**DESIDERATA AND PLAN FOR CALIMA STAR**

**Encoding:**

* UTF8
* BW
* safeBW
* Others?

**Tokenization:**

* Schemes
  + D1
  + D2
  + D3
  + ATB
  + Others
* Tokenization vs raw segmentation
* t+akotub+uwna instead of ta+kotub+uwna

**POS tags:**

* BW
* MADAMIRA
* CATiB
* CATiBex
* A new CALIMA STAR tag set
* Others?

**Phonology:**

say~AratunA => / s a y y aa r a t u n aa/

* CAPHI map (during processing or post processing?)
* Separate Sun and Moon Al
* Reference: Biadsy and Habash and Hirschberg

**Morphology:**

* Completely functional
* A one to one map from pos tags to morphological feature/value pairs
  + NOUN.MPDN => pos:noun gen:m num:p cas:d stt:n
* Reduced MSA option allowed
  + saharap vs saharapu, saharapa, saharapi
* Morphological representation includes:
  + Full functional feature-value (new content)
  + Form-based feature-value (new formulation)
  + Form-based morph-meaning (this is BAMA)

**Clitics:**

* Possible schemes;
  + prc0,1,2,3
  + prc1.5,2.5
  + prc1-9
  + prc:
    - e.g. prc:b/PREP+Al/DET
  + Others?
* Mapping from whichever scheme to the prc0,1,2,3 scheme used in MADAMIRA

**New features:**

* Dialect
* LMM
* Root
* Pattern
* Probability of words out of context (configurable depending on a given text/corpus)
  + solves the problem of lm and ln
* Others?

**Spelling variant:**

* Input and output
* Alif/Ya/Ta/Hamza errors
* AF vs FA

**Possible Output Examples:**

saHaratihim

SEG:saHar-ati+him

TOK:saHar-ap+hum

POS:NOUN.MPDN+PRON.3MP

XYZ: BASE+CLITIC

XYZ:STEM-api+CLITIC

\*DIAC:saHar-ati/NOUN.MPDN+him/PRON.3MP

\*ATB:saHar-api/NOUN.MPDN+hum/PRON.3MP

wasayaktubuwnahA

wa/CONJ+sa/FUT+y-aktub-uwna/VERB.I3MP+hA/PRON.3FS

**Extras:**

* Entries must be easy to add
* UD compatible
* Analysis and generation
* Link to paradigm completion and easy annotation in MADARI
* Consistent across all decisions
* Primary vs secondary fields
  + trim database for output
  + post process?

**Backoff**

* Smarter morphological backoff for unseen words: not just assume it's a PropN (beyond BAMA)

\* Tripartite database in one file (like BAMA, except for single file)

\* Make classes in database aware of inflectional classes which are explicitly defined by us (under MAGEAD) and available as a resource (details to be worked out) (beyond BAMA)

**ARGUMENTS FOR OUR DATABASE AND APPROACH**

* Consistency
  + BAMA is not consistent in tokenization/segmentation: ll => l+Al but thm is not +p+hm?
* Ease of adding entries
  + Prc0,1,2,3 is not easily extended
    - example: lbAllyl and lhyktb
  + Adding new entries requires selection of stemcat which is not intuitive
    - selection of paradigms
    - marking stems and/or roots
* Functional morphology
  + No one is doing complete functional morphology

**SYSTEM AND DATABASE COMPONENTS**

**DATABASE:**

* Lexicon with the columns:
  + lemma
  + gender-number paradigm ID
  + root
  + stem
  + pos
  + rationality
  + gloss
  + link to the original bw lemma
    - are there cases where we might merge two bw lemmas? how to handle?
  + others?
* State-case paradigms
  + for nouns (nominals):
    - derived from the document Google\ Drive/\ NYUAD-CAMEL/CALIMA-STAR/NOUN-GNSC.xlsx
    - structure?
  + for verbs:
    - mood/voice/aspect paradigm
  + closed classes?
  + others?
* Gender-number paradigms
  + nouns (nominals)
    - stem + state-case paradigm ID *OR* pattern + state-case paradigm ID
    - 6 entries each: MS, MD, MP, FS, FD, FP (might not all be valid)
    - add lemma as an entry?
  + verbs
  + closed classes
  + others?
* Clitics
* A model for orthography rules

**SYSTEM:**

* Create the gender-number paradigms and the lexicon
* Create the state-case paradigms
* Link gender-number paradigms to state-case paradigms
* Extend with clitics and apply orthography rules
* Generate db
* Python morphological analyser
* Wrapper taking MADAMIRA input and outputting equivalent input of our own
  + Similar to the wrapper producing segmentation done for Alex’s projects
* Interface to add entries