Morphological Segmentation

Annotation Guidelines

Version 1.3

Linguistic Data Consortium

Table of Contents

1. Introduction

Words are often made up of more than one meaningful part. For example, the English words below can be divided as follows:

dogs = dog + s
laughed = laugh + ed
unhappy = un + happy

In these examples, the ending *–s* tells us that the noun is plural in number. The ending *–ed* indicates that the verb is in the past tense, and the prefix *un-* marks the negative form of the adjective.

In linguistics, the study of how words are composed is called **morphology**. We will call these smaller parts morphological **segments**. In this annotation task, we will break down some words in your language into the segments that make them up.

1.1. Base form

In this task, we will mark how the word we are annotating is different from its **base form**. In English, the singular form of a noun is the base form. Similarly for verbs in English, the unconjugated infinitive form is the base form.

cat
(to) **walk**

Other languages have different base forms. In Hindi, the base form of a verb is the infinitive form minus the infinitive suffix *–nā*. The base form of a noun in Hindi is the word minus any case/number/gender markers. In Russian, it can be the nominative singular form of the noun. In the following examples, the base form is indicated in **bold**. In Spanish, the base form of the verb is the infinitive minus the endings *ar/er/ir.* In Tamil, the base form can be derived by removing the person/number endings from the present tense form. Also for Tamil, the base form for a noun is the nominative singular form.

cāh + nā (Hindi) ‘want’
laṛk + ā (Hindi) ‘boy’
laṛk + ī (Hindi) ‘girl’
drug (Russian) ‘friend’
empez ar (Spanish) ‘to begin’
romp er ‘(Spanish) ‘to break’
tūṅku kir ēṉ (Tamil) ‘I sleep’
paiyaṉ (Tamil) ‘boy’

1.2. Productive elements

In this task, we will only segment on **productive** elements. By productive, we mean morphological elements that are currently in active, widespread use in the language. For example, in English, most nouns have a plural form ending in *–s*, but there are some irregular nouns which have the ending *–en*.

one ox → two oxen
one cow → two cows

One way to test if a segment is productive is to try adding it to a made-up word, such as *wug*.

one wug → two wugs ?
or
one wug → two wugen ?

English speakers would agree that the plural form of the word *wug* should be *wugs* and not *wugen*; this demonstrates that adding *–s* is productive, while adding *–en* is not. Note that this does not imply that all instances of *–en* in English cannot be productive; see example of past participial *–en* in 2.2.

2. Annotation

The following sections describe various ways morphological segments can combine and change, along with how to represent those operations and changes in the annotation scheme.

2.1. Concatenation

When two or more elements occur side by side, with no change to either of them, they are called concatenated. In this annotation task we will indicate concatenation with a space between the two elements.

dog s
laugh ed
un happy
cāh nā (Hindi) ‘to want’
vara sto I sta si ko (Faroese) ‘from your stores?’
nāṭu kaḷ (Tamil) ‘countries’
rot’ ačč ïhu (Amharic) ‘you (pl.) ran’
kwa Zama (Zulu) ‘Zama’s place’
a ŋ kɔ (Akan/Twi) ‘did not go’
wa tu (Swahili) ‘people’

Note that because we do only annotate **productive** processes, in English we would get the following annotations:

cow s

oxen and **not** ox en

In cases where part of an adjacent word is included in the token being annotated, a break is introduced, even if that segment is not meaningful, as occurs with Tamil initial obstruent doubling:

inda p (paiyaṉ) (Tamil) “this (boy)”

2.2. Substitution

In many cases, there is a change in a morphological segment when compared to its **base form**. In this annotation scheme, we will mark the changed element in angled brackets.

g<a>ve base form: give
br<o>k en base form: break
k<i> yā (Hindi) ‘done (masc. sing.)’ base form: kar
p<ue>d o (Spanish) ‘I can’ base form: pod
mara<ṅ> kaḷ (Tamil) ‘trees’ base form: maram
m fua<s> l (Swahili) ‘follower’ base form: fuat

Note that if there is a potential change in the base form in a derived from that we cannot see, that is not explicitly marked in the annotation. In other words, we only tag what has changed and not what can change. For example:

give and **not** g<i>ve

2.3. Insertion and deletion

It is common to have elements inserted and deleted in morphological operations. These insertions and deletions are **not** explicitly marked in the annotation.

Insertion:
running = run ning

Deletion:
tūṅkā (Tamil) ‘to sleep’ = tūṅk ā from tūṅku + ā

3. Additional Considerations

3.1. Meaningful elements

In some cases, it may happen that part of word may be identical to a meaningful morphological segment. We only segment if it **is** that meaningful element. For example:

revisit = re visit
receive = receive and **not** re ceive
ghoṛā (Hindi) ‘horse’ = ghoṛ ā

rājā (Hindi) ‘king’ = rājā and **not** rāj ā

In these examples, the *re-* in *revisit* is meaningful; in *receive* it is an integral part of the word that has no separate meaning of its own. As a result it is not annotated as a segment. In Hindi *ghoṛā* the *–ā* is a case/gender marker indicating the noun is masculine singular direct case, contrast *ghoṛe par* ‘on a horse’; *ghoṛe* ‘horses’; *ghoṛoṃ par* ‘on horses’. On the other hand, the final *-ā* in *rājā* is invariable and does not change for case or number. It is part of the basic form and is not a separate meaningful element.

3.2. No empty brackets

If an element is missing from segment, its absence goes unmarked in the annotation. In other words, do not create an empty bracket.

refus al and **not** refus<> al

3.3. Ambiguous segmentation

In some cases, it may be plausible to include part of a segment with two different segments. In these cases, we will take into consideration the overall symmetry of the paradigm, and favor segmentations which reduce the number of overall forms of noun and verb stems.

Consider the following forms of the Spanish verb *hacer* ‘to do’ with the following annotations:

hac er ‘to do’
hac ía ‘I would do
h<i>c iera ‘he would do
ha ría ‘I would do’
ha go ‘I do’ **not** ha<g> o
ha gamos ‘we do’ **not** ha<g> amos

Here *g* is included in the person/number suffixes rather than on the verb stem. If it had been attached to the verb stem, we would have created an additional form of the basic form of the verb. Proliferation of base forms should be avoided, even at the expense of inflating the number of variant forms for tense/aspect/mood/etc. markers.

3.4. Homographs

Occasionally a word may be written the same as another word which is more or less morphologically complex. In those cases, the annotator is free pick whichever analysis they like.

number (1,2,3…) = number
or
number (more numb) = numb er

3.5. Templatic morphology

Details on how to handle templatic morphology are still an open question as of this version of the guidelines. The most salient open question is the determination of the base form for words. The choice of annotation decisions crucially relies on the determination of the base form. It is also possible that additional notation will be required to account for all derived forms.

However, the currently slated languages to begin morphological segmentation do not exhibit templatic morphology, so this question can be resolved at a later time after consultation and discussion.

3.6. Compounds/Depth

Detailed specification is forthcoming about how to limit annotation depth in morphologically rich languages that also exhibit noun composition. The example provided is:

útflutningsátøkini (Faroese) ‘the export campaigns’

Exhaustive annotation would yield:

út_fl<u>t_n_ing s á_t<ø>k ini

A simplified annotation, limiting depth by not annotating all constituent parts would be:

útflutning s át<ø>k ini

Note that no languages currently slated for annotation exhibit this interaction between richness in morphology combined with prevalent noun-noun composition. The current proposal is to annotate exhaustively until further notice. Further specification can be developed after discussion and consultation, if needed.