TECHNICAL REPORT

No. CCLS-13-03

Title: Annotation Guidelines for Arabic Nominal Gender, Number, and Rationality

Authors: Nizar Habash and Sarah Alkuhlani

Version 11 Nov-18-2011

Annotation Guidelines for Arabic Nominal Gender, Number, and Rationality

Nizar Habash and Sarah Alkuhlani
Columbia University

The annotation task we define here is focused on information relevant to modeling Arabic nominal gender and number computationally. First we define the various facts regarding number and gender in Modern Standard Arabic and then we present the task guidelines and examples.

I. Arabic Gender and Number Facts

1. Arabic nouns inflect for gender (masculine/feminine) and number (singular/dual/plural). We will not address the dual in this task.
2. We distinguish two types of gender/number: form-based gender/number (لفظي) and functional (logical) gender/number (معنوي). For many nouns, the form-based and functional values are the same but not always. For example, كاتب and خليفة are both masculine singular functionally although خليفة has a feminine ending. Similarly, كاتبة and حامل (as in pregnant) are both feminine singular, although حامل looks like a masculine noun. Other examples include words like كتبة and وزراء, which are both masculine plural functionally, but have feminine singular forms. We are concerned only with the functional gender/number (الجنس\العدد المعنوي).
3. Arabic adjectives agree with the nouns they modify in gender and number EXCEPT for plural irrational (non-human, غير عاقل) nouns, which always take feminine singular adjectives.

مكتب مهم	كاتب مهم
مكاتب مهمة	كتاب مهمون
مكتبة مهمة	كاتبة مهمة
مكتبات مهمة	كاتبات مهمات

   This **<u>does not</u>** mean that the adjective مهمة is plural. It only means that it can be used to modify some plural nouns. Note that there are nouns that are semantically rational/human but morphologically not: شعب nation/people (شعوب مهمة).
4. Number quantification in Arabic has many complex rules. We focus on a couple of rules that interact with gender and number in interesting ways.
   a. Numbers 3-10 always take a plural noun: ثلاثة رجالٍ.
   b. Numbers over 10 always take a singular noun:
      أربعة وعشرون رجلاً, مائة رجلٍ , etc.
   c. Numbers 3-10 have masculine and feminine forms: ستة / ست. Masculine numbers are used with nouns whose <u>singular form</u> is

feminine and feminine numbers are used with nouns whose singular form is masculine. For example: خمسة رجالٍ, خمس سيدات, خمس نملاتٍ, but خمسة سجلاتٍ (singular is سجلّ, which is masculine).

5. Some nouns in Arabic are semantically plural but morphologically singular. They do not agree morphologically like plurals: نملٌ كبيرٌ. If نملٌ was plural, it should take a feminine singular adjective since it is irrational: نملٌ كبيرةٌ, which is incorrect. Words like نمل are mass nouns that have a singular form related to it: نملة.
6. Some nouns, which are often thought of as collective, are plural morphologically: عرب (العرب المتفهمون, **not** العرب المتفهم).

II. Task Guidelines

The task is to annotate examples of Arabic nouns with four features: functional gender, functional number, and rationality.

a. **Functional gender** can be M (masculine), F (feminine), B (both), or U (unknown).
b. **Functional number** can be S (singular), D (dual), P (plural), B[1] (S, D or P) or U (unknown)
c. **Rationality** can be R (rational), I (irrational), N (not marked), U (unknown)

The unknown value is only used when the annotator is not sure what the correct answer is, e.g., the annotator does not understand the word, or it is not clear how to apply the tests for different decisions. These cases will be checked later by a supervisor.

Entries in the lexicon you will be annotating will look as follows:

###	كَاتِب	كَاتِب	noun	author, writer
###	كَاتِبَة	كَاتِب	noun	author, writer
###	كَاتِبُون	كَاتِب	noun	authors, writers
###	كَاتِبَات	كَاتِب	noun	authors, writers
###	كُتَّاب	كَاتِب	noun	authors, writers

The entries are automatically clustered to bring together related forms. The first column is the label that you need to modify. The second is the word form of interest. The word is only cited in the nominative form (اسم مرفوع) and with no definite article (ال التعريف). This does not mean exclude other forms. Think of كاتبون, for example, as representing الكاتبون, الكاتبين, كاتبين, etc. The third is the basic lemma/vocable (المفردة). The fourth is the part-of-speech. The last column is the

[1] Value B for number is only used for a small set of closed classes. See section (IV).

English translation. After finishing the annotation, the label column would look like this:

MSR	كَاتِب	كَاتِب	noun	author, writer
FSR	كَاتِبَة	كَاتِب	noun	author, writer
MPR	كَاتِبُون	كَاتِب	noun	authors, writers
FPR	كَاتِبَات	كَاتِب	noun	authors, writers
MPR	كُتَّاب	كَاتِب	noun	authors, writers

Next are some tests for determining the correct value for each feature.

a. Functional number
   i. If the word can be quantified by a number (3-10) **and/or** modified by a plural adjective, it is **plural**, e.g.,
      خمسة رجال **مهمين** ، **خمس** كاتبات **مهمات** ، **خمس** نملات كبيرة
      else the word is singular.
   ii. Some words might look like a plural but they are not, e.g., تشريفاتي. This word is a singular adjective. Its plural form is تشريفاتيون\تشريفاتيات.
   iii. If you are not comfortable making a decision, choose U (unknown).

b. Functional gender
   i. The gender of a **singular** word is the same as the gender of the adjective that can modify it, e.g., رجل كبير, سيدة كبيرة, مكتب كبير, مكتبة كبيرة**.**
   ii. Nouns that can be both masculine and feminine are marked as B (both), e.g., صباح مهذب\مهذبة (أسم علم), طريق طويل\طويلة.
   iii. The gender of a **plural** word is the same as the gender of its **singular**. So, we turn the plural to a singular first to determine gender: مكاتب → مكتب → M; امتحانات → امتحان → M; etc.
   iv. Some nouns are the plural form of more than one singular (with **different** genders): e.g., انشقاقات is plural of انشقاقة (F) and of انشقاق (M); طُرُق is plural of طريق (B) and of طريقة (F) (each of which has an additional unique plural – طرقات and طرائق, respectively). In such cases, assign the value B.
   v. If you are not comfortable making a decision, choose U (unknown).

c. Rationality
   i. If the adjective of the plural noun is feminine singular, the noun is irrational, e.g., امتحانات صعبة.
   ii. The rationality of singular nouns is determined by turning them into the plural first: كلب → كلاب : كلاب خطيرة → Irrational.
   iii. Adjectives take the value N (not marked).

Words such as أهل, عائلة, كتائب have different contexts: الكتائب تعلن [FPI] / الكتائب يعلن [MSI] / الكتائب يعلنون [MPR].

Here is our view: Arabic allows a lot of elision; in fact the constructions above are:

الكتائب تعلن [FPI]
الكتائب يعلن [MSI]
أفراد حزب الكتائب يعلنون [MPR]

So, we will go with the simplest reading when multiple readings can be used. So, عائلة is FSI
العائلة تريد ; العائلات تريد (FSI ; FPI) but أفراد العائلة يريدون

iv. There are some cases with lemma ambiguity. For example, هلتون can refer to the hotel chain or a member of the Hilton family. For these cases, go with the most common reading for that word.

i. If you are not comfortable making a decision, choose U (unknown).

d. Errors

i. The lexicon you will be annotating may contain some errors.

ii. For missing English translations, add the translation preceded by “ADD:” in the English column. For incorrect translations, write “DEL:” just before the word to remove.

iii. For missing entries, add the entry by copying the full line from one of the existing forms first and then modifying it. Add the sequence “ADD:” at the beginning of your label.

iv. For wrong entries, place “ERR” in the label.

Here is an example: let’s pretend the entry you got is this:

###	كَاتِب	كَاتِب	noun	author
###	كَاتِبَة	كَاتِب	noun	author
###	كَاتِبُون	كَاتِب	noun	dancers
###	كَاتِبَات	كَاتِب	noun	dancers
###	كواتبة	كَاتِب	noun	dancers

Here is how you may correct it

MSR	كَاتِب	كَاتِب	noun	author, ADD:writer
FSR	كَاتِبَة	كَاتِب	noun	author, ADD:writer
MPR	كَاتِبُون	كَاتِب	noun	DEL:dancers, ADD:authors, ADD:writers
FPR	كَاتِبَات	كَاتِب	noun	DEL:dancers, ADD:authors, ADD:writers
ERR	كواتبة	كَاتِب	noun	dancers
ADD:MPR	كُتَّاب	كَاتِب	noun	authors, writers

III. Examples

G-N-R = gender-number-rationality

	English	G-N-R
كاتب	Author/writer (male)	MSR
كاتبة	Author/writer (female)	FSR
كاتبون	Authors/writers (male)	MPR
كاتبات	Authors/writers (female)	FPR
كُتّاب	Authors/writers (male)	MPR
كتبة	Authors/writers (male)	MPR
سيد	Gentleman /Mister	MSR
سيدة	Lady	FSR
سيدات	Ladies	FPR
سادة	Gentlemen	MPR
حامل	Carrying (masc.sing.)	MSN
حامل	Pregnant (sing.)	FSR
حوامل	Pregnant (plur.)	FPR
خليفة	Caliph	MSR
خلفاء	Caliphs	MPR
مكتب	office	MSI
مكاتب	offices	MPI
امتحان	exam	MSI
امتحانات	exams	MPI
حكاية	story	FSI
حكايات	stories	FPI
قصة	story	FSI
قصص	stories	FPI
نمل	Ants (uncountable)	MSI
تشريفاتي	Ceremonial (masc.sing.)	MSN
تشريفاتيات	Ceremonial (fem.plur.)	FPN
جيش	army	MSI
جيوش	armies	MPI
تمرة	A palm date	FSI
تمرات	Some palm dates	FPI
تمر	Palm dates	MSI
تمور	Types of palm dates	MPI
أهل	Extended family	MBR
أهالي	Extended families	MPR
العرب	Arabs	MPR

		عشرون X	مئة X	عشرات الX	عشر X	عشرة X	X مهم	X مهمة	X مهمون	X مهمات
		-S-N	-S-N	-P--	FP-N	MP-N	MS--	FS-- -PI-	MPR-	FPR-
كاتب	MSR	Y	Y				Y			
كاتبة	FSR	Y	Y					Y		
كاتبون	MPR			Y		Y			Y	
كاتبات	FPR			Y	Y					Y
كتّاب	MPR			Y		Y			Y	
نملة	FSI	Y	Y					Y		
نملات	FPI			Y	Y			Y		
نمل	MSI			Y			Y			
انشقاق	MSI	Y	Y				Y			
انشقاقة	FSI	Y	Y					Y		
انشقاقات	BPI			Y	Y	Y		Y		
رجل	MSR	Y	Y				Y			
رجال	MPR			Y		Y			Y	
رجالات	MPM			Y		Y			Y	
جيش	MBM	Y	Y				Y		Y	
جيوش	MPM			Y		Y		Y	Y	
تمرة	FSI	Y	Y					Y		
تمرات	FPI			Y	Y			Y		
تمر	MSI						Y			
تمور	MPI			Y		Y		Y		
طفو	MSI						Y			
طفوة	FSI	Y	Y					Y		
طفوات	FPI			Y	Y			Y		
ماء	MSI						Y			
مياه	MPI							Y		

IV. Closed Classes

Closed classes include verbs, numbers, digits, pronouns, and quantifiers. We will discuss each class separately

a- Verbs:
Gender and number functional features match their form-based gender and number features. Therefore, verbs are annotated automatically by assigning them their form-based gender and number features. Rationality feature for verbs is N. If the verb is 1st person, functional gender will be B since the verb could refer to either M or F.

b- Digits: are also annotated automatically as follows:
0 => BBN
1 => BSN
2 => BDN
Other digits (e.g., 21, 482) => BPN
Decimal numbers (e.g.,1.2, 0.5) => BBN

c- Numbers:
Each Number is annotated similarly to digits but with a specific gender value, either F or M depending on its form. However, when a number is an adjective, it behaves differently and does not follow normal noun adjective agreement rules. Such numbers were given the gender and number value B to prevent any inconsistency with the way adjectives agree with nouns.
Ex:
الأيام الأربعة
الرجال الأربعة

Now, lets look at the following example:
الرجال الأربعين
الرجل الأربعين

We believe that الأربعين should have two different lemmas. Due to this limitation in our resourse, we will overcome this by giving the word the value B for both gender and number.

d- Pronouns:
Pronouns were annotated on a case by case basis based on its core semantic meaning. If the pronoun is 1st person pronoun, e.g., for the pronoun أنا, the functional gender will be B since the pronoun could refer to either M or F.

e- Quantifiers:

Quantifiers such as كل , معظم , نصف can be modified by a singular, dual or plural, feminine or masculine word.

كل رجل
كل الرجال
كل فتاة
كل الفتيات

Although in the first two examples, كل has a different meaning (it means “each” in the first example and “all” in the second example), but our resource does not distinguish between them and give them the same lemma. Due to this limitation, quantifiers were given the gender value B and number value B to include both cases.

f- Comparative adjectives:

Comparative adjectives such as أول, أكبر, were given the gender value B and number value B. The reason behind this is that it can be modified by a singular, dual or plural, feminine or masculine word.

أذكى طالب
أذكى الطلاب
أذكى طالبة
أذكى الطالبات

Some comparative adjectives have a feminine form and can only be modified by feminine words such as كبرى, أولى. These were given the gender value F. The number value will still be B since the word can modify or be modified be a singular, dual or plural word.

أولى الطالبات
الطالبة الأولى