Arfath Pasha - use 34 hours / python?

UTF8 as native in databases; but generate BW, SafeBW, etc.

Transcription - phonological: Biadsy and Habash and Hirschberg

CAPHI MAP...

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

* Secondary vs primary?

Al Shams/Moon as different?

Wrapper taking MADAMIRA input…

Six tables good to use

All of the morphology will be functional (but stem is useful)

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

Generate D1,D2 etc.; RAW segmentation, (not TOKAN)

Al/DET+saHar-apu/NOUN.MPDN

Al/DET+saHar-apu/NOUN.MPDN

Configurable morpheme to feature/value pairs?

Different maps to feature/value pairs: prc1,2,3, o 2.5 of 1-9…

prc:b/PREP+Al/DET pos:

Add secondary features: Catib , catibex, LMM etc.

saHaratihim

SEG:saHar-ati+him

TOK:saHar-ap+hum

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

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

POS:NOUN.MPDN+PRON.3MP

XYZ: BASE+CLITIC

XYZ:STEM-api+CLITIC

\*stem,affixes,clitics

\*tok/seg

\*pos

Spelling variation (on inpout and output ktbFA or ktbAF ) and A/Y/p/Hamza errors..

Ease of addition of entries

root.patt,,,

====================

Multidialectal view?

\* common set of POS for MSA and dialects

\* all POS tags are FUNCTIONAL - one to one map of POSTag and Functional feature (NOUN.MSDG) => pos:noun gen:m, …

\* MSA - reduced allowed > saHarap / apu..

\* t+aktub+uwna

\* simplify tags

\* Map from BW - all data from LDC

\* UD compatible?

GOALS/DESIDERATA

\* Allows for analysis and generation (BAMA only does analysis)

\* Morphological representation includes:

- Full functional feature-value (new content)

- Form-based feature-value (new formulation)

- Form-based morph-meaning (this is BAMA)

On inflections and clitics…. Prc1 prc 2? >>>

(link to paradigm completion and easy annotaiton in MADARI)

- Full segmentation and tokenization information based on form-based

morphology; several tokenization schemes (BAMA has one) with consistent

choices (BAMA is not consistent)

- Functional (MAGEAD) and form-based (as we use in paradigm comletion)

pattern/root/vocalism (BAMA does not have)

- English glosses (BAMA has, but can get new)

\* Sufficient lexical coverage (how is that defined?)

- Coverage of ATB\* (BAMA has this, roughly, plus a lot more that is not

in ATB\*)

\* Backoff

- Smarter morphological backoff for unseen words: not just assume it's a

PropN (beyond BAMA)

- Fuzzy orthography for input words (Alif forms, hamza forms) -- Nizar

remembers that there were some specific ideas (perhaps go beyond BAMA);

add common MSA spelling errors: Z/D etc.

\* Probability of words out of context (can be derived from ATB) (not at all

in BAMA) >> YAMAMA!

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

\* Make classes in database aware of inflecional classes which are

explicitly defined by us (under MAGEAD) and available as a resource

(details to be worked out) (beyond BAMA)

OTHER GOALS

\* Can re-annotate the ATB with full morphological representation used in

CALIMA (largely automatic with possible subsequent annotation/error

checking)

HOW DO WE GET THERE

These are preliminary thoughts.

\* Derived from ATB (different from BAMA)

- Gather all forms with segmentation, lemma, features

- Fix the bad represntations in BAMA/SAMA such as the inconsistent

handling of Al and ( +ap and +aY and many more) between tokenization

and segmentation.)

- Do paradigm completion

+ Make use of iconic inflectional classes from Magead

+ Start with Mohammed Tantawy's list of sample verbs for various classes

- Do something special for closed classes

\* Glosses: various posssible sources, including MT techniques

\* Backoff: need to code

\* We need to do evaluation and error correction (LDC?)

\* Lexicon: use other Arabic lexical resources

LEGAL ISSUES

\* We want joint ownership so that we can distribute it freely

\* EGY and IRQ should be similar, design should anticipate dialectal

resources as well

PEOPLE:

\* Columbia: Nizar, Owen, Ramy, Sarah

\* LDC: Chris, Ann B, M Maamouri, ...?

\* Qamus: Tim???

\* Others: Ota, ...

===============

Problems we solve:

* BAMA is not consitent in tokenization/segmentation: ll => l+Al but thm is not +p+hm?
* Prc1234 not easily extended (lbAllyl and lhyktb)