Mohammad, Al-Zaraee, Ammar and Shukairy, Mohammad Arif 2010. An Arabic Morphological Analyzer and Part-Of-Speech Tagger Qutuf 'قُطُوف'. Damascus: Arab International University.

Atwell, Eric 2008. Development of tag sets for part-of-speech tagging. In Anke Ludeling and Merja Kytö (eds.), *Corpus linguistics: an international handbook, volume 1*, Mouton de Gruyter. 501–526.

Atwell, Eric, Demetriou, George, Hughes, John, Schiffrin, Amanda, Souter, Clive and Wilcock, Sean 2000. A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal, International Computer Archive of Modern and medieval English, Bergen* 24: 7–23.

Bamman, David and Crane, Gregory 2008. Building a Dynamic Lexicon from a Digital Library. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008)* Pittsburgh.

Brill, Eric 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics* 21: 543–565.

Cachia, Pierre 1973. *The monitor: a dictionary of Arabic grammatical terms : Arabic-English, English-Arabic / compiled by Pierre Cachia*. Beirut: Librairie du Liban.

Dahdah, Antonie 1987. *A dictionary of Arabic Grammer in Charts and Tables* "معجم قواعد اللغة العربيه في جداول ولوحات". Beirut, Lebanon: Librairie du Liban publisher.

Dahdah, Antonie 1993. *A dictionary of Arabic Grammatical nomenclature Arabic – English* "معجم لغة النحو العربي عربي-انكليزي". Beirut, Lebanon: Librairie du Liban publishers.

Diab, Mona, Hacioglu, Kadri and Jurafsky, Daniel 2004. Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks. *Proceedings of HLT-NAACL*.

Diab, Mona T. 2007. Towards an Optimal POS Tag Set for Arabic Processing. *Proc RANLP*.

Dror, Judith, Shaharabani, Dudu, Talmon, Rafi and Wintner, Shuly 2004. Morphological Analysis of the Qur'an. *Literary and Linguistic Computing* 19: 431–452.

Duh, Kevin and Kirchhoff, Katrin 2005. POS Tagging of Dialectal Arabic: A Minimally Approach. *ACL-05, Computational Approaches to Semitic Languages Workshop Proceedings 55–62*. University of Michigan Ann Arbor, Michigan, USA.

Dukes, Kais, Atwell, Eric and Habash, Nizar 2011. Supervised Collaboration for Syntactic Annotation of Quranic Arabic. *Language Resources and Evaluation Journal (LREJ). Special Issue on Collaboratively Constructed Language Resources*.

Dukes, Kais, Atwell, Eric and Sharaf, Abdul-Baquee. M. 2010. Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank. *Language Resources and Evaluation Conference (LREC 2010)* Valletta, Malta.

Dukes, Kais and Habash, Nizar 2010. Morphological Annotation of Quranic Arabic. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* Valletta, Malta, 19–21 May 2010: European Language Resources Association (ELRA).

Elliott, John and Atwell, Eric 2000. Is anybody out there?: the detection of intelligent and generic language-like features. *JBIS: Journal of the British Interplanetary Society* 53: 7–23.

Freeman, Andrew 2001. Brill's POS Tagger and a Morphology Parser for Arabic. *NAACL 2001 Student Research Workshop, Lancaster University*.

Habash, Nizar, Faraj, Reem and Roth, Ryan 2009. Syntactic Annotation in Columbia Arabic Treebank. *2nd International Conference on Arabic Language Resources & Tools MEDAR 2009* Cairo, Egypt.

Habash, Nizar and Rambow, Owen 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* Ann Arbor, Michigan: Association for Computational Linguistics.

Habash, Nizar and Roth, Ryan M. 2009. CATiB: The Columbia Arabic Treebank. *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* 221–224. Suntec, Singapore.

Hamada, Salwa 2010. مقترح لمعايير وضوابط تقييم المحلّلات الصرفية Evaluation of the Arabic Morphological Analyzers? *Proceedings of The Sixth International Computing science Conference ICCA* Hammamet, Tunisia.

Harmain, Harmain M. 2004. Arabic Part-of-Speech Tagging. *The Fifth Annual U.A.E. University Research Conference* United Arab Emirates.

Johansson, Stig, Atwell, Eric, Garside, Roger and Leech, Geoffrey 1986. *The Tagged LOB Corpus*. Bergen, Norway: Norwegian Computing Centre for the Humanities.

Khoja, Shereen 2001. APT: Arabic Part-of-Speech Tagger. *Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001)* Carnegie Mellon University, Pittsburgh, Pennsylvania.

Khoja, Shereen 2003. APT: An Automatic Arabic Part-of-Speech Tagger. Lancaster, UK: Lancaster University.

Khoja, Shereen, Garside, Porger and Knowles, Gerry 2001. A tagset for the morphosynactic tagging of Arabic. *Corpus Linguistics 2001* Lancaster University, Lancaster, UK.

Leech, Geoffrey and Wilson, Andrew 1999. Standards for Tagsets. In Hans van Halteren (ed.), *Syntactic Wordclass Tagging*. KLUWER Academic Publishers. 55–80.

Maamouri, Mohamed and Bies, Ann 2004. Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*.

Marsi, Erwin, Bosch, Antal van den and Soudi, Abdelhadi 2005. Memory-based morphological analysis generation and part-of-speech tagging of Arabic. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages* 1–8. Ann Arbor: Association for Computational Linguistics.

Monachini, Monica and Calzolari, Nicoletta 1996. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to European languages. Pisa, Italy: Istituto di Linguistica Computazionale -CNR.

Ryding, Karin C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge University Press.

Sawalha, Majdi 2011. Open-source Resources and Standards for Arabic Word Structure Analysis. *School of Computing* Leeds: University of Leeds.

Sawalha, Majdi and Atwell, Eric 2009a. Linguistically Informed and Corpus Informed Morphological Analysis of Arabic. *Proceedings of the 5th International Corpus Linguistics Conference CL2009* Liverpool, UK.

Sawalha, Majdi and Atwell, Eric 2009b.
توظيف قواعد النحو والصرف في بناء محلل صرفي للغة العربية (Adapting Language Grammar Rules for Building Morphological Analyzer for Arabic Language). *Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League*

Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City for Science and Technology (KACST) and Arabic Language Academy. Damascus, Syria.

Sawalha, Majdi and Atwell, Eric 2010. Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text. *Language Resource and Evaluation Conference LREC 2010* Valleta, Malta: European Language Resources Association (ELRA).

Schmid, Helmut and Laws, Florian 2008. Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. *COLING'08* Manchester,UK.

Talmon, Rafi and Wintner, Shuly 2003. Morphological Tagging of the Qur'an. *In Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, an EACL'03 Workshop* Budapest, Hungary.

Teahan, Bill 1998. Modeling English Text. *Department of Computer Science* New Zealand: University of Waikato.

Teufel, Simone, Schmid, Helmut, Heid, Ulrich and Schiller, Anne 1996. Study of the relation between tagsets and taggers. Stuttgart, Germany Institut für maschinelle Sprachverarbeitung, Universität Stuttgart

Tlili-Guiassa, Yamina 2006. Hybrid Method for Tagging Arabic Text. *Journal of Computer Science* 2, 245–248.

Voutilainen, Atro 2003. Part-of-Speech Tagging. In Ruslan Mitkov (ed.), *The Oxford Handbook of Computational Linguistics* 219–232. Oxford University Press.

Wright, W. 1996. *A Grammar of the Arabic Language, Translated from the German of Caspari, and Editted with Numerous Additions and Corrections.* Beirut: Librairie du Liban.

Zibri, Chiraz Ben Othmane, Torjmen, Aroua and Ahmad, Mohamed Ben 2006. An Efficient Multi-agent system Combining POS-Taggers for Arabic Texts. *CICLing 2006*, LNCS 3878.

Zolfagharifard, Ellie 2009. Anti-terror technology tool uses human logic. *The Engineer*.

Authors' addresses: (Majdi Sawalha[1] and Eric Atwell[2])
[1]Computer Information Systems Department
King Abdullah II School of Information Technology
The University of Jordan
Amman 11942
Jordan
E-mail: sawalha.majdi@gmail.com

[2]I-AIBS Institute for Artificial intelligence and Biological Systems
School of Computing
University of Leeds
Leeds LS2 9JT
United Kingdom
E-mail: e.s.atwell@leeds.ac.uk