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

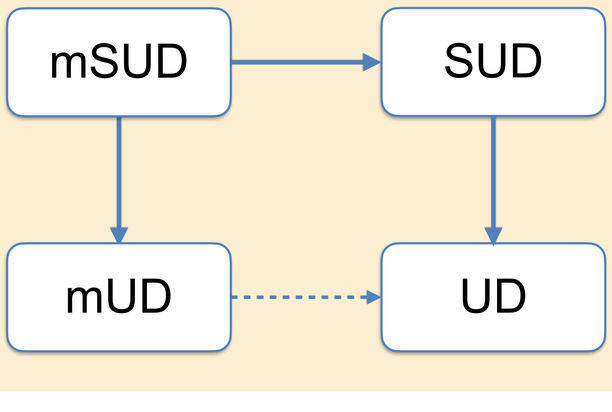
In mSUD

- **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**

Implementation



In release 2.14, three treebanks are **in mSUD**

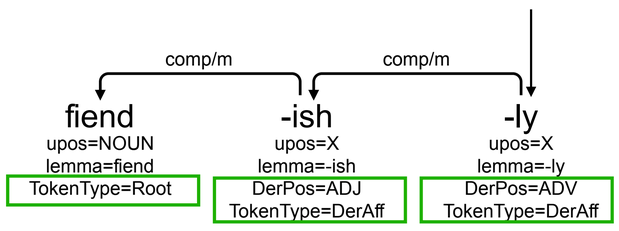
- **mSUD_Beja-NSC**
- **mSUD_Chinese-Beginner**
- **mSUD_Chinese-PatentChar**

Other treebanks are built in mSUD (IGT based)

- **Gbaya**, **Ye'kwana**, **Tuwari**

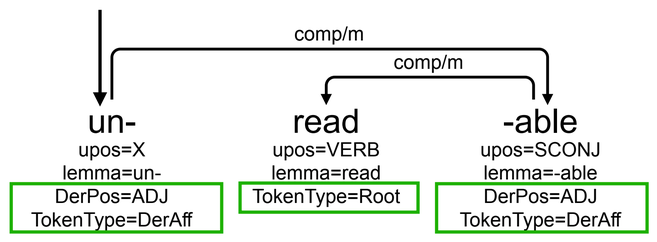
Derivation

a **derivational affix** is **the head**
it controls the **distribution** of the combination



mSUD analysis of the English adverb *fiendishly*

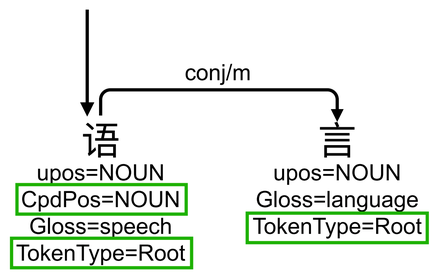
Derivational paths are encoded



mSUD analysis of the English adjective *unreadable*

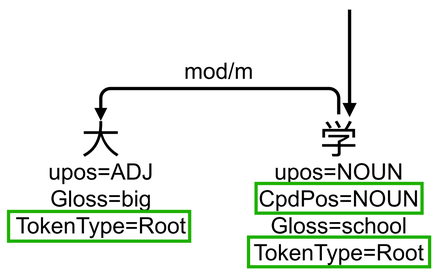
Composition

`conj/m`: Two roots from the **same syntactic and semantic class**



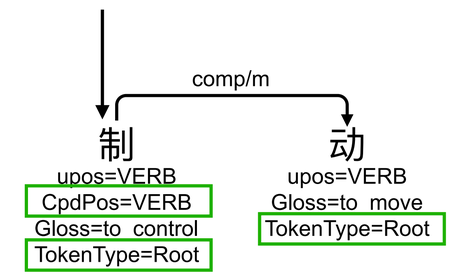
(yǔ yán) 'language', lit. *speech language*

`mod/m`: **Modifier-head relation** between two roots



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

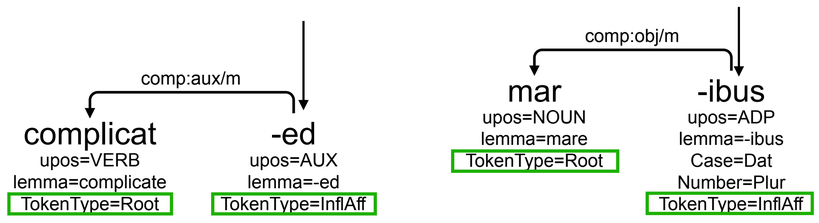
`comp/m`: **predicate-complement** relations



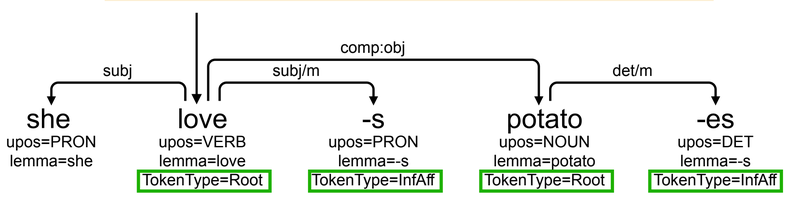
(zhì dòng) 'brake', lit. (to) control (to) move

Inflection

Inflectional affixes which **control the distribution** of the word govern the root (*e.g.* **TAME** affixes and case markers)

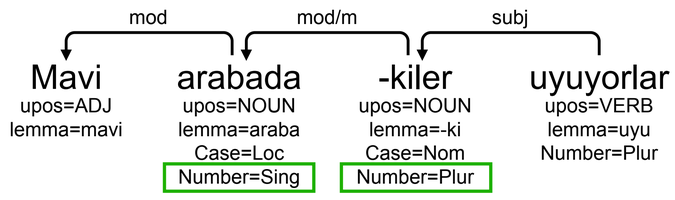


Inflectional affixes are dependents for agreement (no change of the distribution)



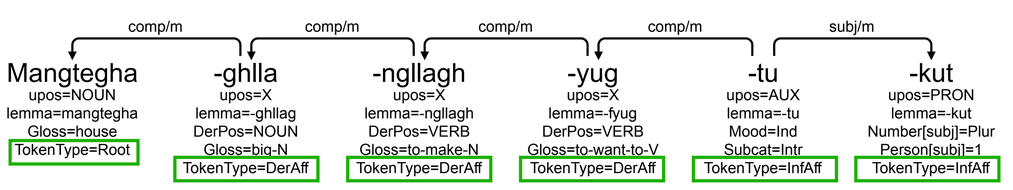
Examples

Turkish inflectional groups (Çöltekin, 2016)



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

Yupik Polysynthetic example (Park et al., 2021)



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

Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
Turino May 25, 2024

LREC-COLING 2024