Take care of your morphs with mSUD annotation format!

Joint Annotation of Morphology and Syntax in Dependency Treebanks

Bruno Guillaume, Kim Gerdes, Kirian Guiller, Sylvain Kahane, Yixuan Li

Word level annotation may be difficult to apply

- **Agglutinative** languages (Turkish)
- Polysynthetic languages (Yupik)
- Languages written without spaces (Chinese)
- Languages with an **oral tradition** (Beja)

We propose **mSUD**

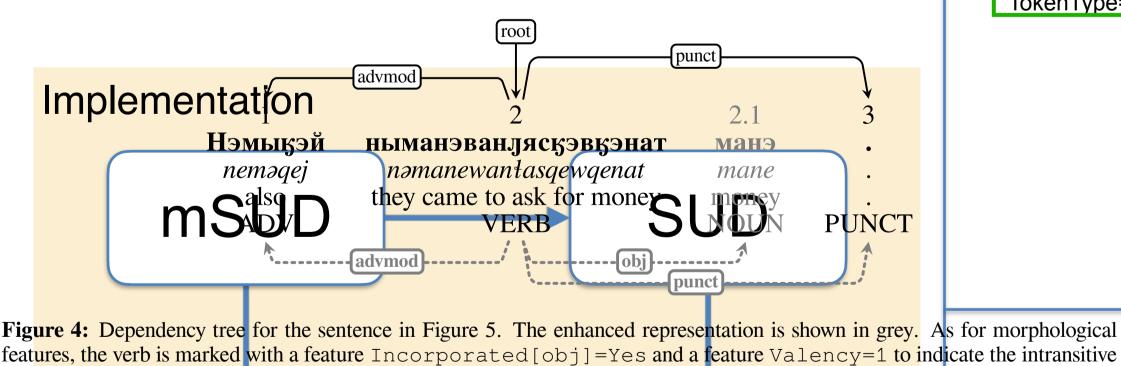
- Morph-level annotation framework
- Convertible to word-based format
- Easier inclusion of IGT-based source data

In mSUD

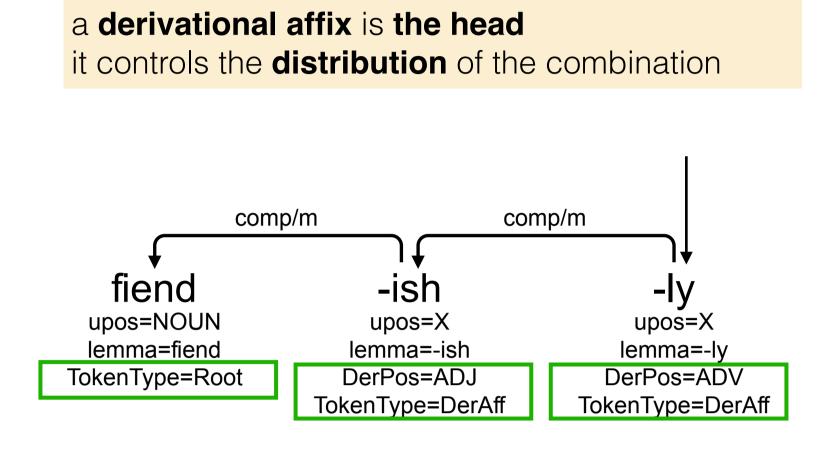
- Two types of dependency
 - regular (e.g. subj)
 - at the morphological (e.g. subj/m)
- A feature **TokenType** (values: **DerAff**, **InflAff**, Root)
- Features for word-level **upos**:
- **DerPos** for **derivational affixes**
- CpdPos for compounds

Three categories of **subword** annotations

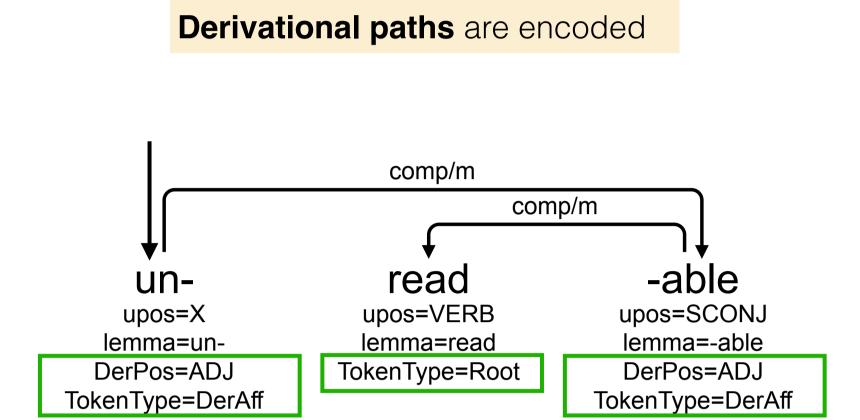
- Derivation
- Composition
- Inflection



Derivation

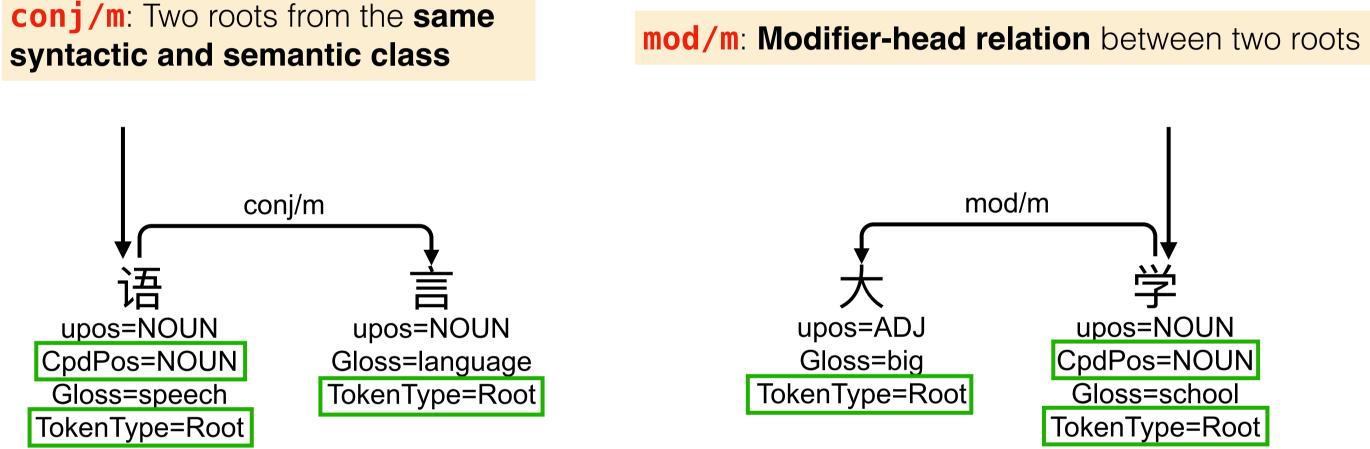


mSUD analysis of the English adverb *fiendishly*

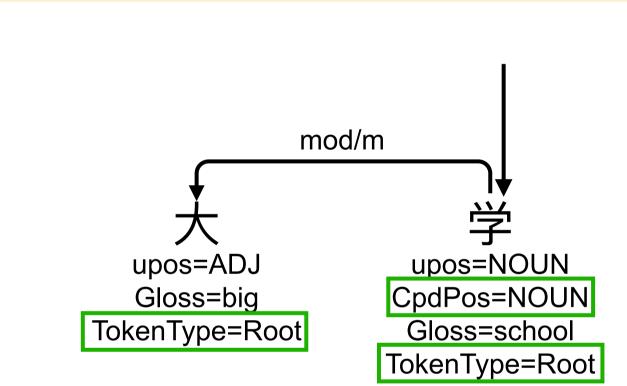


mSUD analysis of the English adjective unreadable

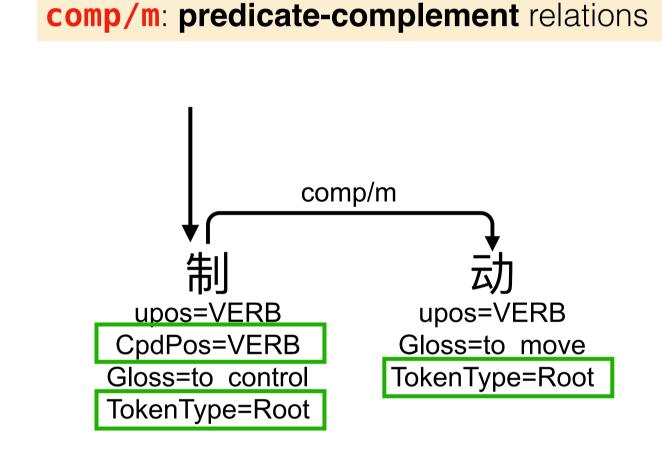
Composition



(yǔ yán) 'language', lit. speech language

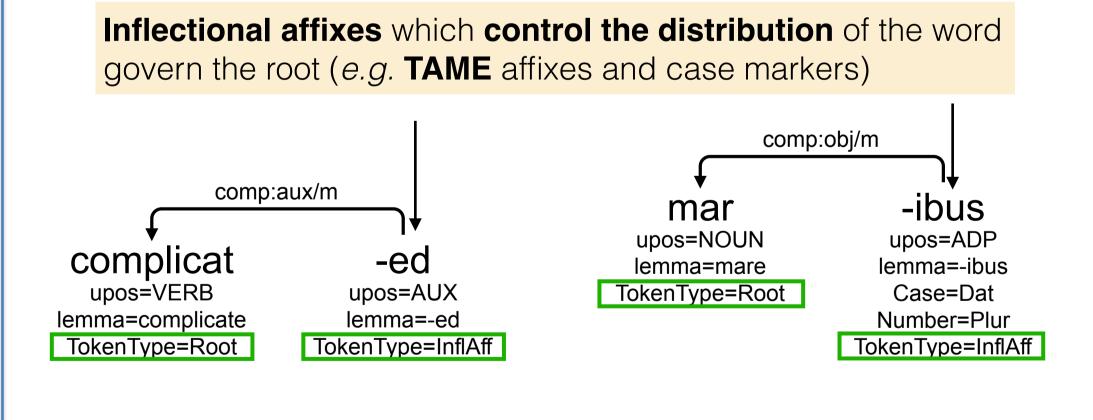


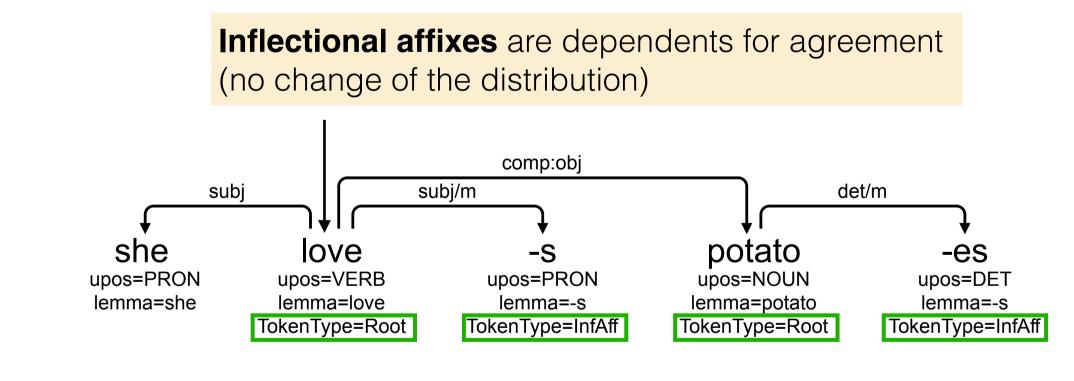
(dà xué) 'university', lit. big school



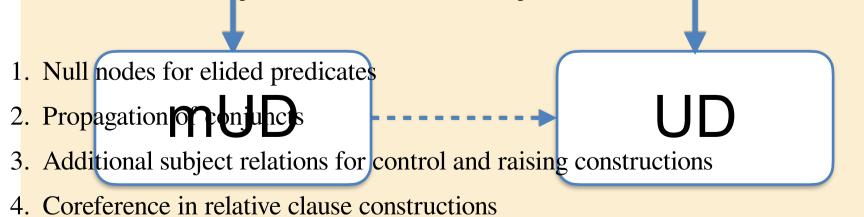
(zhì dòng) 'brake', lit. (to) control (to) move

Inflection





nature of the verb. The incorporated noun in the enhanced representation receives the feature Incorporated=Yes.



5. Modifier labels that contain the preposition or other case-marking information

In order to test our proposed annotation guidelines, we decided to approach a particular lan

cannot be Govered fryethe predicate Tuwari

4 Case study

We propose extending the guidelines for the enhanced representation to allow additional nodes for core arguments of presentation to allow additional nodes for core arguments of presentation to allow additional nodes for core arguments of propose extending the guidelines for the enhanced representation to allow additional nodes for core arguments of propose extending the guidelines for the enhanced representation to allow additional nodes for core arguments of propose extending the guidelines for the enhanced representation to allow additional nodes for core arguments. strictly null nodes — such as those used for elided predicates — as they could only be permitted to represent view or porated execution of trees such incorporated execution of trees such as that in Figure 4 where the incorporated object becomes a node in the enhanced graph.

In an outward sense, the annotation of incorporation has some relation to the annotation of pro-drop

languages, where arguments required by the predicate may not have any form in the syntax and only appear as agreement markers on the verb. However in one important sense it differs in that while for pro-drop languages the potential list of pronoung is from a finite set and can often be inferred mechanically from the verbal agreement, with incorporation the arguments are not a finite set and, barring additional annotation, each of the language of the potential list of pronoung is from a finite set and, barring additional annotation, each of the language of the potential list of pronoung is from a finite set and, barring additional annotation, each of the language of the potential list of pronoung is from a finite set and, barring additional annotation, each of the language of the potential list of pronoung is from a finite set and, barring additional annotation, each of the language of the language

arabadakiler uyuyorlar sleep.PROG.1P car.LOC-ki.PL 'The ones in the blue car are sleeping.'

Loria

Turkish inflectional groups (Çöltekin, 2016)

-kiler

upos=NOUN

lemma=-ki

Case=Nom

Number=Plur

mod/m

arabada

upos=NOUN

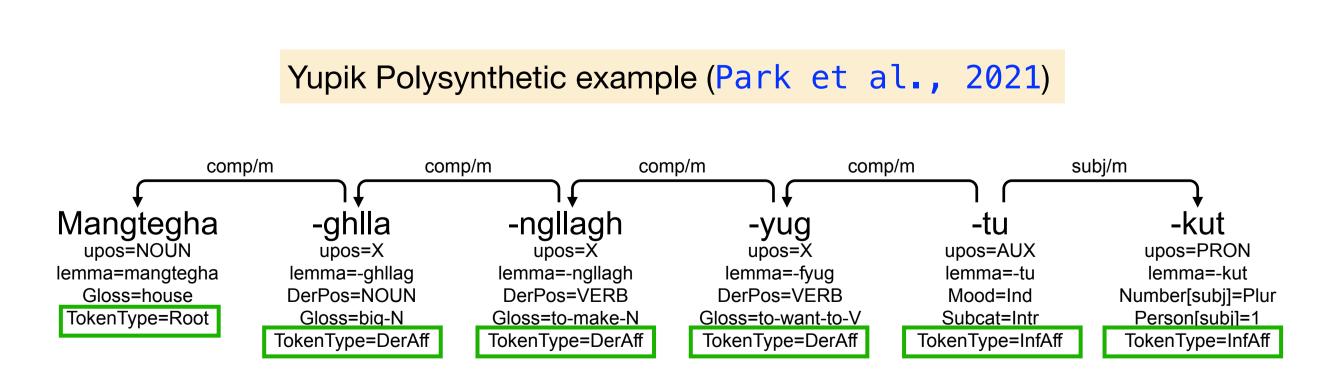
lemma=araba

Case=Loc

Number=Sing

mod

Examples



Mangteghaghllangllaghyugtukut. house-big-to.make-to.want.to-IND.INTR-1PL 'We want to make a big house.'



subj

uyuyorlar

upos=VERB

lemma=uvu

Number=Plur





