Enrichment of Syntactic Dependency Treebanks: Two Experiments with Morphology and Prosody

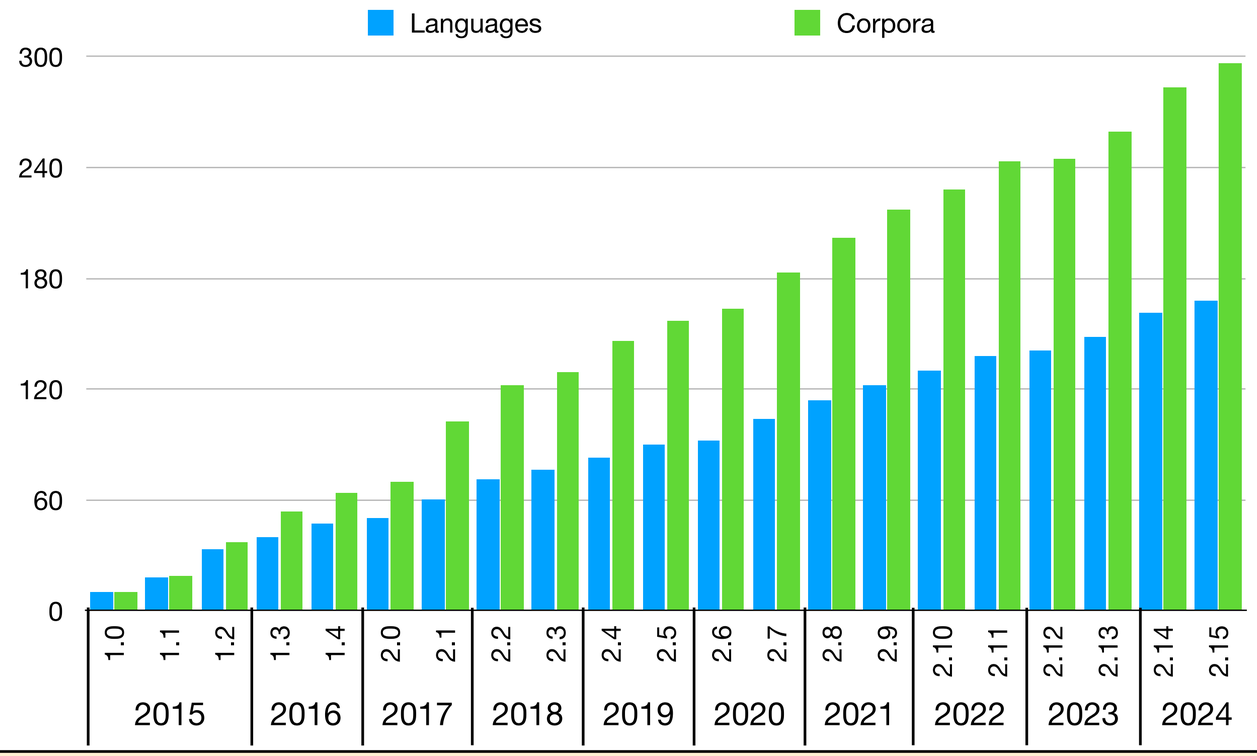
Bruno Guillaume
LORIA
Inria Centre at Université de Lorraine

Orléans -- November 14, 2024

Context

An intense activity of **treebanks production**

- Mainly driven by **Universal Dependencies** project
- New questions on the construction of treebanks for less standardised languages
- New opportunities of exploitation of treebanks for linguistic studies



Version 2.15 (tomorrow):

- 168 languages
- 296 treebanks

What kind of enrichments?

Driven by needs of linguists

- Field linguists working on spoken data
- Linguists working on interaction between prosody and syntax

Joint Annotation of Morphology and Syntax in Dependency Treebanks

Bruno Guillaume[1], Kim Gerdes[2], Kirian Guiller[3], Sylvain Kahane[3], Yixuan Li[4]

LREC-COLING 2024

New Methods for Exploring Intonosyntax: Introducing an Intonosyntactic Treebank for Nigerian Pidgin

Emmett Strickland[1,2], Anne Lacheret-Dujour[1], Sylvain Kahane[1], Marc Evrard[2]
Perrine Quennehen[1], Bernard Caron[3], Francis Egbokhare[4], Bruno Guillaume[5]

Joint Annotation of Morphology and Syntax in Dependency Treebanks

UD requires a **word-based level** annotation but word level segmentation 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

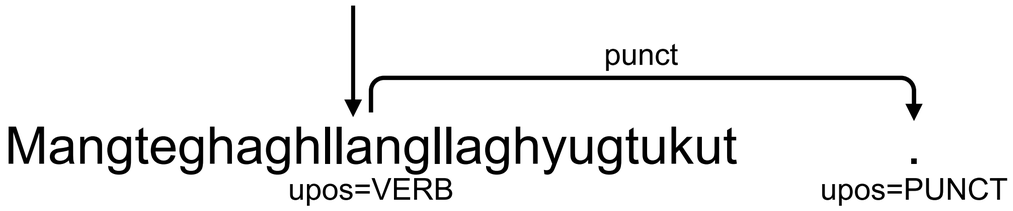
- **Convertible** to existing word-based formats
- Can be used **optionally**, only for languages or contexts where it is needed

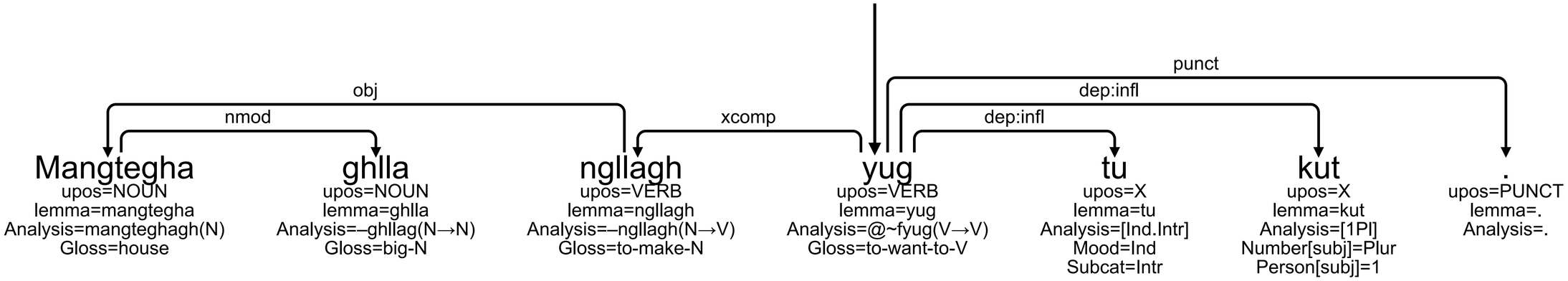
Example with a polysynthetic language

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.'





mSUD: annotation at the morph-level

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

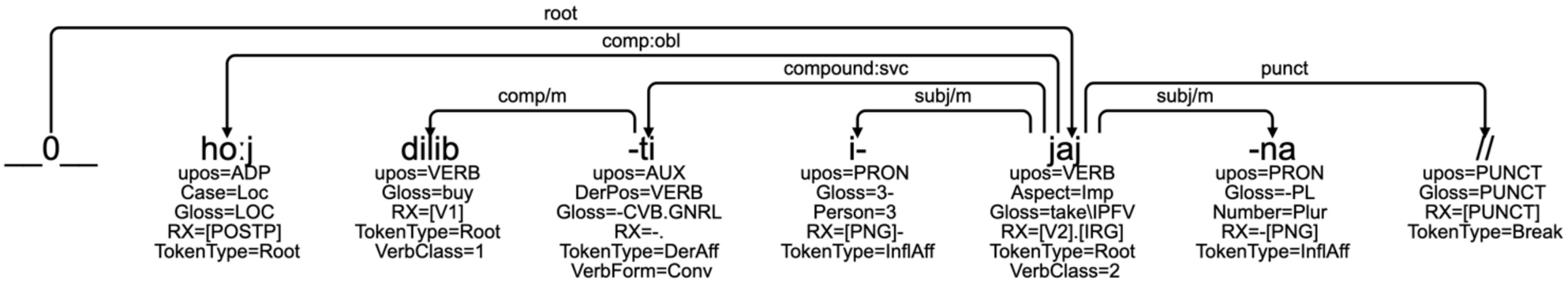
- **Two types** of dependency: **regular** (e.g. `subj`) or at the **morphological** level (e.g. `subj/m`)
- Tokens can be **typed** with a feature `TokenType` with main values `DerAff`, `InflAff`, `Root`
- Two new features to indicate the **final** `upos` on the corresponding word level entity:
  -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

Notes

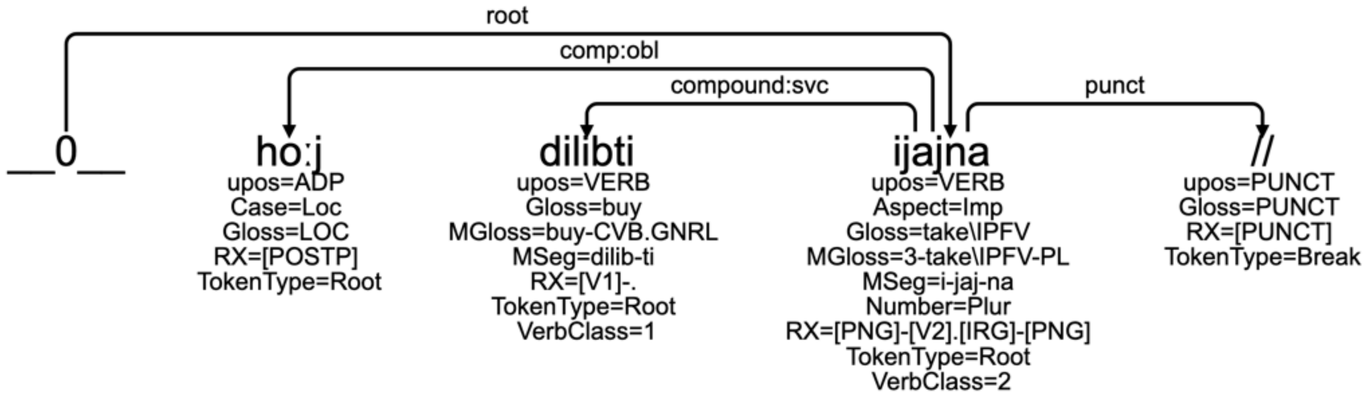
- We also define **mUD** corresponding to the **UD** word-level
- The `Root` feature designates a **core segment** of a word
  This definition is different from the dependency relation "root", which is the head of a sentence

mSUD: annotation at the morph-level

ho:j dilibti ijajna // [en: *buy them from him.*]



mSUD_Beja-Autogramm



SUD_Beja-Autogramm

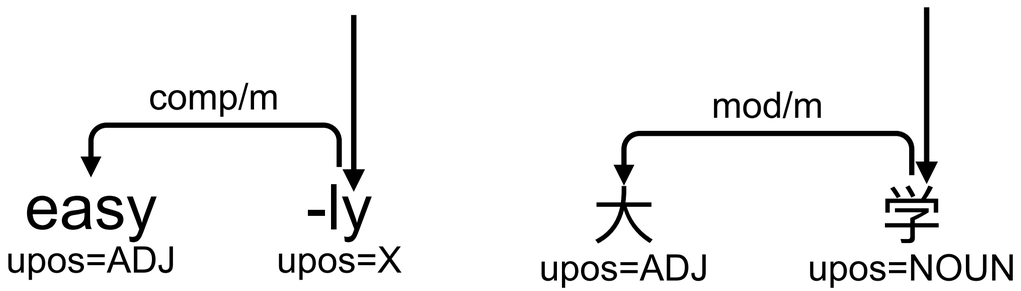
mSUD: annotation at the morph-level

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!
- Language-dependent conventions
  - We add the **'dash' symbol** to make affixes explicit, e.g. when source data is Interlinear Glossed Text (IGT)
  - We do not add the **'dash' symbol** for Chinese or Japanese

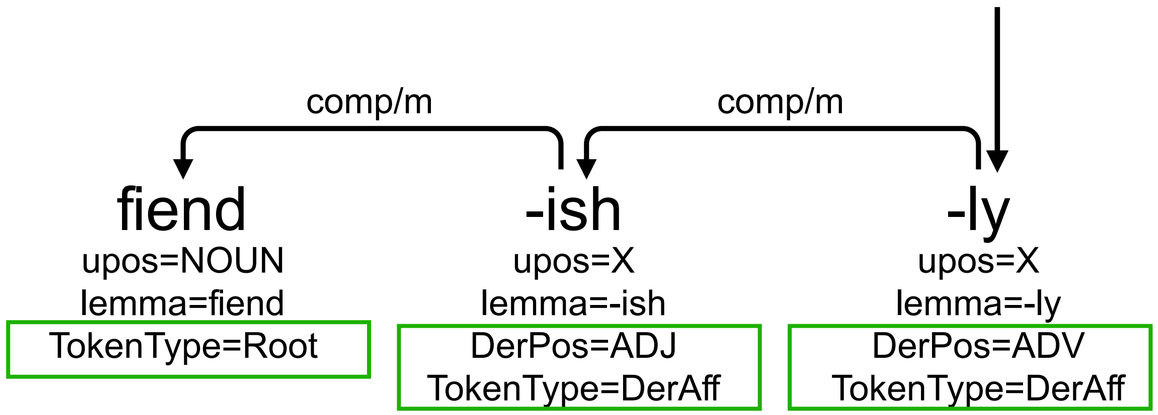


Derivational affixes in mSUD

- 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**:
  The affix determines the POS of the combined morphemes (e.g. a root and an affix)

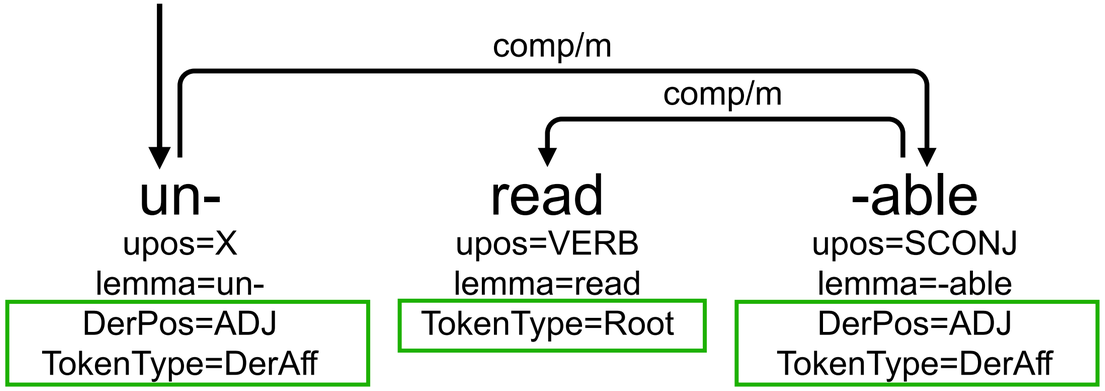
- The **`DerPos`** feature gives the POS of the resulting word



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*

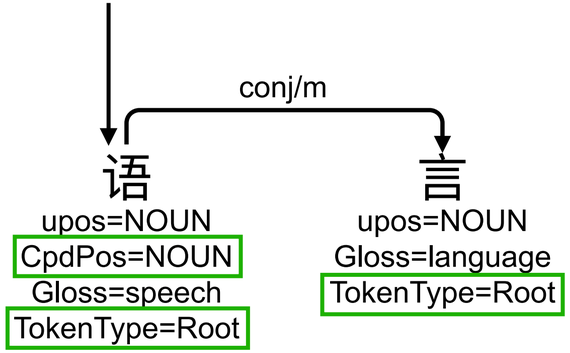
Composition in mSUD - 1/2

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*

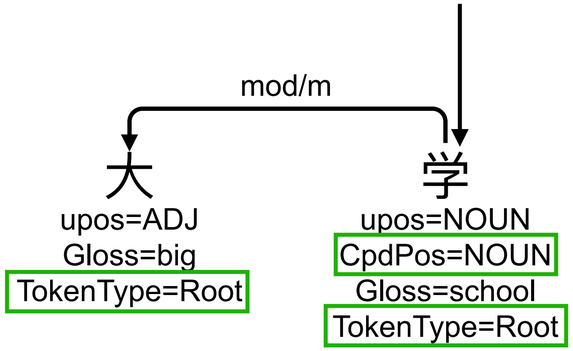
English: NOUN-NOUN wolfhound



- `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*



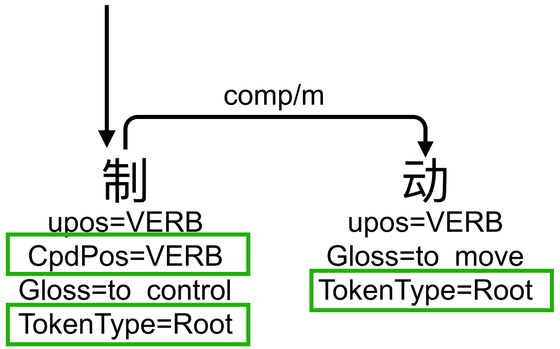
Composition in mSUD - 2/2

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

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

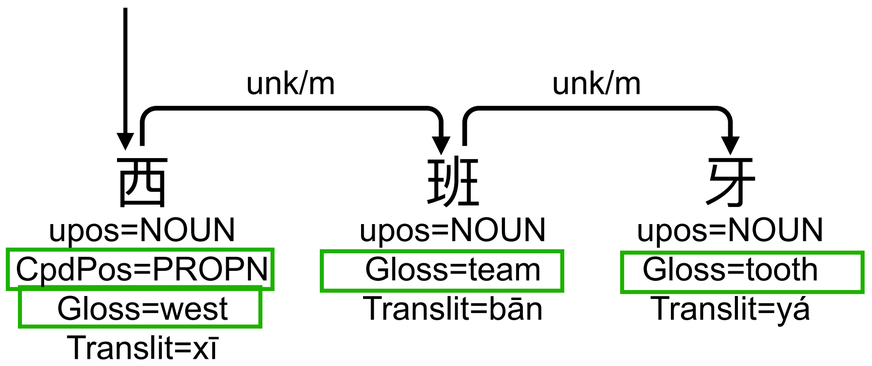
Mandarin: 制 动 (zhì dòng) ‘brake’, lit. (to) control (to) move

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



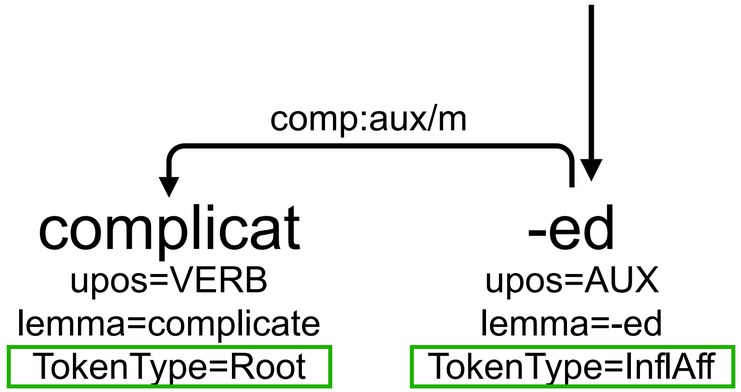
- `unk/m`: **No clear links** between roots

Mandarin: 西班牙 (xībānyá) ‘Spain’, lit. *west team tooth*

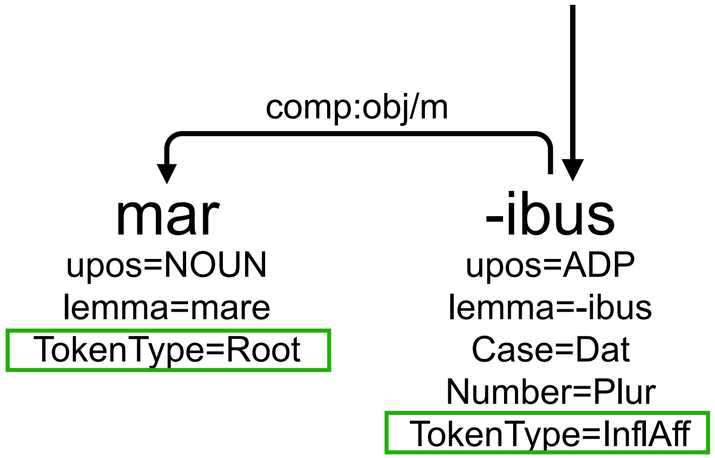


Inflection in mSUD

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



English: *complicated* (past tense)

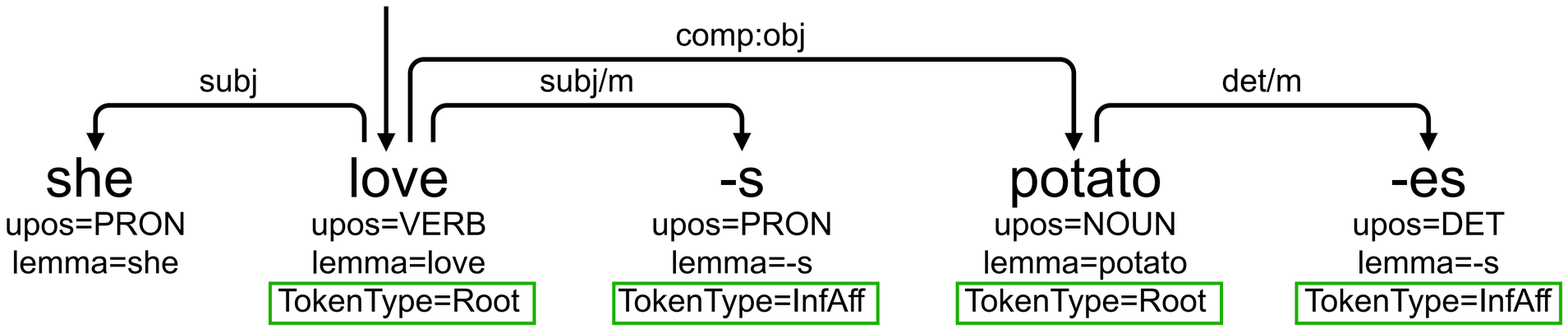


Latin: *maribus* (dative plural)
'to the seas'

Note: There is no need for a equivalent to `DerPos` or to `CpdPos`: the word POS is the one of the root

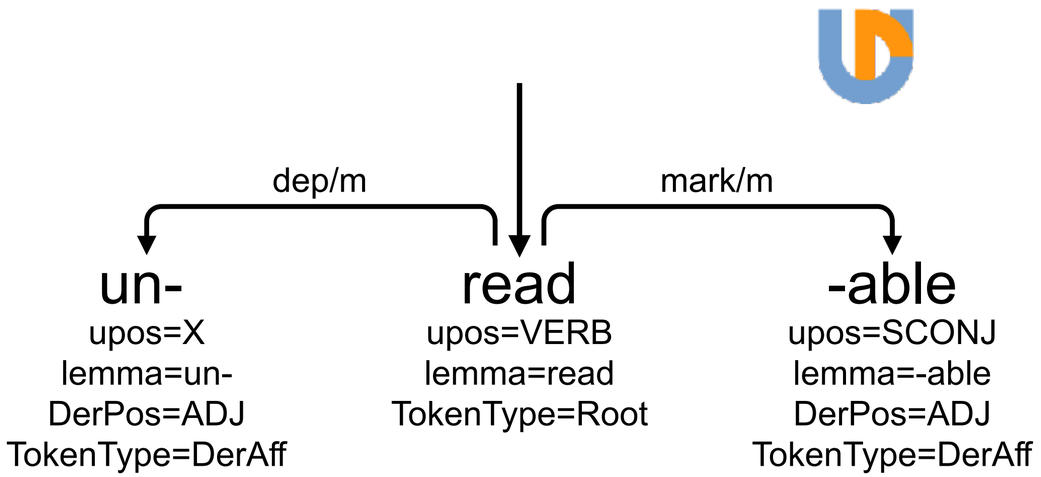
Inflection in mSUD

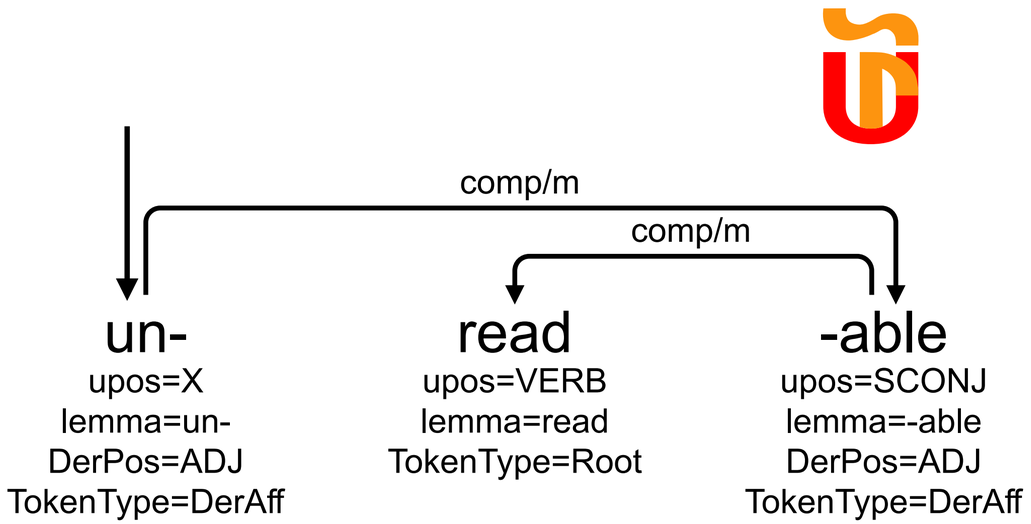
- **Inflectional affixes** are dependents for agreement (no change of the distribution)



mUD: a morph-level annotation of UD

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

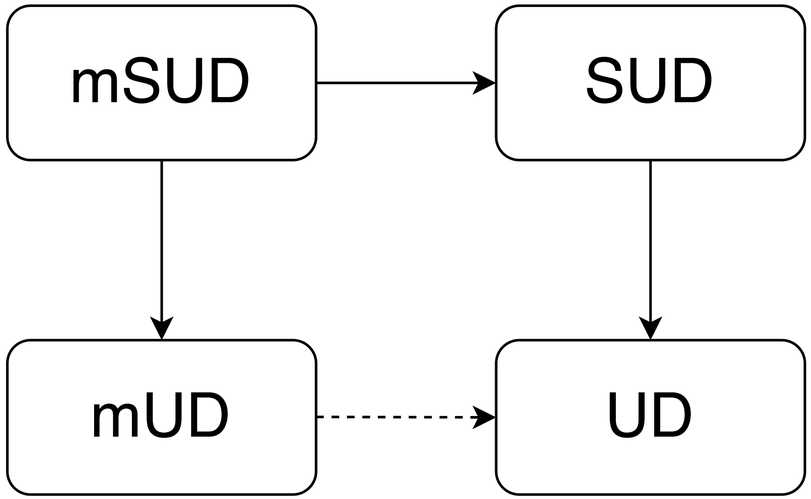




- **Derivational paths** are not fully encoded
- The **order** in which two affixes combine on the same root is **unspecified**
  - It not always possible to compute **the final POS**

Implementation

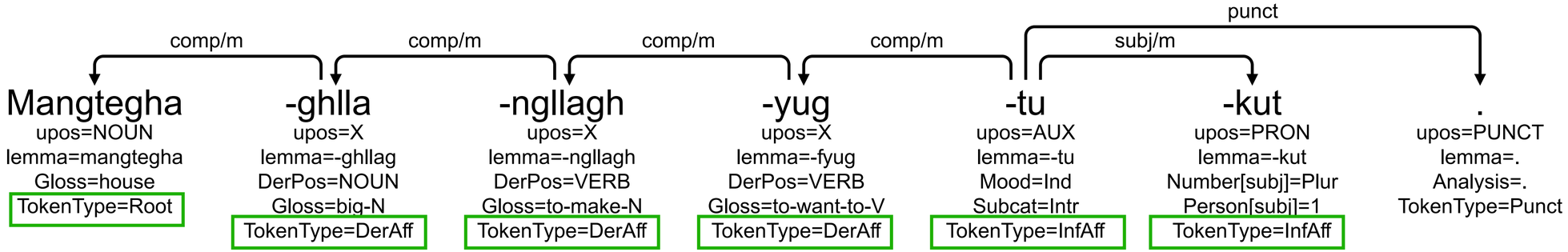
- Two types of **conversion** are used for **treebank maintenance**
  - From **morph-based** to **word-based** (horizontal arrows)
    - **Word boundaries** are encoded in the `/m` extension
    - **Final POS** are computed with `DerPos` and `CpdPos`
  - From (**m**)**SUD** to (**m**)**UD** (vertical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



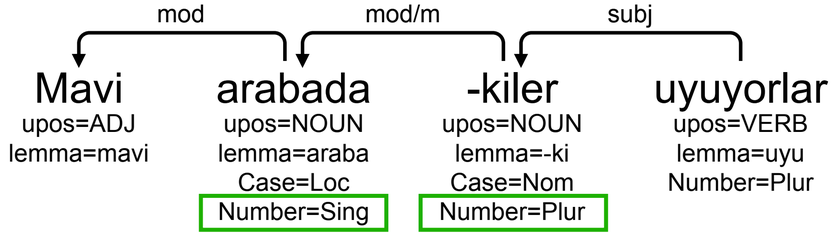
- In release 2.14, three treebanks are **in mSUD**
  - **mSUD_Beja-NSC** grewmatch
  - **mSUD_Chinese-Beginner**
  - **mSUD_Chinese-PatentChar** grewmatch

Application to other treebanks

Yupik Polysynthetic example (Park et al., 2021)



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



Turkish inflectional groups (Çöltekin, 2016)

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

Recent application of mSUD on new data

In release 2.15, three new or augmented treebanks are **in mSUD**

- **mSUD_Beja-Autogramm** much larger **mSUD_Beja-NSC** (with Martine Vanhove)
- new **mSUD_Pesh-ChibErgIS** (with Natalia Cáceres)
- new **mSUD_Northwest_Gbaya-Autogramm** (with Paulette Roulon-Doko)

Other uses of the mSUD format

- Three parallel treebanks in **Mandarin**, **Teochew** and **Taigi** with Pierre Magistry(ANR DiLSi-HN,)
- **Tuwari** (Papua-New Guinea) with Sylvain Loiseau
- **Ika** (Columbia), **Bokota** (Panama) with Natalia's PhD students

Technical issues:

- Several ways to encode morph: `/m`, `TokenType`, `-suff`/`aff-` , `nWord`.
- Ensure consistency between different annotations
- Robustness of the process when there are partial / inconsistent annotation

Intonosyntactic Treebank for Nigerian Pidgin

New Methods for Exploring Intonosyntax: Introducing an Intonosyntactic Treebank for Nigerian Pidgin

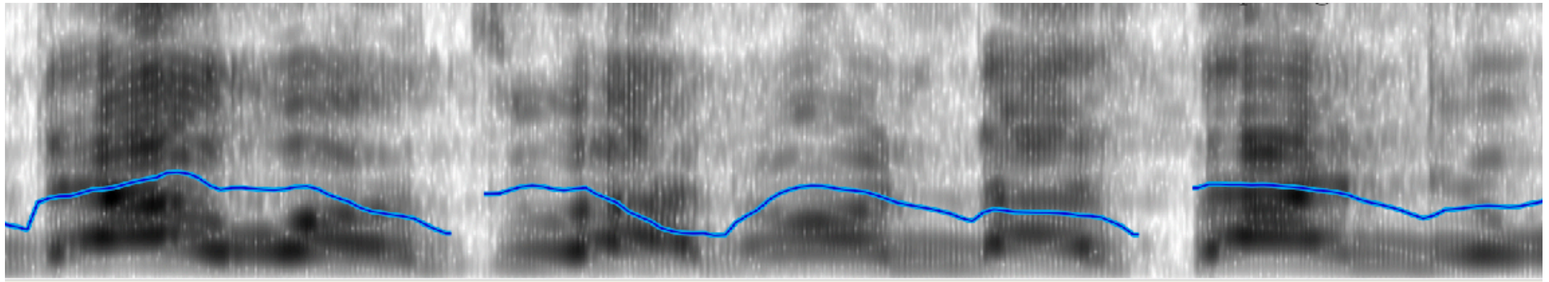
Emmett Strickland[1,2], Anne Lacheret-Dujour[1], Sylvain Kahane[1], Marc Evrard[2]
Perrine Quennehen[1], Bernard Caron[3], Francis Egbokhare[4], Bruno Guillaume[5]

NaijaSynCor: corpus of spoken Nigerian Pidgin

- 30h of transcribed speech
  - Audio recording with word and syllable-level alignments
- 7h with Gold standard syntactic annotation (SUD and UD)
  - 90K tokens and 120K syllables
  - Various genres and speech styles: storytelling, instructions, religious sermon…

Two separate annotation layers

Dat one too, I no understand o. [en: *I don't understand that.*]



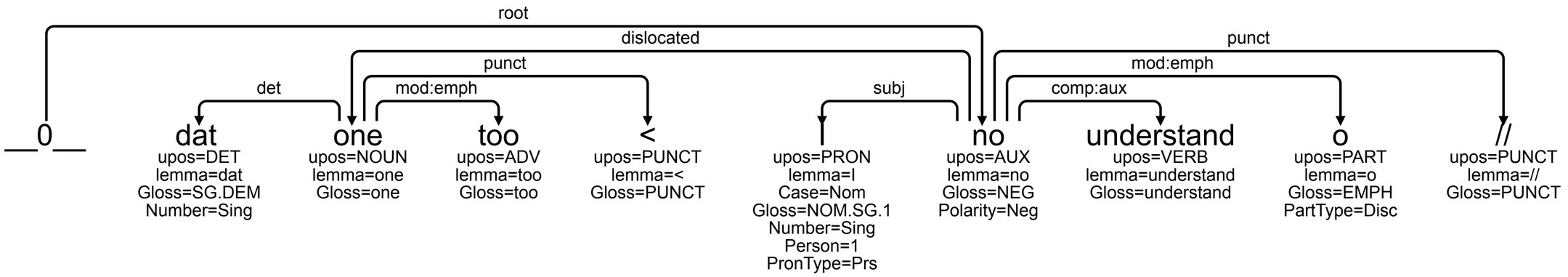
modifiable Tex

108						
dat one too < I no understand o //						
108:1	108:2	108:3	108:5	108	108:7	10
dat	one	too	I	no	understand	o
dat	one	too	i	no	understand	o
da	wO~	twa	nO	da	sta~	do
lH	hl	hl	lh	ll	mm	mm
lh	hl	hl	lh	mm	hm	mm
da	wO~	twa	nO	da	sta~	do

grewatch

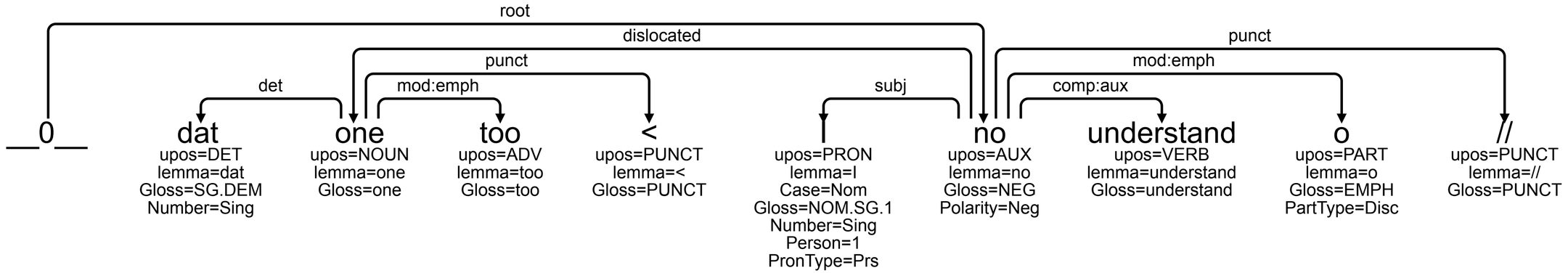
Two separate annotation layers

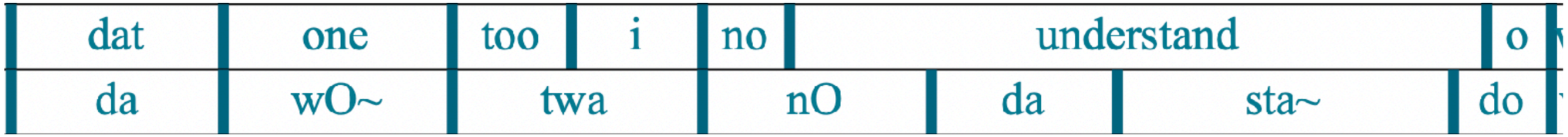
Dat one too, I no understand o. [en: *I don't understand that.*]

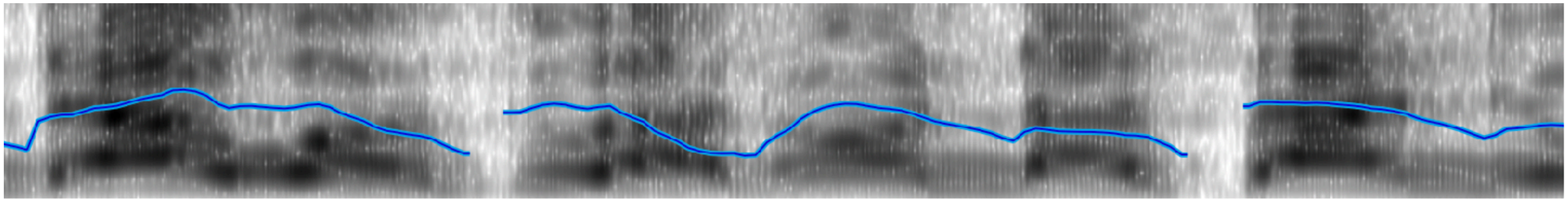


Two separate annotation layers

Dat one too, I no understand o. [en: *I don't understand that.*]

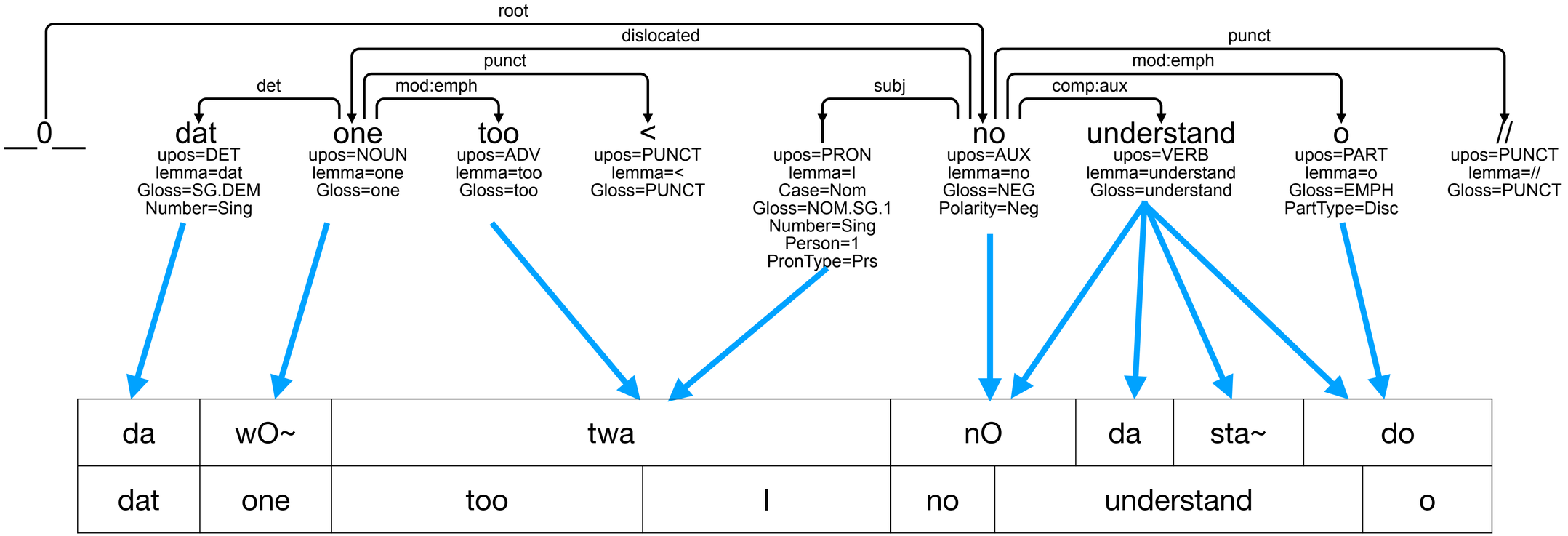






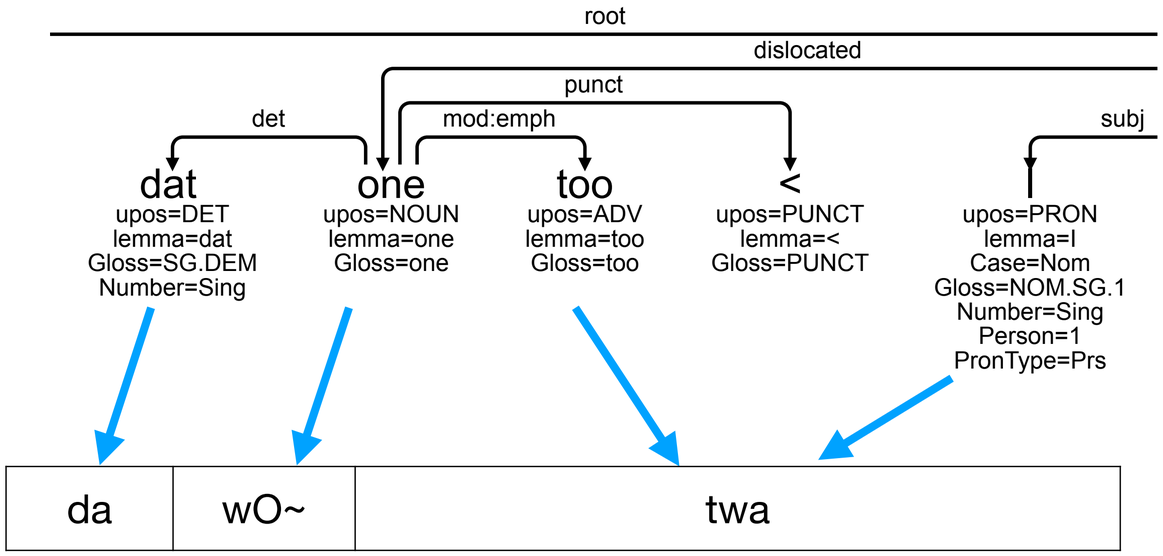
Two separate annotation layers

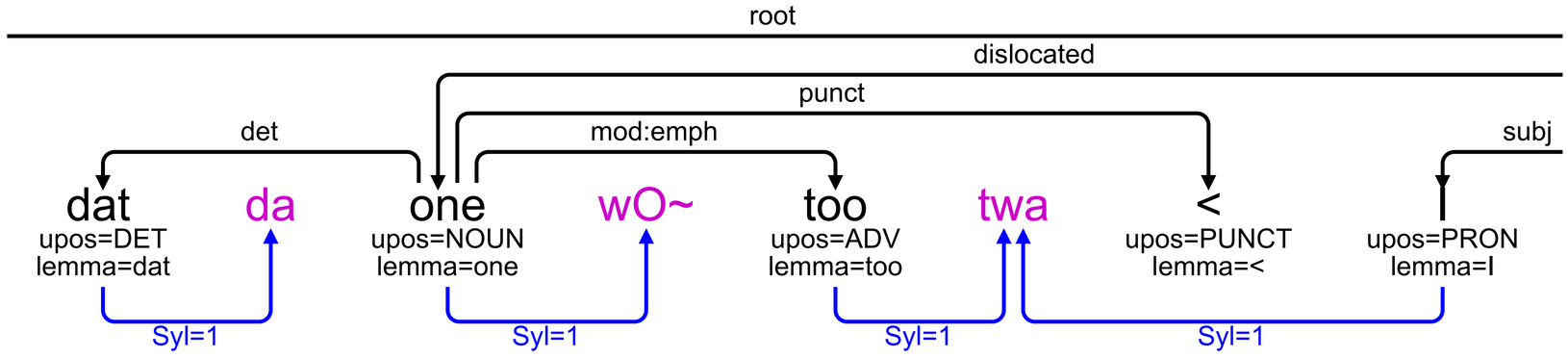
Dat one too, I no understand o. [en: *I don't understand that.*]



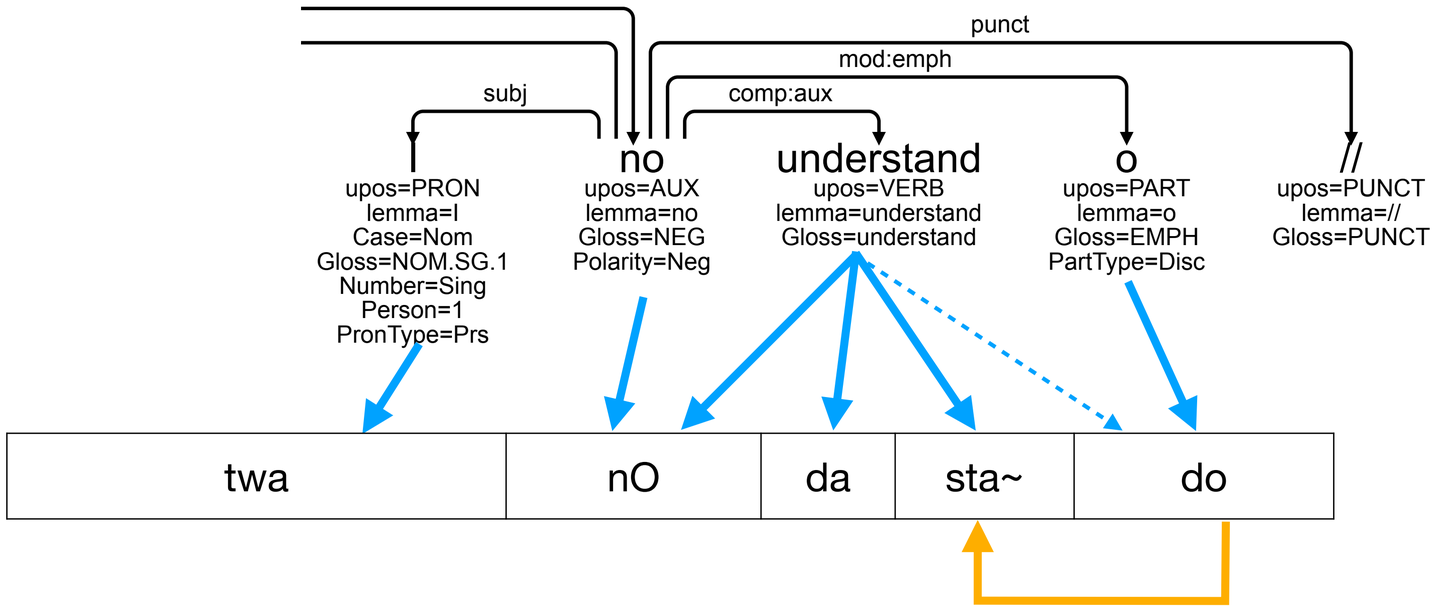
Two annotation layers in the same graph

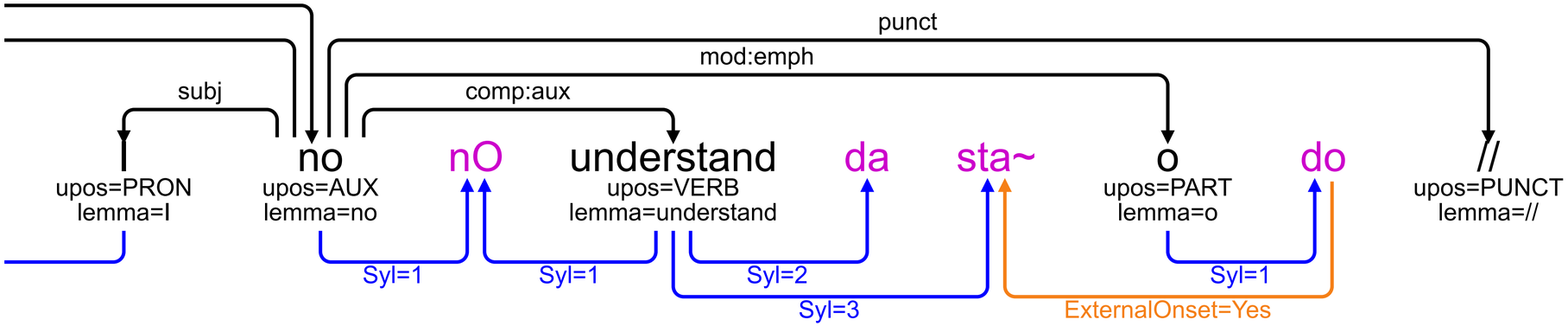
- each syllable is encoded by a new node
- special links encode the mapping node/syllable





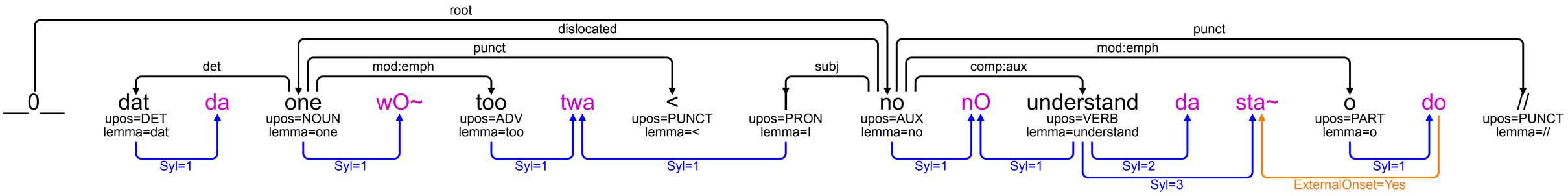
Two annotation layers in the same graph





Two annotation layers in the same graph

Dat one too, I no understand o. [en: *I don't understand that.*]



Two annotation layers in the same graph

All prosodic information availa ctures

Dat one too, I no u at.]

da
AvgAmplitude=78.057
AvgAmplitudeGlobalZscore=1.348
AvgAmplitudeLocalZscore=1.522
AvgAmplitudeNormalized=1
AvgHeightGlo=H
AvgHeightLoc=M
Coda=None
Duration=194
DurationGlobalZscore=0.041
DurationLocalZscore=-0.125
DurationNormalized=0.477
Glo=lH
Loc=lh
MaxAmplitude=80.794
MaxAmplitudeGlobalZscore=1.09
MaxAmplitudeLocalZscore=1.293

In the SUD_Naija-Prosody

- Dependency syntax (following the SUD annotation schema)

- Sociolinguistic variables (available as sentence level metadata)

- Syllable continuous variables
  - F0
  - Duration
  - Amplitude
- Syllable discrete variables
  - Slope (rise, fall, flat)
  - Stylized melodic contours using SLAM model
  - Categorical height values

speaker_age: 16-30

speaker_birthplace: Kwara

speaker_education: Tertiary

speaker_id: Sp425

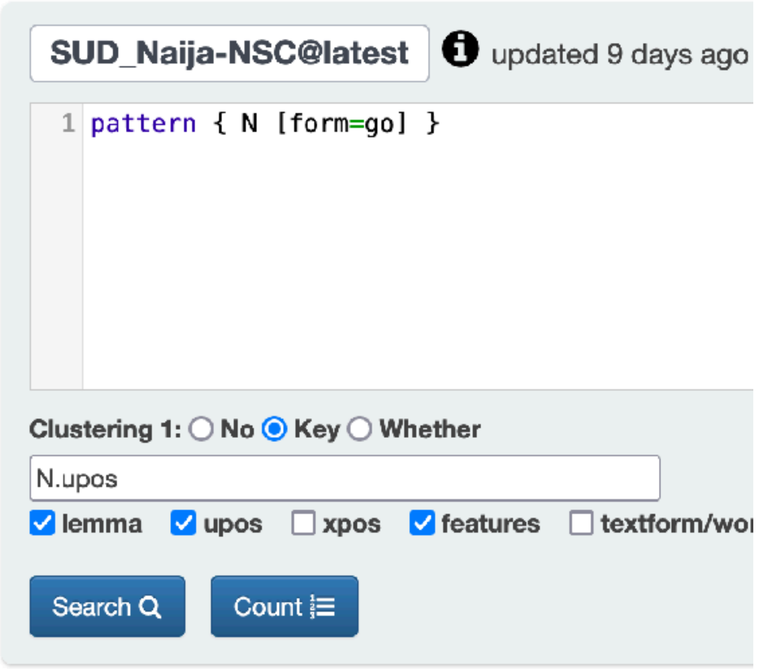
speaker_naija_competency: Excellent

speaker_primary_other_language: Yoruba

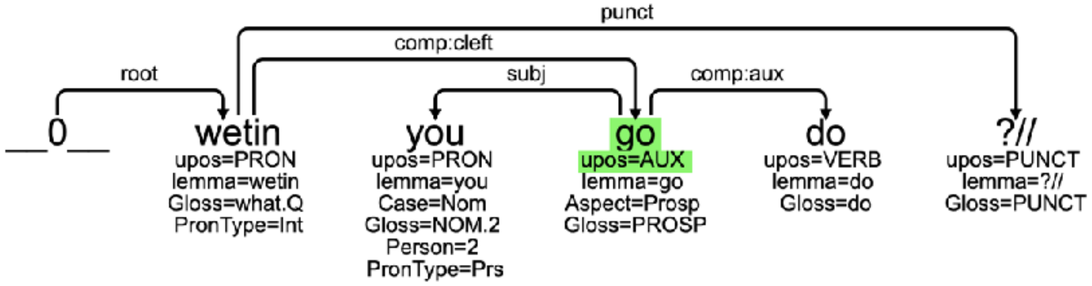
speaker_residence: Oyo

speaker_sex: M

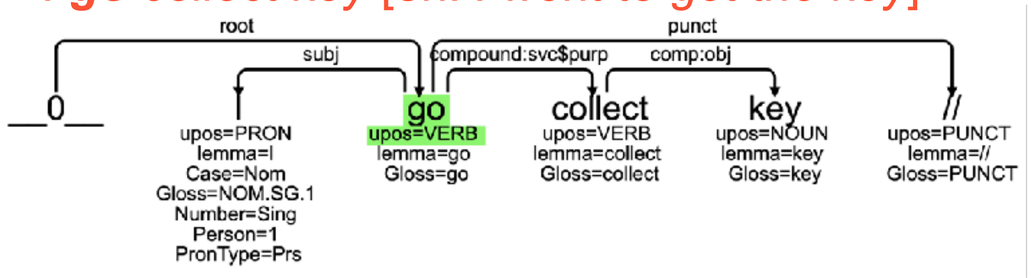
The case of go: exploring a tonal minimal pair



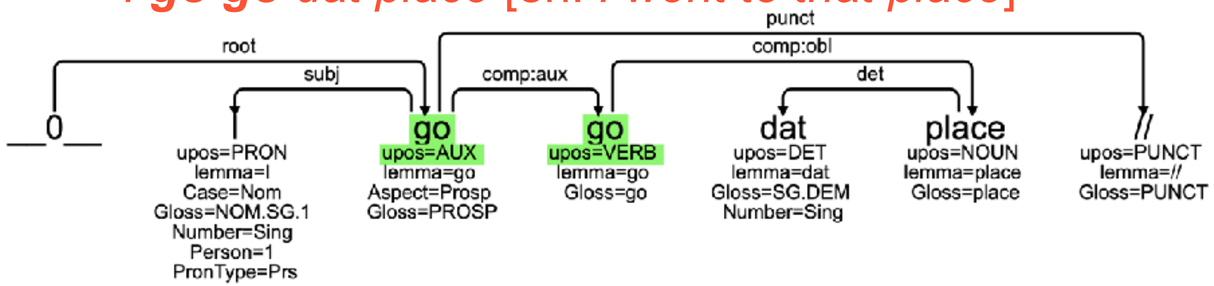
*wetin you **go** do?* [en: *What will you do?*]



*I **go** collect key* [en: *I went to get the key*]



*I **go go** dat place* [en: *I went to that place*]



The case of go: validation of a tonal minimal pair

SUD_Naija-NSC-prosody updated 9 days ago

```
pattern {
  GO [form="go"];
  GO -[Syl=1]-> S;
}
```

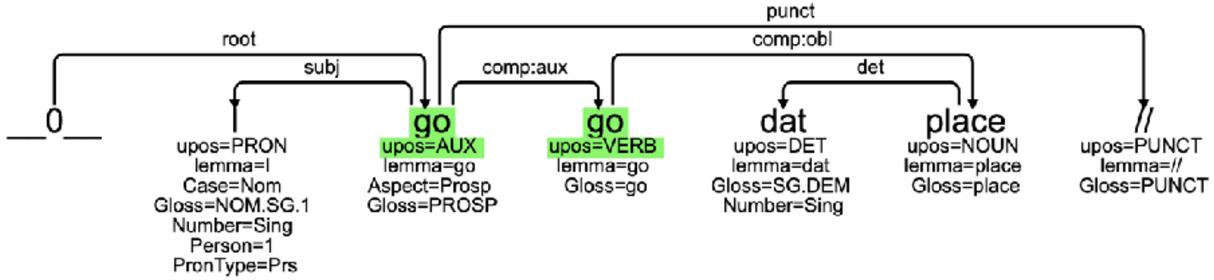
Clustering 1: No Key Whether — GO.upos

Clustering 2: No Key Whether — S.AvgHeightLoc

GO.upos \ S.AvgHeightLoc	1857 M	576 L	282 H	69 __undefined__
1930 AUX	1218	561	90	61
854 VERB	639	15	192	8

Exploring go using by comparing continuous features

```
pattern {
  G01 [form="go"]; G02 [form="go"];
  G01 -[comp:aux]-> G02;
  G01 -[Syl=1]-> S1;
  G02 -[Syl=1]-> S2;
}
```



S2.MeanF0 > S1.MeanF0?

80 Yes | 9 No

grewmatch

S2.Duration > S1.Duration?

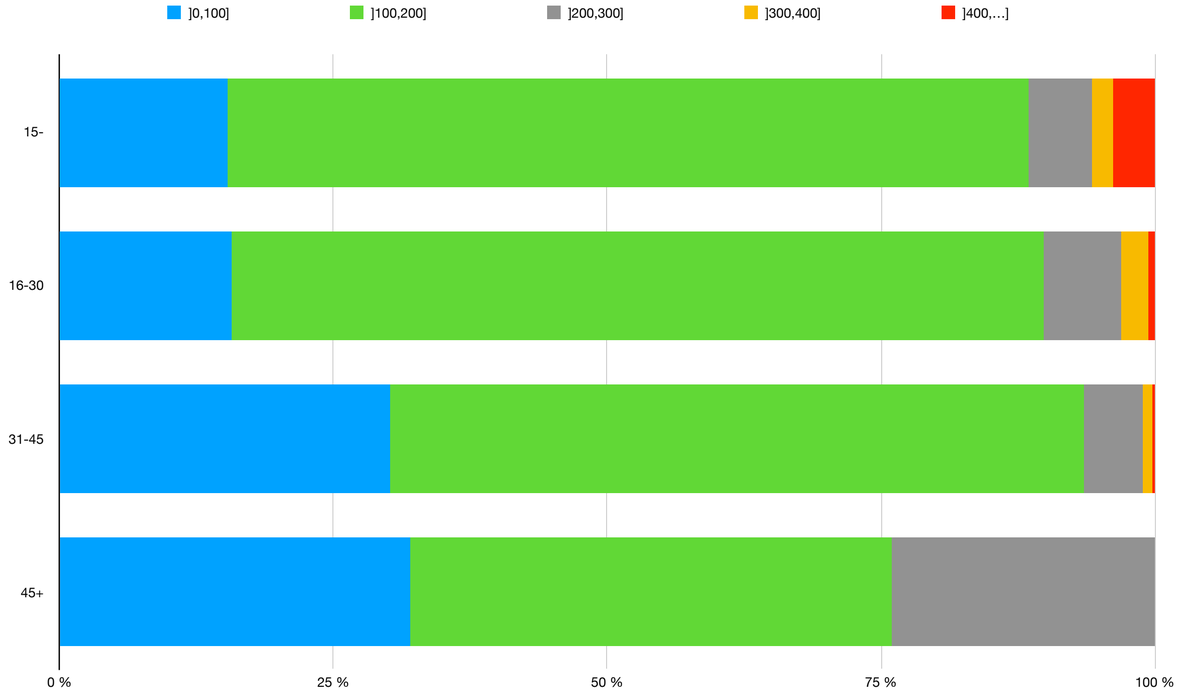
71 Yes | 18 No

grewmatch

Correlation with sociolinguistic variables

Duration of the syllable for the AUX *go* wrt the age of the speaker

	15-	16-30	31-45	45+
]0,100]	8	77	130	8
]100,200]	38	365	273	11
]200,300]	3	35	23	6
]300,400]	1	12	4	
]400,…]	2	3	1	



Work in Progress

SUD_French-Rhapsodie-prosody

How many words can be fused in one syllable?

grewatch

How syllables for one word?

grewatch

Enrichment of Syntactic Dependency Treebanks

- We propose an **mSUD extension** to SUD for **morph-level** based annotation
  - **SUD-style criteria** for deciding the internal mSUD structure of morphs in words
  - Easier **inclusion of IGT-based** source data
  - Word boundaries can be evaluated **during the annotation process**
  - Morph annotation can be applied only **partially** (mix between SUD and mSUD)

- We propose an **graph encoding** of the **syntax / prosody interface**
  - Easier exploration of the interface between annotation layers
  - Applied to **Naija-NSC** and to **French-Rhapsody**

- The process can be applied to other layers
  - UD / Parseme interface .grew .match
  - Constructions .grew .match

Thanks for your attention!

Questions?