Naija Morphological summary

B. Caron. 29/01/2017 (revu 11/02/2019)

# Morphology: General Principles

The UD scheme allows the specification of a complete morpho-syntactic representation that can be applied cross-linguistically. This effectively means that grammatical notions may be indicated via word forms (morphologically) or via dependency relations (syntactically). The morphological specification of a (syntactic) word in the UD scheme consists of three levels of representation:

* A lemma representing the semantic content of the word.
* A part-of-speech tag representing the abstract lexical category associated with the word.
* A set of features representing lexical and grammatical properties that are associated with the particular word form.

Lemmas are typically determined by language-specific dictionaries and lexica. In contrast, the part-of-speech tags and grammatical properties are taken from two universal inventories defined below.

## Lemmas

The LEMMA field should contain the canonical or base form of the word, such as the form typically found in dictionaries.

## Part-of-Speech Tags

The list of [universal POS tags](http://universaldependencies.org/u/pos/index.html) is a fixed list containing 17 tags.  
It is possible that some tags will not be used in some languages. However, the list cannot be extended to cover language-specific extensions. Instead, more fine-grained classification of words can be achieved via the use of [features](http://universaldependencies.org/u/feat/index.html) (see below).

## Features

Features are additional pieces of information about the word, its part of speech and morphosyntactic properties. Every feature has the form Name=Value and every word can have any number of features, separated by the vertical bar, as in Gender=Masc|Number=Sing.

* There are two types of identifiers:
  + feature names = *features*
  + feature values = *values*
* All identifiers (both features and values) consist of English letters or, occasionally, digits 0-9. The first letter is always uppercase. The other letters are generally lowercase, except for positions where new internal words are marked for better readability (e.g. NumType). This makes features distinct from the [universal POS tags](http://universaldependencies.org/u/pos/index.html) (all uppercase) and from the [universal dependency relations](http://universaldependencies.org/u/dep/index.html) (all lowercase).
* A feature of a word should always be fully specified in the data, i.e. both the feature name and the value should be identified: PronType=Prs. Note that the values are not guaranteed to be unique across features, e.g. Sup could denote the superessive case, superlative degree of comparison or supine (a verb form).
* Not mentioning a feature in the data implies the empty value, which means that the feature is either irrelevant for this part of speech, or its value cannot be determined for this word form due to language-specific reasons.
* Canonical ordering: features of one word (appearing on the same line) are always ordered alphabetically; if a feature has multiple values, these are ordered alphabetically, too. This rule facilitates cases when it is necessary to compare feature sets of two words. Alphabetical sorting means that uppercase letters are considered identical to their lowercase counterparts. So for example, Number precedes NumType.
* Description of individual features usually hints what parts of speech the feature is likely to appear with. This information is intended to help understand the typical usage of the feature; however, it is *not a strict rule!* Applicability of features to parts of speech is very language-dependent and it should never be assumed that the feature cannot appear together with a particular POS tag.

### Lexical Features

All of these can be considered attributes of lexemes or lemmas (rather than individual word forms) and they represent a fine-grained sub-classification of words.

### Inflectional Features

These are mostly features of word forms rather than lemmas. There are exceptions: for instance, gender of nouns is usually a lexical feature (all word forms of one lemma have the same gender). However, other parts of speech (adjectives, pronouns, verbs) may inflect for gender because of agreement with nouns.

# Index of Parts of Speech & corresponding features

## Open class words

[ADJ](http://universaldependencies.org/u/pos/ADJ.html) : Adjectives

In Naija, they can both qualify a noun, and be the predicate of a sentence, functioning either as a Predicate Complement of copulas be & na, or as a verb. In both cases, they will retain their POS tag ‘ADJ’.

Most of them have no specific feature, and marked [\_] in the ‘gloss’ field of the dictionary.

The comparative and superlative forms borrowed from English are marked as independent lemmas, with the features [Degree=Cmp] for comparative adjectives:

* Lemmas : ***bigger; cooler; greater,*** etc.
* POS : ADJ
* Feature: Degree=Cmp

Superlative adjectives have the feature [Degree=Sup]

* Lemmas : ***biggest, coolest, greatest***, etc.
* POS : ADJ
* Feature: Degree=Sup

The 1st degree marked in English morphology [Degree=Pos] is left without feature in our dictionary [\_].

The feature [Degree] appears with ADV as well.

ADJ is used for ordinal numbers too:

* Lemmas : ***fourth; second; twenty-ninth***.
* POS : ADJ
* Feature: NumType=Ord

NB: cardinal numbers are tagged with NUM.

[ADV](http://universaldependencies.org/u/pos/ADV.html) : Adverbs

Some adverbs have features. They are:

* Demonstrative adverbs
* Tokens: ***dere; there; den; then***
* Lemmas : ***dere; den***
* POS : ADV
* Feature: there; then\PronType=Dem
* Comparative adverbs, e.g.
* Lemmas : ***earlier, nearer***
* POS : ADV
* Feature: Degree=Cmp
* Superlative adverbs:
* Lemmas : ***earliest, near***
* POS : ADV
* Feature: Degree=Sup
* Pronominal adverbs
* Lemmas : ***how, wherever, when, where***, etc. ***o***
* POS : ADV PART
* Feature: PronType=Int PartType=Disc

The other adverbs are left without feature in our dictionary and annotations.

* Lemmas : ***much, now, o’clock, occasionally***, etc.
* POS : ADV
* Feature: \_

[INTJ](http://universaldependencies.org/u/pos/INTJ.html) : Interjections

Definitions : they have an autonomous illocutionary status, and function as an autonomous utterance.

They don’t have any extra morphological features:

* Lemmas : ***hey, hello, bye-bye, ehen***, etc.
* POS : INTJ
* Feature: \_

[NOUN](http://universaldependencies.org/u/pos/NOUN.html) : Nouns

Their only morphological feature is ‘Number’.

* It is used for reduplicated forms:
* Token: ***beard\_beard***
* Lemma : ***beard***
* POS : NOUN
* Feature: PronType=Red
* and for borrowed suffixed plurals in –s
* Tokens ***bottles; beginners***
* Lemmas : ***bottle; beginner***
* POS : NOUN
* Feature: Number=Plur
* When the meaning of the word is not transparent, an English gloss is added before the morphological features. It is separated from the features by a backslash.
* Token ***brodas***
* Lemma : ***broda***
* POS : NOUN
* Feature: brother\Number=Plur

[PROPN](http://universaldependencies.org/u/pos/PROPN.html) : Proper Nouns

They keep their capital letter(s) in the lemma, and have no morphological feature.

NB: as the transcription is tokenized without capital, the capital letter should be put back by the annotator.

* Lemma : Madam, Lagos, May, N.T.A
* POS : PROPN
* Feature: \_

[VERB](http://universaldependencies.org/u/pos/VERB.html) : Verbs

Naija verbs appear in their lemmatic forms, and hence, don’t carry any morphological features. Verbs appearing in Serial Verb Constructions are tagged with a lexical feature (specifically introduced by us in the UG) which will help identify and retrieve them for further study.

* Lemma : arrange, arrest, bless, burn, etc.
* POS : VERB
* Feature: \_

~~(Modifying) verbs appearing in SVC are marked with the lexical feature [VerbType=Ser]~~

* ~~Lemma : comot, fit, give, pass, reach, come etc.~~
* ~~POS : VERB~~
* ~~Feature: VerbType=Ser~~

(NB: the annotation of Serial Verb Constructions has been moved to the syntax)

However, the Deuber corpus shows many instances of verbs following the English morphology, e.g. ‘***was, is, were, teaches, teaching, taught, comes, coming, came, bringing, brought***’, etc. They will be tagged using the English set of morphological features. These are:

* Mood: Imp (Imperative); Ind (Indicative)
* VerbForm: Fin (Finite), Ger (Gerundive), Inf (Infinitive).
* Tense: Past (Past), Pres (Present)

If in doubt, run a search of the the **English(UDv1.3)** corpus on the UD website, and follow the examples you will find there: <http://bionlp-www.utu.fi/dep_search/?treeset=UD> (NB: don’t forget to select the English treeset before lauching your search on any token).

## Closed class words

[ADP](http://universaldependencies.org/u/pos/ADP.html) : Adpositions

In Naija, these are prepositions. Following Huddleston and Pullum’s analysis of English parts of speech, they keep their [ADP](http://universaldependencies.org/u/pos/ADP.html) tag even when they are not followed by an NP and inherit the head function, and work as an ADV.

* Lemma : about, along, as, at, before, dan (=than), from, like, down, etc.
* POS : ADP
* Feature: \_

[AUX](http://universaldependencies.org/u/pos/AUX_.html) : Auxiliaries

An auxiliary is a function word that accompanies the lexical verb of a verb phrase and expresses grammatical distinctions not carried by the lexical verb, such as person, number, tense, mood, aspect, voice or evidentiality. It is often a verb (which may have non-auxiliary uses as well) but many languages have nonverbal TAM markers and these should also be tagged AUX.

NB: In UG, the class AUX also include copulas (in the narrow sense of pure linking words for nonverbal predication). In Naija, I have taken the decision to tag them as PART.

AUX in Naija are grouped according to their features:

* Tense

|  |  |  |
| --- | --- | --- |
| Lemma | ***bin*** | ***go*** |
| Feature | Tense=Past | Tense=Prosp |

* Aspect

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Lemma | ***come/con*** | ***dey*** | ***don*** | ***never*** | ***take*** |
| Feature | Aspect=Real | Aspect=Imp | Aspect=Perf | Aspect=Perf|Polarity=Neg | Aspect=Inch |

* Mood

|  |  |  |
| --- | --- | --- |
| Lemma | ***make/meh*** | ***for*** |
| Feature | Mood=Jus | Mood=Cond |

Example:

* Lemma : ***go; dey; make***, etc.
* POS : AUX
* Feature: Tense=Prosp; Aspect=Imp; Mood=Jus

Auxiliaries borrowed from English (e.g. ***can, must, shall, could, will***) have the only feature: [VerbForm=Fin]

* Lemma : ***can, must, shall, could, will*** etc.
* POS : AUX
* Feature: VerbForm=Fin

[CCONJ](http://universaldependencies.org/u/pos/CCONJ.html) : Coordinating conjunctions

Coordinating conjunctions in Naija (e.g. ***abi/tabi, and, both, but, or, plus, eida***) have no special morphological feature.

[DET](http://universaldependencies.org/u/pos/DET.html) : Determinants

Two categories of determinants are marked with features : articles (***di*** is a definite article, and ***one*** is a Specific Indefinite article), demonstratives (***dis***, ***dese***, ***dat***, ***dose***), and the emphatic morpheme ***sef[[1]](#footnote-1)***.

* Lemma : ***di ; one***, etc. ***sef***
* POS : DET DET DET
* Feature: Definite=Def|PronType=Art Definite=Spec|PronType=Art PronType=Emp
* Lemma : ***dis ; dese***, etc.
* POS : DET DET
* Feature: PronType=Dem|Number=Sing Number=Plur|PronType=Dem
* Lemma : ***dat ; dose***, etc.
* POS : DET DET
* Feature: Number=Sing|PronType=Dem Number=Plur|PronType=Dem

[NUM](http://universaldependencies.org/u/pos/NUM.html) : Numerals

This part of speech concerns cardinal numbers.

* Lemma : ***one***, ***seven, thirteen, etc.*** etc.
* POS : NUM
* Feature: NumType=Card

NB: Some linguists prefer to use the ADJ part of speech for cardinal as well as ordinal numbers.

[PART](http://universaldependencies.org/u/pos/PART.html) : Particles

Three types of particles (neither AUX, nor DET, ADV, or ADJ, but a little of all that), covering Polarity (***no***, ***not*** for negation), Emphasis (***sef*** a noun modifier; ***o*** as a sentence modifier), and Copulas (***na***, ***dey*** & ***be***).

They are differenciated by the lexical feature PartType, which takes two values: Cop (=Copula) & Disc (=Discourse).

* Lemma : ***no, not na***, ***dey***, ***be o, sha***
* POS : PART PART PART
* Feature: Polarity=Neg PartType=Copula PartType=Disc

[PRON](http://universaldependencies.org/u/pos/PRON.html) : Pronominals

Two types of pronominals are tagged : personal pronouns, and adverbial pronouns.

* Personal pronouns

The morphological features of personal pronouns (PronType=Prs) are complex. They tag Case (Nominative= Independent and subject pronouns; Accusative = other pronouns, e.g. direct objects and pronouns following prepositions); Number (Sing/Plur), Person (1, 2, 3), and Possessive (Boolean feature Yes).

|  |  |  |  |
| --- | --- | --- | --- |
| me | mi | 1sg. Nominative (unbound) | Case=Nom|Number=Sing|Person=1|PronType=Prs |
| I (a) | à | 1sg. Nominative (bound) | Case=Nom|Number=Sing|Person=1|PronType=Prs |
| me | mi/mì | 1sg. Accusative | Case=Acc|Number=Sing|Person=1|P ronType=Prs |
| ma | mà | 1sg. Possessive | Number=Sing|Person=1|Poss=Yes|PronType=Prs |
| you | yu/yù | 2sg. Nominative | Case=Nom|Number=Sing|Person=2|PronType=Prs |
| you | yu/yù | 2sg. Accusative | Case=Acc|Number=Sing|Person=2|PronType=Prs |
| your | yɔ̀ | 2sg. Possessive | Number=Sing|Person=2|Poss=Yes|PronType=Prs |
| (h)im | im/ìm | 3sg. Nominative | Case=Nom|Number=Sing|Person=3|PronType=Prs |
| e | ì | 3sg. Nominative (@) | Case=Nom|Number=Sing|Person=3|PronType=Prs |
| (h)im | hìm | 3sg. Possessive | Number=Sing|Person=3|Poss=Yes|PronType=Prs |
| am | am/àm | 3sg. Accusative | Case=Acc|Number=Sing|Person=3|PronType=Prs |
| we | wi/wì | 1pl. Nominative | Case=Nom|Number=Plur|Person=1|PronType=Prs |
| us | ɔs/ɔ̀s | 1pl. Accusative | Case=Acc|Number=Plur|Person=1|PronType=Prs |
| our | àwa | 1pl. Possessive | Number=Plur|Person=1|Poss=Yes|PronType=Prs |
| una | ùnà | 2pl. Nominative | Case=Nom|Number=Plur|Person=2|PronType=Prs |
| una | ùnà | 2pl. Accusative | Case=Acc|Number=Plur|Person=2|PronType=Prs |
| una | ùnà | 2pl. Possessive | Number=Plur|Person=2|Poss=Yes|PronType=Prs |
| dem | dɛm/dɛm | 3pl. Nominative | Case=Nom|Number=Plur|Person=3|PronType=Prs |
| de | dè | 3pl. Nominative (@ ?) | Case=Nom|Number=Plur|Person=3|PronType=Prs |
| dem | dɛm/dɛ̀m | 3pl. Accusative | Case=Acc|Number=Plur|Person=3|PronType=Prs |
| dem | dɛ̀m | 3pl. Possessive | Number=Plur|Person=3|Poss=Yes|PronType=Prs |
| we-we | wi-wi | 1pl. Reciprocal | Number=Red|Person=1|Poss=Yes|PronType=Rcp |
| una-una | ùnà-ùnà | 2pl. Reciprocal | Number=Red|Person=2|Poss=Yes|PronType=Rcp |
| dem-dem | dɛm-dɛm | 3pl Reciprocal | Number=Red|Person=3|Poss=Yes|PronType=Rcp |

* Adverbial pronouns

They are tagged with the feature PronType=Int to reflect their main use as an interrogative pronoun.

* Lemma : ***wetin***, ***what, which, who, whoever*** ***whose***
* POS : PRON PRON
* Feature: what\PronType=Int Poss=Yes|PronType=Int

[SCONJ](http://universaldependencies.org/u/pos/SCONJ.html) : Subordinating conjunctions

Subordinating conjunctions in Naija (e.g. ***after, if, since, so sotay, that, though, unless, until, wey, weda (=whether), whereas, because, derfore (=therefore)***) have no special morphological feature.

## Others

[PUNCT](http://universaldependencies.org/u/pos/PUNCT.html)= Punctuation signs

[SYM](http://universaldependencies.org/u/pos/SYM.html)= Symbols

NB : does not apply to our corpus, which is oral.

[X](http://universaldependencies.org/u/pos/X.html)

Applies to foreign words, codeswitching, etc. and words that cannot be analysed.

# References

Huddleston, Rodney & Geoffrey K. Pullum. 2008. *The Cambridge Grammar of the English Language*. 2nd ed. Cambridge: Cambridge University Press.

1. ***Sef*** may need further analysis to confirm its status as DET. [↑](#footnote-ref-1)