Joint Annotation of Morphology and Syntax in Dependency Treebanks

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

LREC-COLING 2024

22-24 May, 2024 ••• Turino, Italia

Why annotate at the morph level?

Development of **morpho-syntactic treebanks** in many new languages

- Boosted by the **UD project**
- UD requires a **word-based level** annotation

Word level annotation is **difficult to apply** in many contexts

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

Our proposal: a **morph-level annotation** format

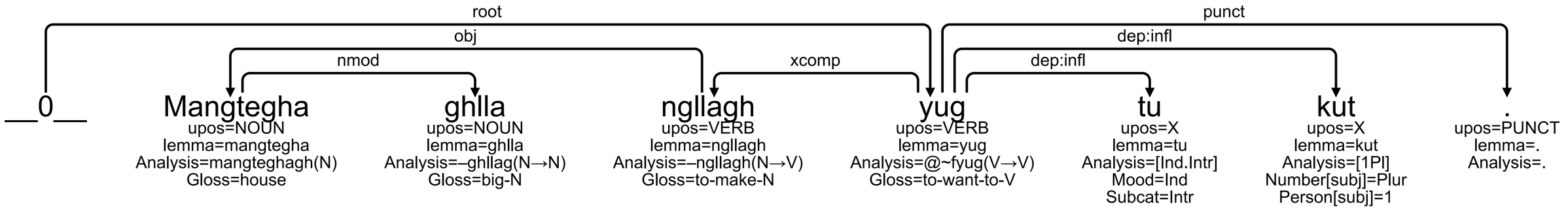
- **Convertible** to existing word-based formats
- Can be used **optionally**, only for languages or contexts where is it needed

Example with a polysynthetic language

Some UD treebanks have already used some morph-based annotation

- **UD_Yupik-SLI** Park et al., 2021

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



Our Proposal: mSUD

Allow for a morph-level annotation that can be converted to word-level

- We define **mSUD** as the morph-level annotation corresponding to the word-level **SUD**

In mSUD

- **Two types** of dependency: **regular** (e.g. `subj`) or at the **morphological** (e.g. `subj/m`)
- Tokens can be **typed** with a feature `TokenType` with main values `DerAff`, `InflAff`, `Root`
- Two new features to indicate the **final** `upos` on the corresponding word level entity:
  - `DerPos` for **derivational affixes**
  - `CpdPos` for **compounds**

Notes

- We also define **mUD** corresponding to the **UD** word-level
- By ***root***, we mean to a core segment of a word.
  This definition is different from the root, which is the head of a sentence

Our Proposal: mSUD

Three categories of **subword** annotations

- **Derivation**
- **Composition**
- **Inflection**

Notes

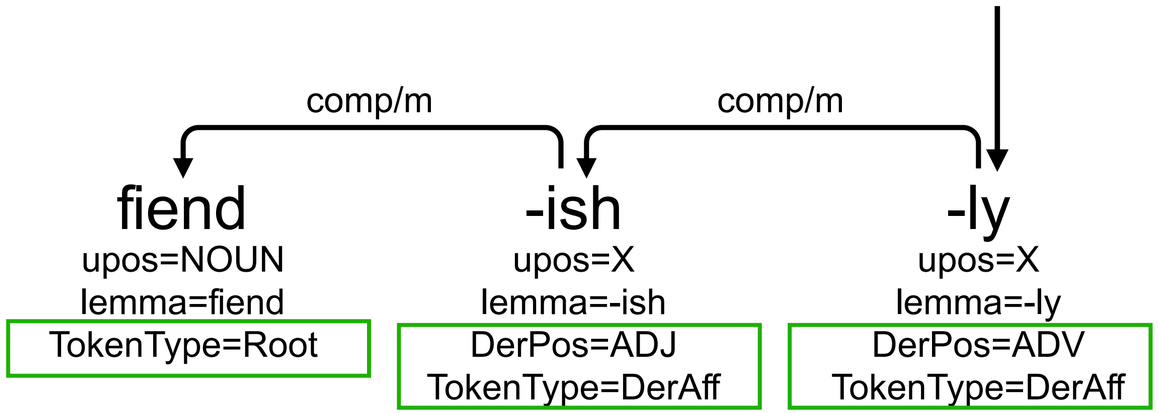
- We use some **English** examples to make it easier to read, even if the mSUD annotation is not particularly relevant to English!
- Depending on the language:
  - We may add the **‘dash’ symbol** to make suffixes explicit when annotation, e.g. when source data is *Interlinear Glossed Text* (IGT)
  - We may not add the **‘dash’ symbol** for Chinese or Japanese

Derivational affixes in mSUD

- SUD uses **distributional criteria** to select the **head** of a phrase
- The **head** of a phrase is the element that **controls its distribution**

- At the morph-level, a **derivational affix** is **the head**:
  it is the affix that decides what is the POS of the combination between a root and an affix

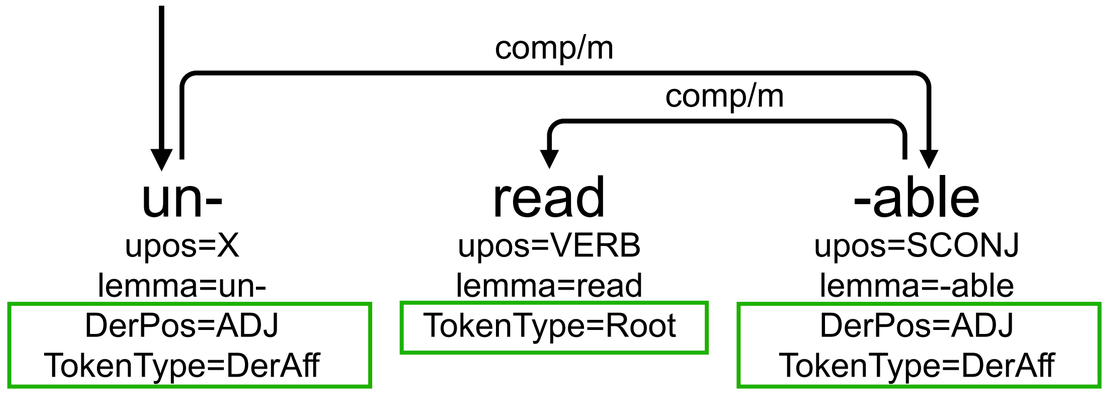
- The **`DerPos`** feature gives the POS of the resulting word



mSUD analysis of the English adverb *fiendishly*

Derivational paths in mSUD

- The analysis reveals the **internal structure of the word**
  - The root *read* combines first with the suffix *able*
  - and then with the prefix *un* (*un* cannot combine with the verbal root)
- **Derivational paths** are encoded



mSUD analysis of the English adjective *unreadable*

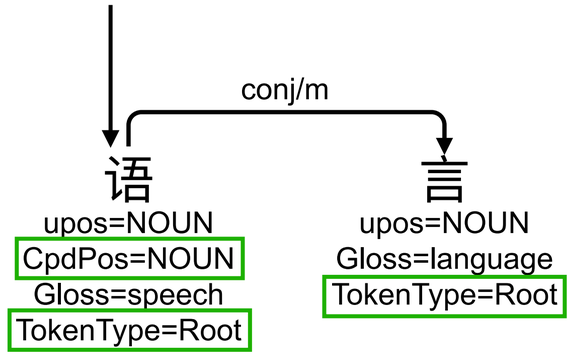
Composition in mSUD - 1/2

Compounds are words formed by **combining of two or more roots**

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

Mandarin: 语言 (yǔ yán) 'language', lit. *speech language*

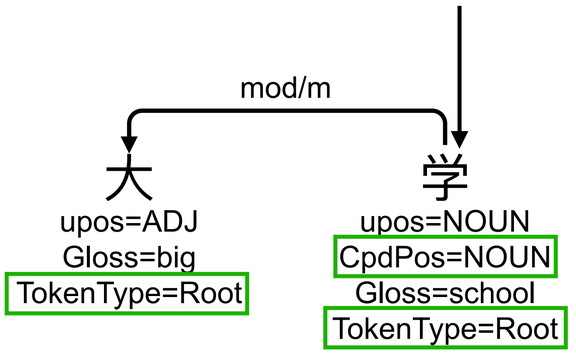
English: NOUN-NOUN wolfhound



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

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

German: ADJ-NOUN *Hochschule* 'university', lit. *high school*



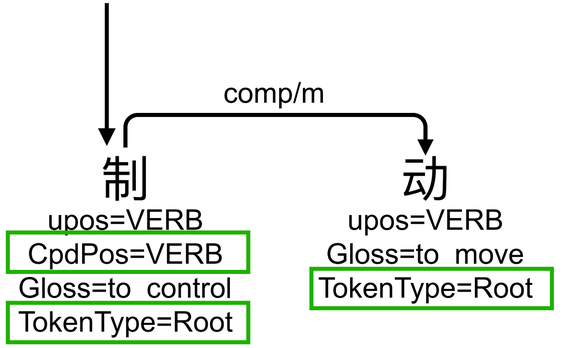
Composition in mSUD - 2/2

Compounds are words formed by **combining of two or more roots**

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

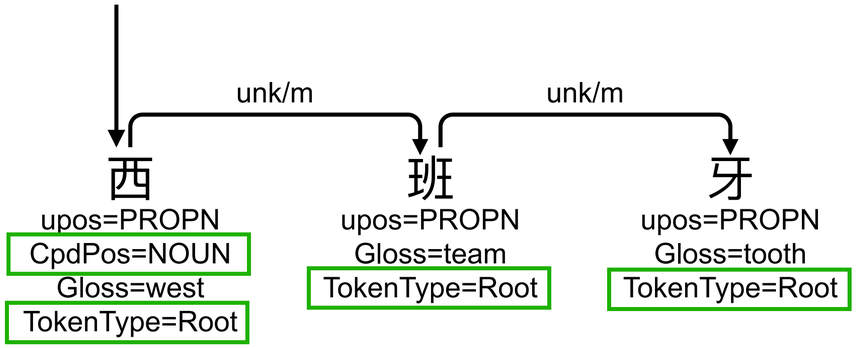
Mandarin: 制 动 (zhì dòng) ‘brake’, lit. (to) control (to) move

German: NOUN-VERB *Autofahren* ‘driving (a car)’, lit. *car driving*.



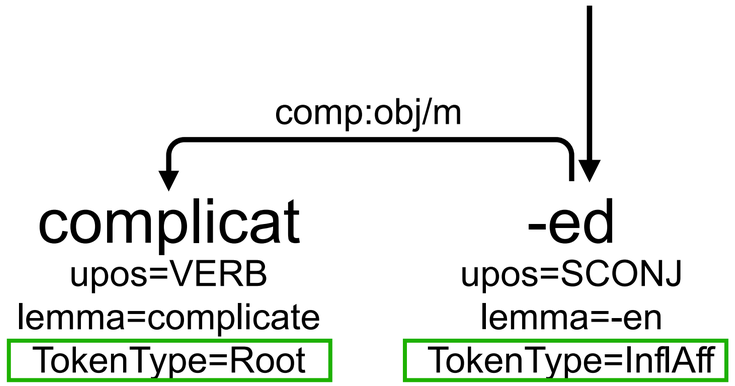
- `unk/m`: **No clear links** between roots

Mandarin: 西班牙 (xī-ba-nyá) ‘Spain’, lit. *west team tooth*

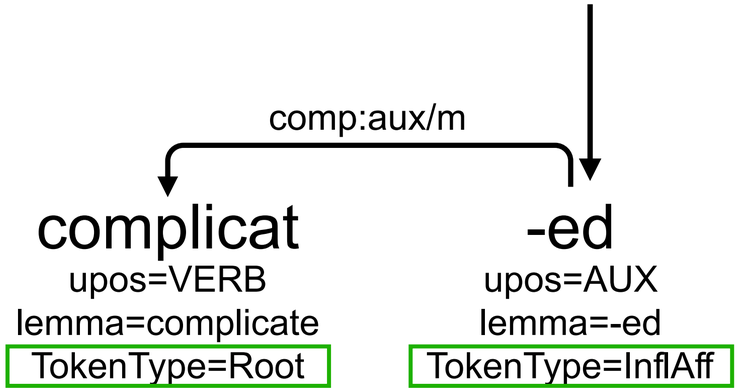


Inflection in mSUD

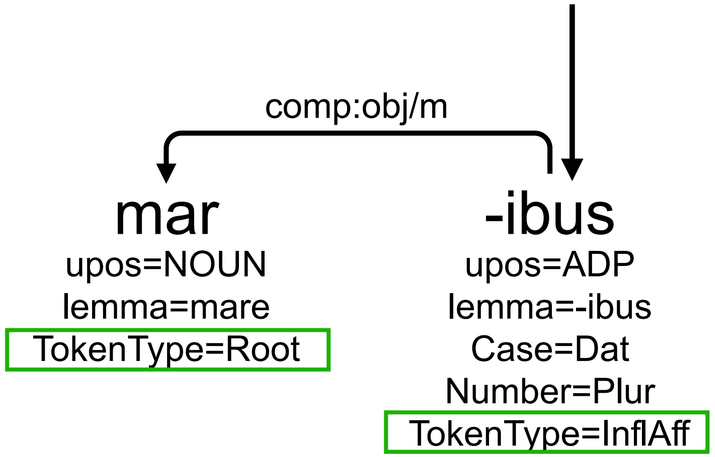
- **Inflectional affixes** govern the root when they **control the distribution** of the word
  - **TAME** affixes
  - **Case** markers



English: *complicated* (past participle)



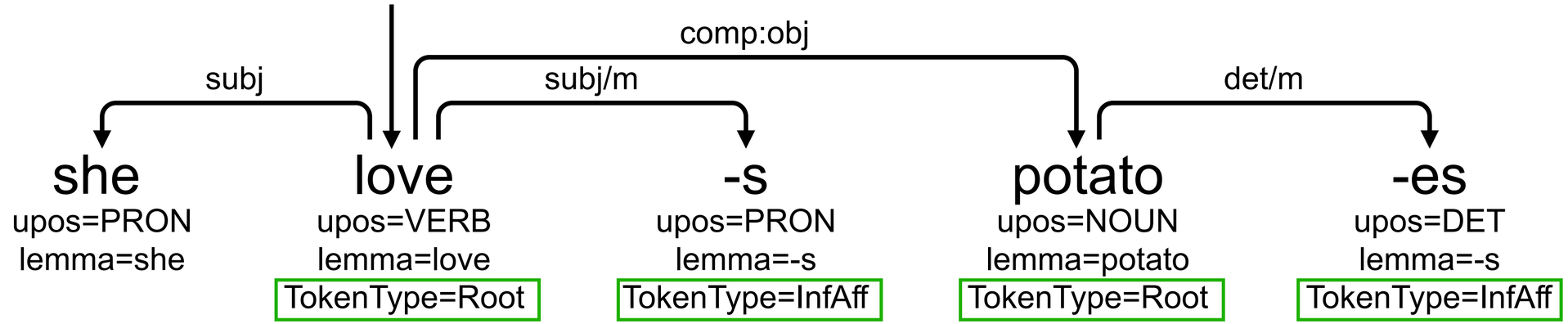
English: *complicated* (past tense)



Latin: *maribus* (dative plural)

Inflection in mSUD

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

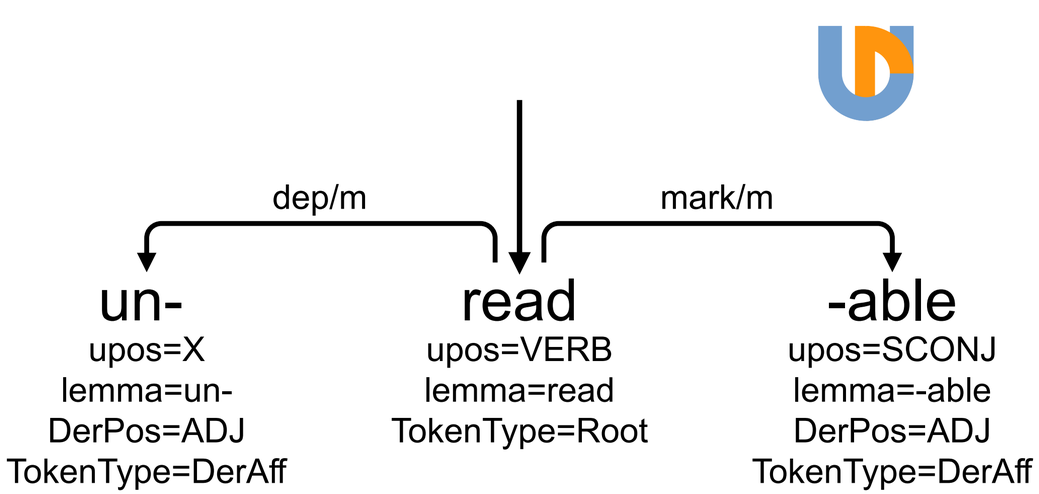


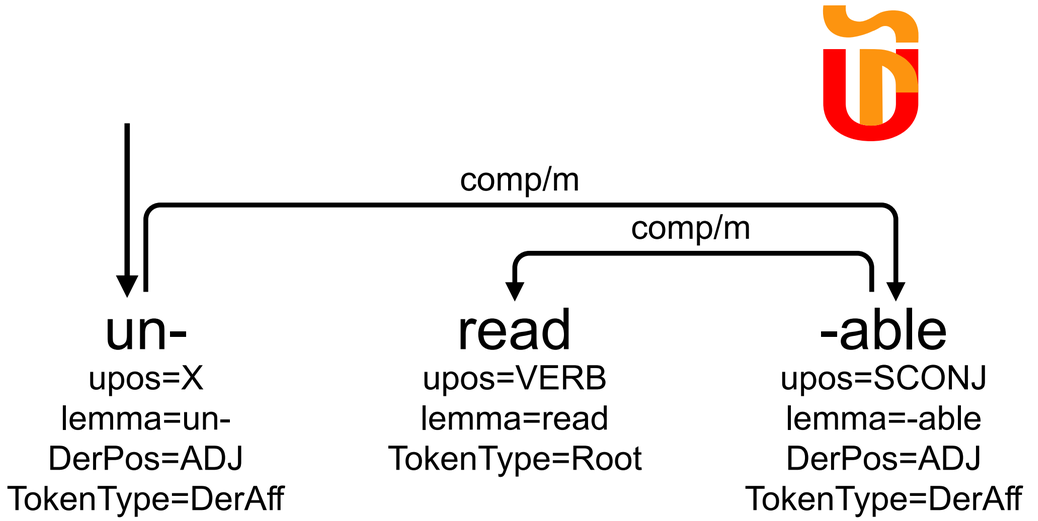
Note

- There is no need for a equivalent to **`DerPos`** or to **`CpdPos`**: the POS in unchanged in inflection

mUD: a morph-level annotation of UD

- Similarly, we can define **mUD**, a UD-style annotation at morph level
  - UD: **semantic words** are **heads ➔ root** tokens are **heads**, **affixes** are **dependents**

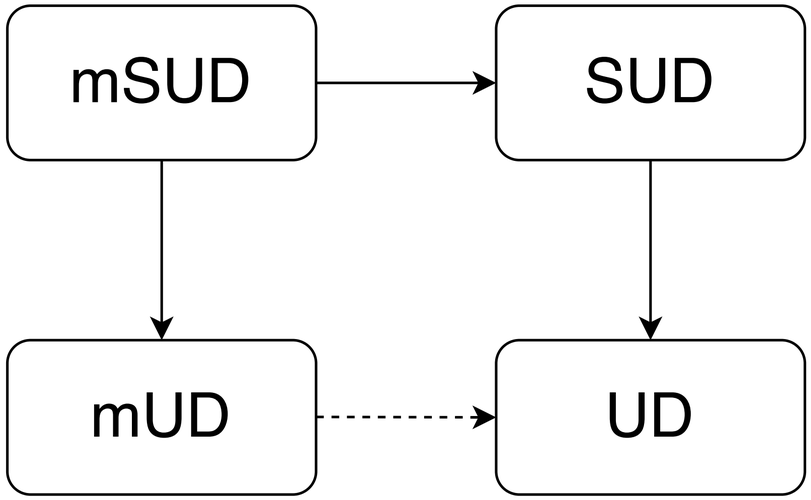




- **Derivational paths** are not fully encoded
- The **order** in which two affixes combine on the same root in **unspecified**
  - It not always possible to compute **the final POS**

Implementation

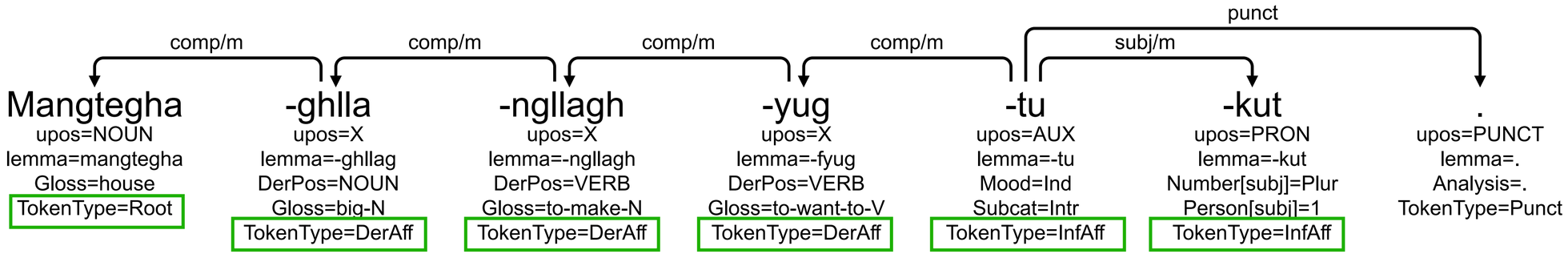
- Two types of **conversion** are used for **treebank maintenance**
  - From **morph-based** to **word-based** (horizontal arrows)
    - **Word boundaries** are encoded in the `/m` extension
    - **Final POS** are computed with `DerPos` and `CpdPos`
  - From **mSUD** to **mUD** (vertical arrows)
    - Adaptation of the conversion given in `Gerdes et al. 2018`



- 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**

Application to other treebanks

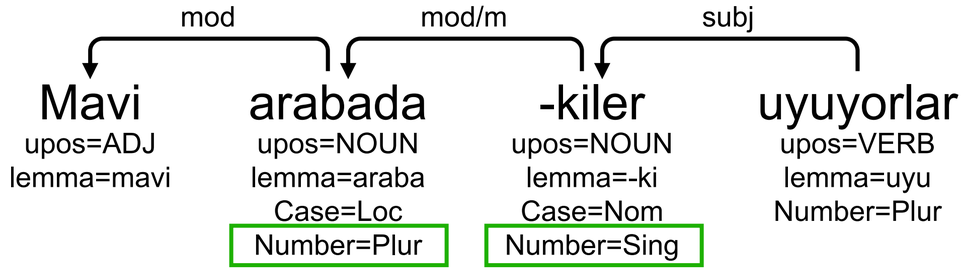
Yupik Polysynthetic example (Park et al., 2021)



Turkish inflectional groups (Çöltekin, 2016)

- Partial annotation at the morph level
  - Conflicting inflectional features
  - Different syntactic relations

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



Joint Annotation of Morphology and Syntax in Dependency Treebanks

- We have proposed an **mSUD extension** to SUD for **morph-level** based annotation
  - **SUD-style criteria** for deciding the internal mSUD structure of morphs in words
  - Encoding the **derivational path**
- Three mechanisms for describing **subword annotation**
  - Derivation
  - Composition
  - Inflection
- **Automatic convesion** to existing **word-based** formats
- A similar **mUD extension to UD** is also described
- It can be applied only for some languages or some treebanks
- Easier **inclusion of IGT-based** source data