Hello, and welcome to MTAAC. If you’re reading this, it’s because you’ve been roped into manually annotating texts for our gold corpus. Please find a summary of this assignment below.

TL;DR Annotations

1. Open the prepared .conll file in Libre Office.

2a. If blank, complete the SEGM and XPOSTAG fields.

2b. If pre-annotated, double-check populated fields and fill in the blanks.

3. Save as .tsv/.csv (see below)

4. ?????

5. Profit

For the preparation of MTAAC’s gold corpus, we will annotate 5.2% of the total corpus, about 4,000 texts, with morphological and lexical information. 20% of these texts will be accompanied by English translation, the rest will be untranslated.

Annotation files will be in a version of the CoNLL-U format specialized for our corpus, known as CDLI-CoNLL. CDLI-CoNLL arranges the text with one word per row and columns to track information about the words.

The format of the annotation file is called CDLI-CoNLL

Basically it arranges the text as one word per line with columns for different information about the words.

It is designed for data entry and storage of lexical and morphological information and will live in a second field beside the ATF. (We are planning on developing a tool that verifies that both are synchronized, repairs some discrepancies, and warns us of other problems.)

The file can be edited in Excel, Calc or Sheets: import as UTF-8 TSV file format.

The columns are as follows: (See P222901.conll for an example)

ID: all information about the surface, column, line and word token (o.col1.1.1; o.1.1 if there is no column) Only the column number is optional.

FROM: token from text, atf transliteration

SEGM: normalized form of the token (amar-inim-du10-ga-ta => amarinimduga.ta) BUT using the dictionary entry word as the stem part

XPOSTAG: ORACC ETCSRI morphological tags, separated by periods and replacing "STEM" with the part of speech or named entity tag. eg (amar-inim-du10-ga-ta => PN.ABL) Implicit morphemes should be put in square brackets.

A human annotator will face 2 types of files:

1) A file like P222901\_empty.conll (in the first part of the work only)

Task : fill in all fields like in P222901.conll

2) A pre-annotated file like P222901\_empty.conll (when we have enough data to generate pre-annotations)

Tasks :

- Verify generated annotations

- Check for lines where there are suggestions in additional columns, choose best option and copy and paste to appropriate field

- Fill in remaining fields

The part of speech tags are here :

<<https://docs.google.com/spreadsheets/d/1Is7MGG0h8h0vfHj9C9mnWOD2utPeuvm1ZeYb1dsaejg/edit?usp=sharing>>

The ORACC ETCSRI morphological tags can be found here :

<<https://docs.google.com/spreadsheets/d/1y0_y9HDQNwH0VqDCjjYuUpFsugw4GEybu6Pu01I_D9c/edit?usp=sharing>>

(Both in the annotation folder)

We have the ETSCRI data as an example so it is possible to search for forms to help, I could also extract all variants in a list prepared in CDLI-CoNLL format to make it even easier.

See ETCSRI\_Q001379\_annotation\_data.conll for the type of information we have from ETCSRI.

There will be some slight changes to this workflow and model as we advance, eg the POS for verbs will be a bit more elaborate, also during testing I will certainly discover problems to address...

I'm working on the documentation of this topic this week, I'll send you the link when ready.

in libre office, make sure to check ONLY the “tab” option as a separator when opening the .conll files (should only have to do this once, then it will remember as default). This way you can simply command+s save the file and it will preserve the format. If Libre issues a warning, click “use CSV” or whatever it says. You can uncheck the box to save save this step so that a warning does not arise.