Arabic Part-of-Speech/Morphological Analysis Guidelines

The Penn Arabic Treebank uses a level of annotation more accurately described as morphological analysis than as part-of-speech tagging. (For further background, see also the description of Tim Buckwalter's lexicon and morphological analyzer at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49.)

Instructions to annotators.

A few things to keep in mind when part-of-speech tagging Arabic text:

- ?? As a general rule, your task is just to select a solution from the list given – if a correct solution is not available on the list, just click the "no match" button. Don't worry about writing comments for now.
- ?? Do not tag the "byline" information at the beginning of the file (the date, location, etc. of the original written newspaper article). Start with the beginning of the actual text in the article.
- ?? Tag or comment every word in the file (after the byline). Don't skip any!
- ?? If anything goes wrong or you have any kind of problem with the interface, the program or the files, please send e-mail to Hubert or to Kazuaki.
- ?? You can see what the part-of-speech annotation tool looks like here.

How to tag an Arabic word:

- ?? First, look at the highlighted word in Arabic script in the window at the top of the "tool."
- ?? Sound out (in your head! :-) what the word should be, including its case ending. Read just enough to be sure you understand the meaning. Now you know whether the word is a noun, verb, etc.
- ?? Then, look at the provided solutions in the lower window, starting with the voweled word.
- ?? Once you have identified which voweled word you want, look only at the choices for the category (noun, verb, etc.) you have already chosen in your mind.
- ?? Choose the row that has the rest of the correct analysis (case, number, gender, etc.) for the word.
- ?? The final step is to double-click on that row and make sure that the tagged selection appears in the "Selection" box.

Dealing with problematic cases:

- ?? When there is a problem with the file (for example, if the Arabic words do not correspond to the transliterated/tagged words), (1) send Hubert e-mail with the full file name/number, (2) check the file back in without working on it any more, and (3) take a new file to work on.

- ?? Passive verb – click on the "Passive form" button. If you identify a verb as passive, but there is no passive solution given, click on the "Passive form" button. This will automatically register the passive solution in the comment box, and you can go on to the next word.
- ?? Foreign names – click on the "Is a name" button. If you identify a noun as a foreign name/proper noun and that solution is not given, click on the "Is a name" button to automatically add that comment, and go on to the next word.
- ?? No correct solution in the list – click on the "No match" button. If there is no correct solution given in the list, click on the "No match" button to automatically add that comment to the comment box, and go on to the next word.
- ?? If there is one correction to be made, go ahead and fix it. But if there are three corrections to be made (for example, if there are problems with the gloss and the transliteration and the tag), don't take the time to fix it all – just click on "No match" and go on.
- ?? For any other problems or questions, check to see if the answer is in the "POS Questions & Answers" section below. If the answer is not there, make a note of the question (or send it to Ann or Mohamed), click on the "No match" button, and go on to the next word.

POS Questions & Answers

Divided/compound proper names in Arabic (Abdul Ahmed, e.g.): Label all parts of the name with the "Is a name" button.

Idioms: (for example, in what in them = 'included'): Label each word independently for its own part of speech (ignore the idiomatic meaning).

Don't focus too much on the translation/gloss. The gloss is useful and important as an indicator of what the tag is if other structural indicators don't tell you, but it is not so important in and of itself. Put something in the comment line if the gloss is really bad, but if the gloss is understandable as is, just let it go.

Wrong vowel: use the "Should be u" "Should be a" "Should be i" buttons. Don't worry about (and no need to comment on) which syllable has the wrong vowel at this point, since it will be obvious to the corrector.

Missing hamza: use the "Hamza problem" button.

Typos: use the "Typo" button.

Noun vs. Adjective: If the word is really an adjective, but there is no ADJ solution given, use the "NOUN -> ADJ" button. Similarly, if the word is really a noun, but there is no NOUN solution given, use the "ADJ -> NOUN" button.