# Morphological annotations questions and problems

## Word types

Adjectives:

Few true adjectives exist in Sumerian. One that comes to mind is “gal” (big, great). Zolyomi classified a couple of words as adjectives such as “nig2-tuku” (the one acquiring things: “rich”) which is thus a participle. In fact, almost all instances of adjectives in ETCSRI are participles formed with the verb tuku (to take, acquire) or gal2 (to be, exist). Where do we draw the line? What should we consider an adjective and why? This distinction should probably be made so it is easier to process the text after annotation as humans understand it’s not a black or white situation.

Compound verbs:

How do we manage compound verbs? Should we give a special POS to the object part of the compound? What about the lemma?

compound verbs? e.g. al ak (P134626)

Are we going to indicate the parts of compound verbs are linked? One option is to provide the comp with all the options on the backend and manually annotate the text more simply by just marking the constituent nominal and verbal components like normal. Thus the computer should be able to render a contextually based analysis without us having to do any extra work in the annotation.

POS antecedent: ETCSRI marks both “original” POS and “actual” POS, eg nominalized verbs are marked as noun and verb, distinctly. Same goes for numerals that are used as nouns and not adjectives. Do we need that information? If yes, in which cases? How to we tag it?

Participles: Connects with problems above: How should we manage participles? Christian mentioned

## Word definition

Word segmentation is not 100% equivalent in CDLI and ETCSRI

Lemma

ORACC uses lemma[sense] to create unique entries for words with a specific sense. Should we add [sense] to the lemma in the SEGM field? In the ETCSRI has a column for the sense so I can add it to the ETCSRI model of segmentations

How should normalize forms? with or without morphemes, overhanging consonants, etc?

Should tag-tag-a be SEGM tag-tag-a or tag-a, XPOSTAG

engma/linking with epsd lexicon

When there is variation between the “CDLI version” and the ePSD, how do we want to handle it? For example, the common phrase /pisan-dub-ba/ on CDLI = bisajdubak; we must decide how this will be entered in the SEGM field.

If possible, we should create a general rule for the occurrence of engma to make things simple.

double genitive = GEN.GEN

We want to include determinatives in the segmentation but NOT phonetic complements (which should be come through in the normalization anyway).

Names building that is not a temple, e.g. wall, palace? Following ETCSRI, currently tagging as GN “geographical name” but

How does Christian want us to keep record of which bits are damaged or reconstructed while we’re annotating? We can segment as if they are unbroken, but there may be some contradiction in different uses of square brackets, e.g. to indicate reconstructed text or supplemented morphemes.

For verbs, we will (for now) not indicate transitivity in our annotations. ePSD includes in its entries whether or not a given verb is *generally* transitive or intransitive, and ETCSRI includes this in its tags. This binary classification may not accurately reflect the character of ambitransitive verbs in their various contexts, so we will omit it in our own tags. Information about the “typical” transitivity of verbs ought to remain in the lexicon, while the contextual transitivity should be made self-evident via morphological tagging.

In order to train the computer as best we can, morphemes that are not expressed in the text must be reconstructed in the segmentation according to our analysis. For example, in P116238 there may be implied genitival relationships that are not explicitly indicated with -ak. Accurately tagging the morphology in such a case requires making the determination whether or not this relationship is *really* genitival and necessitates the addition of [-ak] in the segmentation. This should be quick when it comes to annotating those texts already given in English translation (provided the morphology is accurately represented in translation) but may become prohibitively time consuming when annotating untranslated texts without clear rules.