A tagset for the morphosyntactic tagging of Arabic

Shereen Khoja   Roger Garside   Gerry Knowles
Computing Department   Computing Department   Linguistics Department
Lancaster University   Lancaster University   Lancaster University

1. Introduction

Principles for allocating grammatical parts-of-speech for words have been around for fourteen centuries, and the principles used today vary very little from the principles used then.Yet there is no publicly available grammatically annotated Arabic corpus. The first step for the annotation of corpora is the compilation of a tagset that can accurately describe and cover the whole of the language. Since the principles for describing and tagging Arabic already exists, these have been used to compile the tagset that we describe here.

The Arabic tagset we describe does not follow the EAGLES recommendations for the morphosyntactic annotation of corpora, but this is to be expected since Arabic is very different from the languages for which EAGLES was designed, and belongs to the Semitic family rather than the Indo-European one. Following a normalised tagset and the EAGLES recommendations would not capture some of Arabic's relevant information, such as the jussive mood of the verb and the dual number that are integral to Arabic. Another important aspect of Arabic is inheritance, where all subclasses of words inherit properties from the classes they are derived.

Before describing the tagset, we will give a brief summary of the Arabic language followed by the justification for our proposed tagset. We will also compare the Arabic tagset with the EAGLES guidelines and describe other similar work in this area. Finally we will present an example of Arabic text that uses this tagset, and present the full tagset with examples.

2. A brief description of the Arabic language

Arabic is a Semitic language, and its basic feature is that most of its words are built up from, and can be analysed down to roots. The exceptions to this rule are common nouns and particles. Around 64% percent of roots are composed of three consonants (sometimes called radicals)(Thalouth et al., 1986), and there are also roots consisting of two, four and five consonants.

Words are built up from these roots by following fixed patterns[1] that add prefixes, suffixes and infixes to the word. For example, the Arabic word *mdrs*[2] "teacher" is built up from the root *drs* "study" by following the pattern *m123*[3], where the number 1 is replaced by the first consonant in the root, number 2 is replaced by the second consonant in the root, and number 3 is replaced by the third consonant in the root. Following the same pattern the word *mdkhn*[4] "smoker" is built up from the root *dkhn* "smoke".

Prefixes and suffixes can be added to the words that have been built up from roots (these are called stems) to add number or gender (among other things) to these words. For example, adding the Arabic suffix *wn* to the word *mdrs* "teacher" gives the word *mdrswn* "teachers" which indicates the masculine plural. The feminine plural of that same word is *mdrsat*.

This system that studies how words are constructed from roots, and describes the patterns they follow is called *ṣrf*[5], which in English could be called derivational morphology (Wehr, 1974).

---

[1] These are sometimes called measures or binyan, and in Arabic are called *awzan*.

[2] The transliteration notation used throughout this paper is the same one described in (Haywood, 1988; pp. 3, 4). Transliterated Arabic words will be written in italics, and the English translation will be placed within quotation marks next to the Arabic word. These transliterated words do not include notations for the short vowels. Short vowels are described at the end of section 2.

[3] In Arabic, the three numbers are represented by the Arabic consonants (f, ', l), but we have used the numbers instead to make it clearer for non-Arabic readers.

[4] The Arabic consonant represented here by *kh* has the sound of the ch found in the Scottish word loch.

[5] ṣ here is an emphatic and stronger s.

Arabic contains three genders (much like English), masculine, feminine and neuter. It also contains three persons, one to describe the speaker, one to describe the person being addressed and one to describe the person that is not present. Arabic differs from Indo-European languages in that it contains three numbers instead of the more common two numbers. So as well as singular and plural, there is also the dual that is used for describing the actions of two people. All these attributes are taken into account when constructing the tagset.

When teaching Arabic grammar to students, Arabic grammarians and linguists have always used the Arabic system of inflectional morphology called *I'rab*. Basically, students are presented with a sentence, and for each word they must provide a description of it based on its position within the sentence. For example, given the sentence *drs alwld* "the boy studied", students would have to say that the first word is the indeclinable, indicative, perfect verb, while the second word is the nominative subject.

We have based our Arabic tagset on this system of language teaching. As such we distinguish between the three moods of the verb (indicative, subjunctive, and jussive), and the three case forms of the noun (nominative, accusative, and genitive) (Haywood, 1998).

The moods of the verb and the cases of the nouns should all be clear apart from the jussive mood of the verb. The jussive is needed to express a command in the first and third person. This mood is realised in Arabic by rejecting the final vowel and is sometimes called the apocopated imperfect. The subjunctive and jussive moods are only applied to the imperfect verb since perfect verbs are always in the indicative (Wright, 1974).

The Arabic alphabet consists of 28 consonants but 3 of these are also used as long vowels. These are the same as the "ea" in beak, the "oo" in root, and the "a" in tar. Arabic also contains three short vowels that are equivalent to the "i" in sit, the "oo" in foot, and the "a" in bat. The short vowels do not normally appear in Arabic text, but when they do appear, (such as in children's books), they are represented by diacritics that are written above or below the consonants they follow.

For example, in normal (unvowelised) Arabic text, the word pronounced '*darasa*[6] "he studied" would actually be written in Arabic using the three consonants only as *drs*. This way the word could then also be pronounced '*durisa*' "It was studied" or '*dars*' "lesson" where the second vowel is the silent or absent vowel. It is left to the reader of Arabic to decide which of the words is actually intended by looking at the context it appears in.

When describing the tagset in the following sections the only difference between some forms of the word (e.g. subjunctive and jussive moods of the verb) is the diacritic. The method used to determine the tag of the words in this case would be to use the context of that word in the sentence, the same technique used by the human reader.

3. Justification for the proposed morphosyntactic tagset

The morphosyntactic tagset that is described here follows the tagging system that has been used for around fourteen centuries by all students of Arabic (both young and old).

Arabic grammarians describe Arabic as being derived from three main categories: noun, verb and particle. In fact there is a famous poem that actually starts with the verse: the word or language is one of three, noun, verb and particle. It is from these three main categories that the rest of the language is derived. For most native speakers of Arabic and also many learners of the language, this is a very natural way of describing the language. Early on at school children are taught that the basis of language is this threesome, and in fact when they come to learn a second language (such as English) they try to apply this theory to that language.

---

[6] This is a transcription not a transliteration as we include the short vowels that are not normally present in the Arabic text. It is sometimes necessary for us to transcribe words in this paper to indicate the changes in the vowels. Transcribed words will be in italics and single quotes, while transliterated words will be in italics only.

But the most important aspect of this system of describing Arabic is that all the subclasses of these three main classes inherit properties from the parent classes. For example, all subclasses of the noun inherit the *tnwyn* "nunation"[7] when in the indefinite which is one of the main properties of the noun. Therefore, when compiling the tagset, it is important to highlight this factor of the Arabic language so that it is obvious which categories inherit which properties. The diagram in Appendix C illustrates this inheritance.

Arabic is very rich in categorising words, and contains classes for almost every form of word imaginable. For example, there are classes for nouns of instruments, nouns of place and time, nouns of activity and so on (Mace, 1999). If we tried to use all the subclasses described by Arabic grammarians, the size of the tagset would soon reach more than two or three hundred tags. For this reason, we have chosen only the main classes and subclasses. But because of the way all the classes inherit from others, it would be quite simple to extend this tagset to include more subclasses, or simplify it and make it smaller.

Arabic could be described using an Indo-European tagset, but by doing so we would lose many of its properties. Let us try forcing Arabic on to an Indo-European tagset for a moment. One of the main categories in Indo-European language is the adjective, but Arabic does not include adjectives as one of its main parts-of-speech. An adjective in Arabic is actually a noun that happens to describe something, and in the same way that you would say in English that a word is an object or subject, in Arabic you would say that it is the adjective *ṣifa* and the item being described is the thing being described *almwṣwf*. By using an Indo-European tagset, we lose the fact that adjectives inherit "nunation" from nouns.

The point is that Arabic grammar has been studied for centuries, and the principles of describing the language already exist. Since so much knowledge is readily available, it is logical to derive our tagset from this wealth of information. The alternative to this is to base the Arabic tagset on an Indo-European one, but by doing this we may lose a lot of the information that an Arabic tagset would give us. Also, by moulding Arabic to fit an Indo-European language, we might distort the way Arabic is perceived by its native speakers.

4. Description of the proposed morphosyntactic tagset of Arabic

In this section we are going to describe the tagset we have compiled for the Arabic language. The tagset contains 177 tags, 103 nouns, 57 verbs, 9 particles, 7 residual, and 1 punctuation.

The tagset follows the inheritance property described in section 3, and a diagram to further clarify this is shown in Appendix C. A full description of each of the tags and examples of Arabic words that take those tags now follows.

The five main categories for words are:

1. N [noun]	2. V [verb]	3. P [particle]
4. R [residual]	5. PU [punctuation]	

The residual category contains foreign words, mathematical formulae and numbers. The punctuation category contains all punctuation symbols, both Arabic and foreign such as (?,!"‹؟).

The subcategories of noun are:

1.1. C [common]	1.2. P [proper]	1.3. Pr [pronoun]
1.4. Nu [numeral]	1.5. A [adjective]	

Adjectives are nouns that describe the aspects of an object. Adjectives inherit the properties of nouns, so they take "nunation" when in the indefinite and can take the definite article when definite. For example, *alwld ṣghyr* "The small boy" contains the adjective *ṣghyr* "small". This adjective can take the

---

[7] Nunation is the doubling of the vowels at the end of nouns and all subclasses of nouns when they are indefinite. This doubling has the effect of adding a final "n" to the pronunciation, so that, *kitabu* becomes *kitabun* "book".

definite article as in '*darasa alwaladu aḻṣaghyr'* "the small boy studied", and it can also have "nunation" as in '*ḥasanu ṣaghyrun'* "Hassan is small".

Examples of these subcategories include:
- Singular, masculine, accusative, common noun such as *ktab* "book" in the sentence '*akhadha alwaladu kitaban'* "the boy took a book".
- Singular, masculine, genitive, common noun such as *ktab* "book" in the sentence '*darastu min kitabin'* "I studied from a book".
- Singular, feminine, nominative, common noun such as *ktab* "book" in the sentence '*hadhihi madrasatun'* "this is a school".

The subcategories of the pronoun are:

> 1.3.1. P [personal]   1.3.2. R [relative]   1.3.3. D [demonstrative]

The personal pronouns can be detached words such as *hwa* "he", or attached to a word in the form of a clitic. The attached pronouns can be attached to nouns to indicate possession, to verbs as direct object, or attached to prepositions such as *fyh* "in it".

Some examples of pronouns include:
- Third person, singular, masculine, personal pronoun such as *hwa* "him".
- Singular, feminine, demonstrative pronoun such as *hdhh* "this".

The subcategories of the relative pronoun are:

> 1.3.2.1. S [specific]   1.3.2.2. C [common]

Examples of relative pronouns include:
- Dual, feminine, specific, relative pronoun such as *alltan* "who".
- Plural, masculine, specific, relative pronoun such as *alldhyn* "who".
- Common, relative pronoun such as *'men'* "who".

The subcategories of the numeral are:

> 1.4.1. Ca [cardinal]   1.4.2.O [ordinal]   1.4.3. Na [numerical adjective]

Numerical adjectives describe the number of sides to a shape, for example *thmany* "octagonal". They also indicate a pair or couple when describing two people as in *thna`y*.

Examples of numerals include:
- Singular, masculine, nominative, indefinite cardinal number such as '*arba'atun'* "four".
- Singular, masculine, nominative, indefinite ordinal number such as '*rabi'un'* "fourth".
- Singular, masculine, numerical adjective such as '*ruba'iyun'* "of four".

The linguistic attributes of nouns that have been used in this tagset are:

(i)	Gender:	M [masculine]	F [feminine]	N [neuter]
(ii)	Number:	Sg [singular]	Pl [plural]	Du [dual]
(iii)	Person:	1 [first]	2 [second]	3 [third]
(iv)	Case:	N [nominative]	A [accusative]	G [genitive]
(v)	Definiteness:	D [definite]	I [indefinite]	

Verbs are categorised into three main parts:

> 1. P [perfect]   2. I [imperfect]   3. Iv [imperative]

The definition of perfect verbs not only includes (i) the equivalent of English past tense verbs (i.e. to describe acts completed in some past time) but also (ii) describes acts which at the moment of speaking have already been completed and remain in a state of completion, (iii) describes a past act that often took place or still takes place (i.e. commentators are agreed (have agreed and still agree)), (iv)

describes an act which is just completed at the moment by the very act of speaking it (I sell thee this), and (v) describes acts which is certain to occur that it can be described as having already taken place (mostly used in promises, treaties and so on) (Wright, 1974).

The imperfect does not in itself express any idea of time, it merely indicates a begun, incomplete, or enduring existence either in present, past or future time. While the imperative verbs order or ask for something to be done in the future.

Examples of verbs include:
- First person, singular, neuter, perfect verb *'kasartu'* "I broke".
- First person, singular, neuter, indicative, imperfect verb *'aksiru'* "I break"
- Second person, singular, masculine, imperative verb *'aksir'* "Break!"

The verbal attributes that have been used in our tagset are:

(i)	Gender:	M [masculine]	F [feminine]	N [neuter]
(ii)	Number:	Sg [singular]	Pl [plural]	Du [dual]
(iii)	Person:	1 [first]	2 [second]	3 [third]
(iv)	Mood:	I [indicative]	S [subjunctive]	J [jussive]

The two most notable verbal attributes that are fundamental to Arabic but do not normally appear in Indo-European tagsets are the dual number, and the jussive mood that have been described previously in section 2. A more detailed comparison between the Arabic tagset and EAGLES is described in section 5 below.

The subcategories of particle are:

1.1. Pr [prepositions]	1.2. A [adverbial]	1.3. C [conjunctions]
1.4. I [interjections]	1.5. E [exceptions]	1.6 N [negatives]
1.7. A [answers]	1.8. X [explanations]	1.9. S [subordinates]

Examples of particles include:
- Prepositions *fy* "in"
- Adverbial particles *swf* "shall"
- Conjunctions *w* "and"
- Interjections *ya* "you"
- Exceptions *swa* "Except"
- Negatives *lm* "Not"
- Answers *ajl* "yes"
- Explanations *ay* "that is"
- Subordinates *lw* "if"

5. Comparison between the Arabic tagset and EAGLES

The EAGLES tagset is traditional and conventionally based on Latin. There does not seem to be a rationale for why it contains the categories it does. On the other hand, Arabic has always been described as having certain categories, and that the subcategories inherit properties from the parent categories. The Arabic tagset thus captures generalisations that cannot be found in any other language.

The Arabic tagset has three major categories (five if residual and punctuation are included), while EAGLES describes eleven major categories (thirteen if residual and punctuation are included) (Leech, 1996). The eleven major categories in EAGLES are adjective, pronoun/determiner, article, adverb, adpositions, conjunctions, interjections and unique. The unique category is applied to categories with a unique or very small membership, which are 'unassigned' to any of the standard part-of-speech tags.

Although two of the EAGLES major categories are major categories in Arabic (i.e. noun and verb), the other categories are described as subcategories of the major categories in the Arabic tagset. For example pronouns and numerals are subcategories of the nouns, while conjunctions and interjections

are subcategories of the particle. One major category in EAGLES is the article. This is not described in Arabic as a category at all, and definiteness is just a linguistic feature that is realised by the definite article.

Another difference between the Arabic tagset and EAGLES is the number. Arabic marks words not just for singular and plural but also for dual. There is also the difference in the verb tenses or aspects. Verbs are defined in ancient Arabic grammar as being perfect, imperfect or imperative. This classification is an important part of Arabic, and trying to mould Arabic verbs to fit the traditional past, present and future tenses of Indo-European languages would be very unnatural.

Arabic verbs also contain a mood that is not described anywhere in the EAGLES tagset. This is the jussive mood. The purpose of the jussive is to express a command in the first or third person[8] as in '*ḥeena yaḥḍuru yalbas thiyaban naẓyfatan*' which means "when he attends, let him (he must) wear clean clothes", where the jussive is the word *'yalbas'* "wear". Also, there is no negative imperative in Arabic, so the negative particle followed by the jussive is used in its place, such as '*la taktub'* "do not write", where the jussive is the word *'taktub'* "write".

The way that definiteness is handled in Arabic is quite different to the way it is handled in EAGLES, and this difference is apparent when comparing the tagsets. In Arabic nouns are marked for definiteness by the prefix that is the definite article, unlike in English where the article itself can be definite or indefinite. In Arabic there is no indefinite article, and instead "nunation" is used.

6.   Related work

A tagset that has been similarly derived is discussed briefly in (El-Kareh et al., 1999), where the authors state that:

> "..Arabic words are classifies into three main classes, namely, verb, noun and particle. Verbs are subclassified into three subclasses; nouns into forty six subclasses (e.g. Active participle, Passive participle, Exaggeration pattern, Adjectival noun, Adverbial noun, Infinitive noun, Common noun, Pronoun, Quantifier, etc.) and particles into twenty three subclasses (e.g. additional, resumption, Indefinite, Conditional, Confirmational, Prohibition, Imperative, Optative, Reasonal, Dubious, etc.)"

The authors then go on to add the linguistic features since they do not include them in the main part-of-speech tags. The linguistic features they add include perfectness, transitivity and voice for verbs, number and gender for nouns, and agglutination[9] for particles.

When describing our tagset we have decided to include these linguistic features in the part-of-speech tag, though not as a subcategory of the main tags. So, although we would say that perfect verbs are a subcategory of the verb tag, we say that masculine is an attribute of that verb (not a subcategory).

El-Kareh et al (1999) do not describe the tagset in any more detail, and it is not clear from the brief description above what some of the categories are.

7.   Other improvements or additions to the tagset

The current Arabic tagset can be extended to include many other features such as transitivity and voice[10] for verbs, and derivation[11] for nouns. The noun category can also be subcategorised to include many of the subclasses described by Arabic grammarians that lend meaning to the word, for example, nouns of instrument and nouns of place and time.

Since Arabic contains three different kinds of plurals, these could also be included in the tagset. The three types of plurals are: the masculine sound plural, the feminine sound plural, and the broken plural. The first two are recognised in Arabic morphologically by suffixes, while the last, is derived by the following of fixed patterns.

---

[8] A command in the second person is expressed using the imperative verb (see section 4).
[9] Agglutination here means whether the particle is a detached particle or one that can only appear attached to a word.
[10] Arabic has two voices, the active and the passive.
[11] Nouns in Arabic can be derived from other nouns or they can be derived from verbs.

8. Summary and conclusions

In this paper we describe a morphosyntactic tagset that is derived from the ancient Arabic grammar. The tagset does not follow the traditional Indo-European tagset that is based on Latin but is instead based on the Semitic tradition of analysing language.

We have shown that in the Arabic tagset, all the subcategories inherit properties from the parent categories thus capturing the generalisations of the language.

Appendix A shows the result of applying the tagset to a sample of real Arabic text, which proves that the tagset is sufficient in describing Arabic in some detail, and the full tagset can be found in Appendix B and contains transcribed Arabic examples, as well as the English translation, for all the tags.

9. References

El-Kareh S, Al-Ansary S 2000 An Arabic Interactive Multi-feature POS Tagger. In *Proceedings of the, ACIDCA conference*, Monastir, Tunisia, pp 204-210.

Haywood J, Nahmad H 1998 *A new Arabic grammar of the written language*. London, Lund Humphries.

Leech G, Wilson A 1996 *Recommendations for the Morphosyntactic Annotation of Corpora* EAGLES Report EAG-TCWG-MAC/R. http://www.ilc.pi.cnr.it/EAGLES96/annotate/annotate.html

Mace J 1999 *Arabic verbs and essential* grammar. London, Hodder & Stoughton.

Thalouth B, Al-danan A 1986 A comprehensive Arabic morphological analyser-generator. In *Proceedings of Arab summer school*, Syria.

Wehr H 1980 *A dictionary of modern written Arabic*. London, Macdonald and Evens Ltd.

Wright, W 1974 *A Grammar of the Arabic* Language. Lebanon, Librairie du Liban.

Appendix A.   Applying the tagset to some real Arabic text

This paragraph is an extract from Al-Jazirah newspaper dated 1/1/1998. Note that Arabic is written from right to left, hence the change in paragraph direction.

بعث_VPSg3M خادم_NCSgMNI الحرمين_NCDuMAD الشريفين_NCDuMGD الملك_NCSgMND فهد_NP بن_
_PPr الى_NCSgFGI تهنئة_NCSgFNI برقية_NP سعود_R آل_NCSgMAD العزيز_NCSgMAI عبد_NCSgMNI
فخامة_NCSgFGI الرئيس_NCSgMGD الكسندر_RF كواسنيفيسكي_RF رئيس_NCSgMNI جمهورية_NCSgFGI
بولندا_RF بمناسبة_PPr_NCSgFGI اليوم_NCSgMAD الوطني_NCSgMND لبلاده_NCSgMND_NPrPSg3M_NCPlFGI_PPr
وأعرب_PC_VPSg3M الملك_NCSgMND المفدى_NCSgMAD باسمه_NPrPSg3M_NCSgMGI_PPr وباسم_
السعودية_NCSgFGD العربية_NCSgFGD المملكة_NCSgFGD وحكومة_PC_NCSgFGI شعب_NCSgMGI PC_PPr_NCSgM
لفخامته_NCSgMAI متمنياً NCPlMND التهاني NCSgFNI اخلص عن_PPr NCSgFGD
ولشعب_PC_NCSgFGD والسعادة_NCSgFGD الصحة_NCSgFGD دوام_NCSgMNI PPr_NCSgFGI_NPrPSg3M
PU_. النائم_NCSgMN الازدهار_NCSgMND الصديق_NCSgMND بولندا_RF PC_PPr_NCSgMGI

Appendix B.   The full tagset

Tag	Description of word category	Example (Arabic)	Transcription	Translation
NCSgMNI	Singular, masculine, nominative, indefinite common noun	كتابٌ	*kitabun*	book
NCSgMAI	Singular, masculine, accusative, indefinite common noun	كتاباً	*kitaban*	book
NCSgMGI	Singular, masculine, genitive, indefinite common noun	كتابٍ	*kitabin*	book
NCSgMND	Singular, masculine, nominative, definite common noun	الكتابُ	*alkitabu*	the book
NCSgMAD	Singular, masculine, accusative, definite common noun	الكتابَ	*alkitaba*	the book
NCSgMGD	Singular, masculine, genitive, definite common noun	الكتابِ	*alkitabi*	the book
NCSgFNI	Singular, feminine, nominative, indefinite common noun	مدرسةٌ	*madrasatun*	school
NCSgFAI	Singular, feminine, accusative, indefinite common noun	مدرستاً	*madrasatan*	school
NCSgFGI	Singular, feminine, genitive, indefinite common noun	مدرسةٍ	*madrasatin*	school
NCSgFND	Singular, feminine, nominative, definite common noun	المدرسةُ	*almadrasatu*	the school

NCSgFAD	Singular, feminine, accusative, definite common noun	المدرسةَ	*almadrasata*	the school
NCSgFGD	Singular, feminine, genitive, definite common noun	المدرسةِ	*aladrasati*	the school
NCDuMNI	Dual, masculine, nominative, indefinite common noun	كتابان	*kitaban*	two books
NCDuMAI	Dual, masculine, accusative, indefinite common noun	كتابين	*kitabain*	two books
NCDuMGI	Dual, masculine, genitive, indefinite common noun	كتابين	*kitabain*	two books
NCDuMND	Dual, masculine, nominative, definite common noun	الكتابان	*alkitaban*	the two books
NCDuMAD	Dual, masculine, accusative, definite common noun	الكتابين	*alkitabain*	the two books
NCDuMGD	Dual, masculine, genitive, definite common noun	الكتابين	*alkitabain*	the two books
NCDuFNI	Dual, feminine, nominative, indefinite common noun	مدرستان	*madrasatan*	two books
NCDuFAI	Dual, feminine, accusative, indefinite common noun	مدرستين	*madrasatain*	two schools
NCDuFGI	Dual, feminine, genitive, indefinite common noun	مدرستين	*madrasatain*	two schools
NCDuFND	Dual, feminine, nominative, definite common noun	المدرستان	*almadrasatan*	the two schools
NCDuFAD	Dual, feminine, accusative, definite common noun	المدرستين	*almadrasatain*	the two schools
NCDuFGD	Dual, feminine, genitive, definite common noun	المدرستين	*almadrasatain*	the two schools
NCPlMNI	Plural, masculine, nominative, indefinite common noun	كتبٌ - مسلمون	*muslimoon – kutubun*	Muslims – books
NCPlMAI	Plural, masculine, accusative, indefinite common noun	كتبا - مسلمين	*muslimeen – kutuban*	Muslims – books
NCPlMGI	Plural, masculine, genitive, indefinite common noun	كتبٍ - مسلمين	*muslimeen – kutubin*	Muslims – books
NCPlMND	Plural, masculine, nominative, definite common noun	الكتبُ - المسلمون	*almuslimoon – alkutubu*	the Muslims – the books
NCPlMAD	Plural, masculine, accusative, definite common noun	الكتبَ - المسلمين	*aluslimeen – alkutuba*	the Muslims – the books
NCPlMGD	Plural, masculine, genitive, definite common noun	الكتبِ - المسلمين	*almuslimeen – alkutubi*	the Muslims – the books
NCPlFNI	Plural, feminine, nominative, indefinite common noun	مسلماتٌ–مدارسٌ	*madarisun – muslimaatun*	schools – Muslims
NCPlFAI	Plural, feminine, accusative, indefinite common noun	مسلماتاً–مدارساً	*madarisan – muslimaatan*	schools – Muslims
NCPlFGI	Plural, feminine, genitive, indefinite, common noun	مسلماتٍ–مدارس	*madarisin – muslimaatin*	schools – Muslims
NCPlFND	Plural, feminine, nominative, definite common noun	المسلماتُ–المدارسُ	*almadarisu – almuslimaatu*	the schools – the Muslims
NCPlFAD	Plural, feminine, accusative, definite common noun	المسلماتَ–المدارسَ	*almadarisa – almuslimaata*	the schools – the Muslims
NCPlFGD	Plural, feminine, genitive, definite common noun	المسلماتِ–المدارسِ	*almadarisi – almuslimaati*	the schools – the Muslims
NP	Proper noun	شيرين–جده	*Jiddah – Shyryn*	Jeddah – Shereen
NPrPSg1	First person, singular, neuter, personal pronoun	كتابي - ضربني–أنا	*ana- kitaabee – darabanee*	Me – my book – he hit me
NPrPSg2M	Second person, singular, masculine, personal pronoun	كتابكَ–أنتَ	*anta – kitaabuka*	You – your book
NPrPSg2F	Second person, singular, feminine, personal pronoun	كتابك–أنتِ	*anti – kitaabuki*	You – your book
NPrPSg3M	Third person, singular, masculine, personal pronoun	هو - كتابه	*kitaabahu – huwa*	His book – him
NPrPSg3F	Third person, singular, feminine, personal	هي - كتابها	*kitaabuhaa – hiya*	Her book –

				her
NPrPDu2	Second person, dual, neuter, personal pronoun	كتابكما–أنتما	*antumaa – kitaabakumaa*	You two – your book
NPrPDu3	Third person, dual, neuter, personal pronoun	كتابهما–هما	*humaa – kitaabahumaa*	Those two – their book
NPrPPl1	First person, plural, neuter, personal pronoun	كتابنا–نحن	*naḥnu – kitaabunaa*	Us – our book
NPrPPl2M	Second person, plural, masculine, personal pronoun	كتابكم–أنتم	*antum – kitaabakum*	You – your book
NPrPPl2F	Second person, plural, feminine, personal pronoun	كتابكن–أنتن	*antunna – kitaabakunna*	You – your book
NPrPPl3M	Third person, plural, masculine, personal pronoun	كتابهم–هم	*hum – kitaabahum*	Them – their book
NPrPPl3F	Third person, plural, feminine, personal pronoun	هن - كتابهن	*kitaabahunna – hunna*	Their book – them
NPrRSSgM	Singular, masculine, specific, relative pronoun	الذي	*allathi*	Who
NPrRSSgF	Singular, feminines, specific, relative pronoun	التي	*allati*	Who
NPrRSDuM	Dual, masculine, specific, relative pronoun	اللذين–اللذان	*alladhaani – alladhaini*	Who
NPrRSDuF	Dual, feminine, specific, relative pronoun	اللتين–اللتان	*allataani – allataini*	Who
NPrRSPlM	Plural, masculine, specific, relative pronoun	اللذين - اللائي	*allaiy – alladheena*	Who
NPrRSPlf	Plural, feminine, specific, relative pronoun	اللاتي - اللائي	*allaaiy - allatee*	Who
NPrRC	Common, relative pronoun	مهما– ما –من	*men – maa – mahmaa*	Who – what
NPrDSgM	Singular, masculine, demonstrative pronoun	ذلك– ذاك – ذا –هذا	*hadhaa – dhaa – dhaaka – dhaalika*	This – that
NPrDSgF	Singular, feminine, demonstrative pronoun	– تلك – ذي – ذه – هذي –هذه تيك–تاك	*haadhihi – haadhee – dhih – dhy – tilka – taaka – teeka*	This – that
NprDDuM	Dual, masculine, demonstrative pronoun	– هذين – ذانك – ذان –هذان ذينك–ذين	*haadhani – dhaani – dhaanika – hadhaini –dhaini*	This – that
NprDDuF	Dual, feminine, demonstrative pronoun	تين – هتين – تانك – تان –هتان تينك–	*haatani – taani- taanika – haataini*	This – that
NPrDPl	Plural, neutral, demonstrative pronoun	أولائك – أولاء – أولى –هؤلاء أولاك– أولالك –	*haaolaai – olaa- olaaika- olaalika – olaaka*	Those
NNuCaSgM	Singular, masculine, cardinal number	أربع	*arba'*	Four
NNuCaSgF	Singular, feminine, cardinal number	أربعة	*arba'a*	Four
NNuOrSgM	Singular, masculine, ordinal number	رابع	*raabi'*	Fourth
NNuOrSgF	Singular, feminine, ordinal number	رابعة	*raabia*	Fourth
NNuNaSgM	Singular, masculine, numerical adjective	رباعي	*rubaa'y*	Of four
NNuNaSgF	Singular, feminine, numerical adjective	رباعية	*rubaa'iya*	Of four
NACSgMNI	Singular, masculine, nominative, indefinite adjective	سعيدٌ	*sa'ydun*	happy
NACSgMAI	Singular, masculine, accusative, indefinite adjective	سعيدا	*sa'ydan*	happy
NACSgMGI	Singular, masculine, genitive, indefinite adjective	سعيدٍ	*sa'ydin*	happy
NACSgMND	Singular, masculine, nominative, definite adjective	السعيدُ	*alsa'ydu*	the happy
NACSgMAD	Singular, masculine, accusative, definite adjective	السعيدَ	*alsa'yda*	the happy
NACSgMGD	Singular, masculine, genitive, definite adjective	السعيدِ	*alsa'ydi*	the happy
NACSgFNI	Singular, feminine, nominative, indefinite adjective	سعيدةٌ	*sa'ydatun*	happy

NACSgFAI	Singular, feminine, accusative, indefinite adjective	سعيدتًا	*sa'ydatan*	happy
NACSgFGI	Singular, feminine, genitive, indefinite adjective	سعيدةٍ	*sa'ydatin*	happy
NACSgFND	Singular, feminine, nominative, definite adjective	السعيدةُ	*alsa'ydatu*	the happy
NACSgFAD	Singular, feminine, accusative, definite adjective	السعيدةَ	*alsa'ydata*	the happy
NACSgFGD	Singular, feminine, genitive, definite adjective	السعيدةِ	*alsa'ydati*	the happy
NACDuMNI	Dual, masculine, nominative, indefinite adjective	سعيدان	*sa'ydan*	two happy
NACDuMAI	Dual, masculine, accusative, indefinite adjective	سعيدين	*sa'ydain*	two happy
NACDuMGI	Dual, masculine, genitive, indefinite adjective	سعيدين	*sa'ydain*	two happy
NACDuMND	Dual, masculine, nominative, definite adjective	السعيدان	*alkitaban*	the two happy
NACDuMAD	Dual, masculine, accusative, definite adjective	السعيدين	*alsa'ydain*	the two happy
NACDuMGD	Dual, masculine, genitive, definite adjective	السعيدين	*alsa'ydain*	the two happy
NACDuFNI	Dual, feminine, nominative, indefinite adjective	سعيدتان	*sa'ydatan*	two happy
NACDuFAI	Dual, feminine, accusative, indefinite adjective	سعيدتين	*sa'ydatain*	two happy
NACDuFGI	Dual, feminine, genitive, indefinite adjective	سعيدتين	*sa'ydatain*	two happy
NACDuFND	Dual, feminine, nominative, definite adjective	السعيدتان	*alsa'ydatan*	the two happy
NACDuFAD	Dual, feminine, accusative, definite adjective	السعيدتين	*alsa'ydatain*	the two happy
NACDuFGD	Dual, feminine, genitive, definite adjective	السعيدتين	*alsa'ydatain*	the two happy
NACPlMNI	Plural, masculine, nominative, indefinite adjective	سعيدون	*sa'ydoon*	happy
NACPlMAI	Plural, masculine, accusative, indefinite adjective	سعيدين	*sa'ydeen*	happy
NACPlMGI	Plural, masculine, genitive, indefinite adjective	سعيدين	*sa'ydeen*	happy
NACPlMND	Plural, masculine, nominative, definite adjective	السعيدون	*alsa'ydoon*	the happy
NACPlMAD	Plural, masculine, accusative, definite adjective	السعيدين	*alsa'ydeen*	the happy
NACPlMGD	Plural, masculine, genitive, definite adjective	السعيدين	*alsa'ydeen*	the happy
NACPlFNI	Plural, feminine, nominative, indefinite adjective	سعيداتٌ	*sa'ydaatun*	happy
NACPlFAI	Plural, feminine, accusative, indefinite adjective	سعيداتاً	*sa'ydaatan*	happy
NACPlFGI	Plural, feminine, genitive, indefinite, adjective	سعيداتٍ	*sa'ydaatin*	happy
NACPlFND	Plural, feminine, nominative, definite adjective	السعيداتُ	*alsa'ydaatu*	the happy
NACPlFAD	Plural, feminine, accusative, definite adjective	السعيداتَ	*alsa'ydaata*	the happy
NACPlFGD	Plural, feminine, genitive, definite adjective	السعيداتِ	*alsa'ydaati*	the happy
VPSg1	First person, singular, neuter, perfect verb	كسرتُ	*kasartu*	I broke
VPSg2M	Second person, singular, masculine, perfect verb	كسرتَ	*kasarta*	You broke
VPSg2F	Second person, singular, feminine, perfect verb	كسرتِ	*kasarti*	You broke
VPSg3M	Third person, singular, masculine, perfect verb	كسر	*kasara*	He broke
VPSg3F	Third person, singular, feminine, perfect verb	كسرتْ	*kasarat*	She broke
VPDu2	Second person, dual, neuter, perfect verb	كسرتما	*kasartumaa*	You (two) broke

VPDu3M	Third person, dual, masculine, perfect verb	كسرا	*kasaraa*	They (two) broke
VPDu3F	Third person, dual, feminine, perfect verb	كسرتا	*kasarataa*	They (two) broke
VPPl1	First person, plural, neuter, perfect verb	كسرنا	*kasarnaa*	We broke
VPPl2M	Second person, plural, masculine, perfect verb	كسرتم	*kasartum*	You broke
VPPl2F	Second person, plural, feminine, perfect verb	كسرتن	*kasartunna*	You broke
VPPl3M	Third person, plural, masculine, perfect verb	كسروا	*kasaroo*	They broke
VPPl3F	Third person, plural, feminine, perfect verb	كسرن	*kasarna*	They broke
VISg1I	First person, singular, neuter, indicative, imperfect verb	أكسرُ	*aksiru*	I break
VISg1S	First person, singular, neuter, subjunctive, imperfect verb	أكسرَ	*aksira*	I break
VISg1J	First person, singular, neuter, jussive, imperfect verb	أكسرْ	*aksir*	I break
VISg2MI	Second person, singular, masculine, indicative, imperfect verb	تكسرُ	*taksiru*	You break
VISg2MS	Second person, singular, masculine, subjunctive, imperfect verb	تكسرَ	*taksira*	You break
VISg2MJ	Second person, singular, masculine, jussive, imperfect verb	تكسرْ	*taksir*	You break
VISg2FI	Second person, singular, feminine, indicative, imperfect verb	تكسرين	*taksiryna*	You break
VISg2FS	Second person, singular, feminine, subjunctive, imperfect verb	تكسري	*taksiry*	You break
VISg2FJ	Second person, singular, feminine, jussive, imperfect verb	تكسري	*taksiry*	You break
VISg3MI	Third person, singular, masculine, indicative, imperfect verb	يكسرُ	*yaksiru*	He breaks
VISg3MS	Third person, singular, masculine, subjunctive, imperfect verb	يكسرَ	*yaksira*	He breaks
VISg3MJ	Third person, singular, masculine, jussive, imperfect verb	يكسرْ	*yaksir*	He breaks
VISg3FI	Third person, singular, feminine, indicative, imperfect verb	تكسرُ	*taksiru*	She breaks
VISg3FS	Third person, singular, feminine, subjunctive, imperfect verb	تكسرَ	*taksira*	She breaks
VISg3FJ	Third person, singular, feminine, jussive, imperfect verb	تكسرْ	*taksir*	She breaks
VIDu2I	Second person, dual, neuter, indicative, imperfect verb	تكسران	*taksiraani*	You break
VIDu2S	Second person, dual, neuter, subjunctive, imperfect verb	تكسرا	*taksiraa*	You break
VIDu2J	Second person, dual, neuter, jussive, imperfect verb	تكسرا	*taksiraa*	You break
VIDu3MI	Third person, dual, masculine, indicative, imperfect verb	يكسران	*yaksiraani*	They break
VIDu3MS	Third person, dual, masculine, subjunctive, imperfect verb	يكسرا	*yaksiraa*	They break
VIDu3MJ	Third person, dual, masculine, jussive, imperfect verb	يكسرا	*yaksiraa*	They break
VIDu3FI	Third person, dual, feminine, indicative, imperfect verb	يكسران	*yaksiraan*	They break
VIDu3FS	Third person, dual, feminine, subjunctive, imperfect verb	يكسرا	*yaksiraa*	They break
VIDu3FJ	Third person, dual, feminine, jussive, imperfect verb	يكسرا	*yaksiraa*	They break
VIPl1I	First person, plural, neuter, indicative, imperfect verb	نكسرُ	*naksiru*	We break
VIPl1S	First person, plural, neuter, subjunctive, imperfect verb	نكسرَ	*naksira*	We break
VIPl1J	First person, plural, neuter, jussive, imperfect verb	نكسرْ	*naksir*	We break
VIPl2MI	Second person, plural, masculine, indicative, imperfect verb	تكسرون	*taksiroon*	You break

VIPl2MS	Second person, plural, masculine, subjunctive, imperfect verb	تكسروا	*taksiroo*	You break
VIPl2MJ	Second person, plural, masculine, jussive, imperfect verb	تكسروا	*taksiroo*	You break
VIPl2FI	Second person, plural, feminine, indicative, imperfect verb	تكسرن	*taksirna*	You break
VIPl2FS	Second person, plural, feminine, subjunctive, imperfect verb	تكسرن	*taksirna*	You break
VIPl2FJ	Second person, plural, feminine, jussive, imperfect verb	تكسرن	*taksirna*	You break
VIPl3MI	Third person, plural, masculine, indicative, imperfect verb	يكسرون	*yaksiroon*	They break
VIPl3MS	Third person, plural, masculine, subjunctive, imperfect verb	يكسروا	*yaksiroo*	They break
VIPl3MJ	Third person, plural, masculine, jussive, imperfect verb	يكسروا	*yaksiroo*	They break
VIPl3FI	Third person, plural, feminine, indicative, imperfect verb	يكسرن	*yaksirna*	They break
VIPl3FS	Third person, plural, feminine, subjunctive, imperfect verb	يكسرن	*yaksirna*	They break
VIPl3FJ	Third person, plural, feminine, jussive, imperfect verb	يكسرن	*yaksirna*	They break
VIvSg2M	Second person, singular, masculine, imperative verb	أكسرْ	*aksir*	Break!
VIvSg2F	Second person, singular, feminine, imperative verb	أكسري	*aksiry*	Break!
VIvDu2	Second person, dual, neuter, imperative verb	أكسرا	*aksiraa*	Break!
VIvPl2M	Second person, plural, masculine, imperative verb	أكسروا	*aksroo*	Break!
VIvPl2F	Second person, plural, feminine, imperative verb	أكسرن	*aksirna*	Break!
PPr	Prepositions	في- مع - من - ل -ك	*ka – li – min – ma'a – fy*	As – for – from – with – in
PA	Adverbial particles	لن- سوف – ثم -إذا	*idhaa – thumma – swf – ln*	And then – then – shall – won't
PC	Conjunctions	و- حتى -ف	*f – ḫta – w*	So – so – and
PI	Interjections	أيتها-يا	*ya – aytha*	You
PE	Exceptions	سوى-غير	*ghyr – swa*	Except
PN	Negatives	لم- لا	*la – lm*	Not
PW	Answers	لا-أجل	*la – ajl*	No – yes
PX	Explanations	أي	*ay*	That is
PS	Subordinates	لو-ما	*ma – lw*	If
RF	Residual, foreign	روجور	*rwjwr*	Roger
RM	Residual, mathematical	÷	*/*	/
RN	Residual, number	3	*3*	3
RD	Residual, day of the week	الاثنين	*alithnyn*	Monday
Rmy	Residual, month of the year	محرم	*mhrm*	muharram
RA	Residual, abbreviation	واس	*was*	e.g.
RO	Residual, other	آل	*Aal*	
PU	Punctuation	؟	*?*	?

Appendix C. The tagset hierarchy