Improving Surface-syntactic Universal Dependencies (**SUD**): MWEs and deep syntactic features

surfacesyntacticud.github.io

Treebanks and Linguistic Theories
SyntaxFest — August 26-30 2019

Kim Gerdes
Almanach (Inria), LPP (CNRS)
Sorbonne Nouvelle
kim@gerdes.fr

Bruno Guillaume
Université de Lorraine, CNRS,
Inria, LORIA, Nancy, France
bruno.guillaume@inria.fr

Sylvain Kahane
Modyco,
Université Paris Nanterre & CNRS
sylvain@kahane.fr

Guy Perrier
Université de Lorraine, CNRS,
Inria, LORIA, Nancy, France
guy.perrier@loria.fr

SUD

SUD stands for **Surface-syntactic Universal Dependencies**

- Presented in 2018 at the UD workshop
- Used in some corpus annotation tasks (Naija, French, Chinese)
- Used in some experiments presented at the SyntaxFest!

Today's presentation:

- Recall the SUD principles
- Refinement with deep syntactic features on edges
- Encoding of MWEs in SUD

General principles of **SUD**

SUD follows UD on:

- Tokenisation
- POS tagging
- Morphological features

SUD departs from UD on dependency relations definition:

- Heads definition
- Set of relations

SUD heads

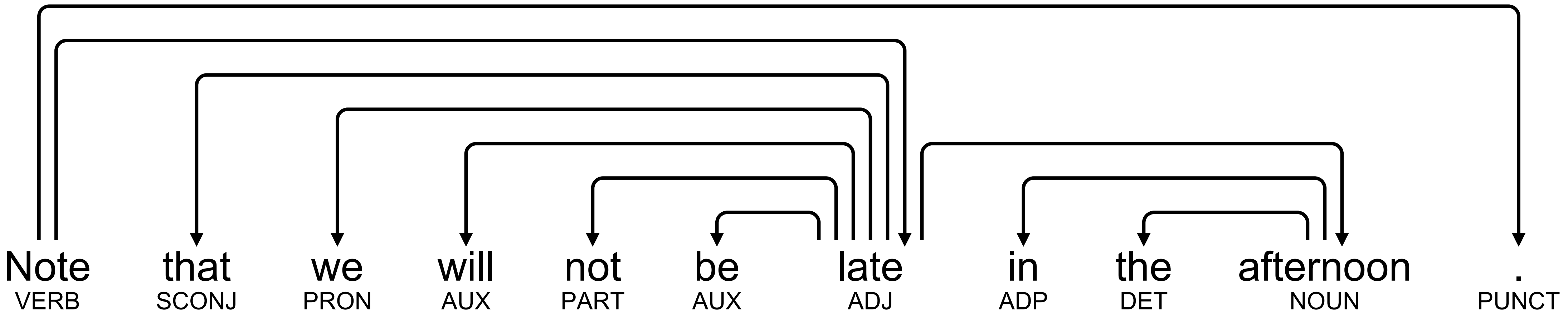
Heads:

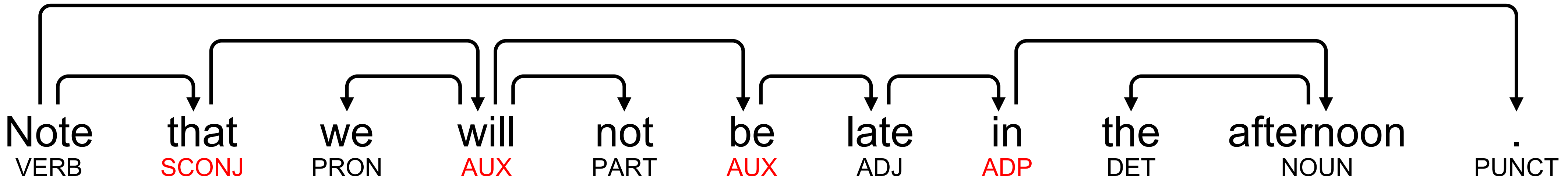
- Distributional criteria (Bloomfield, Hudson, Mel’čuk) favours functional heads ⇒ ADP, AUX, SCONJ are heads
- String analysis of coordination

SUD heads

Heads:

- Distributional criteria (Bloomfield, Hudson, Mel’čuk) favours functional heads ⇒ **ADP, AUX, SCONJ are heads**
- String analysis of coordination

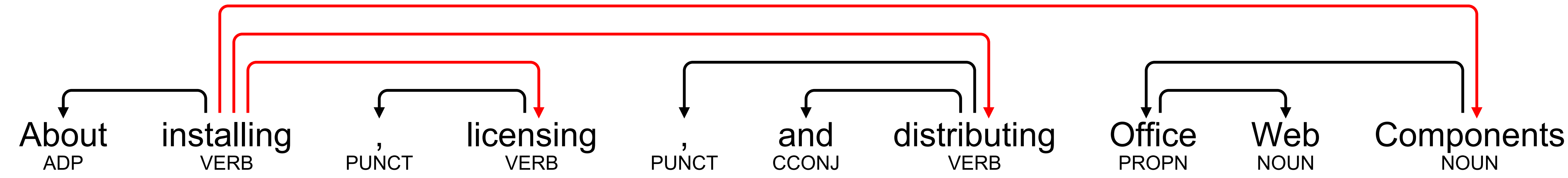


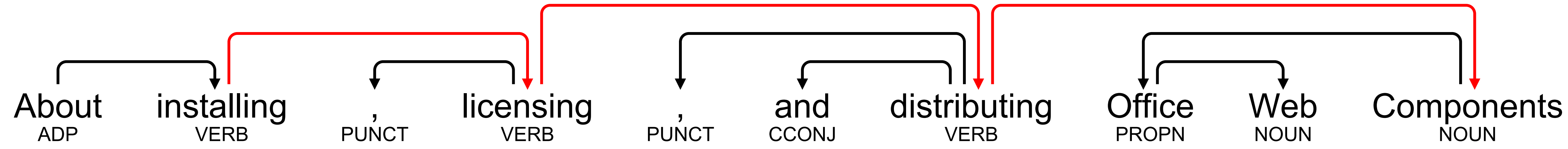


SUD heads

Heads:

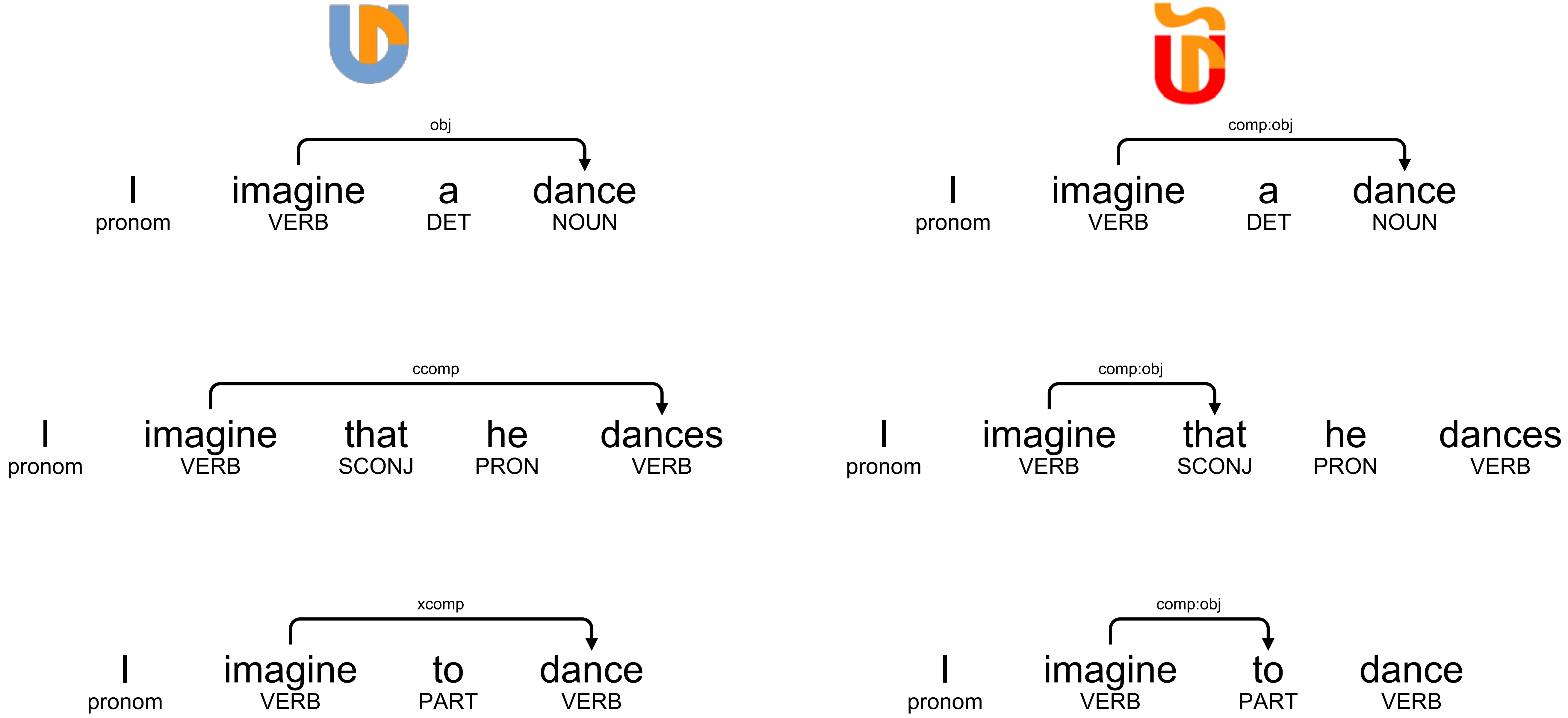
- Distributional criteria (Bloomfield, Hudson, Mel’čuk) favours functional heads ⇒ ADP, AUX, SCONJ are heads
- **String analysis of coordination**





SUD relations

- **Functional criteria**: Two units that commute in the same syntactic position must be linked to their governor by the same relation



obj

I imagine a dance

pronom VERB DET NOUN

comp:obj

I imagine a dance

pronom VERB DET NOUN

ccomp

I imagine that he dances

pronom VERB SCONJ PRON VERB

comp:obj

I imagine that he dances

pronom VERB SCONJ PRON VERB

xcomp

I imagine to dance

pronom VERB PART VERB

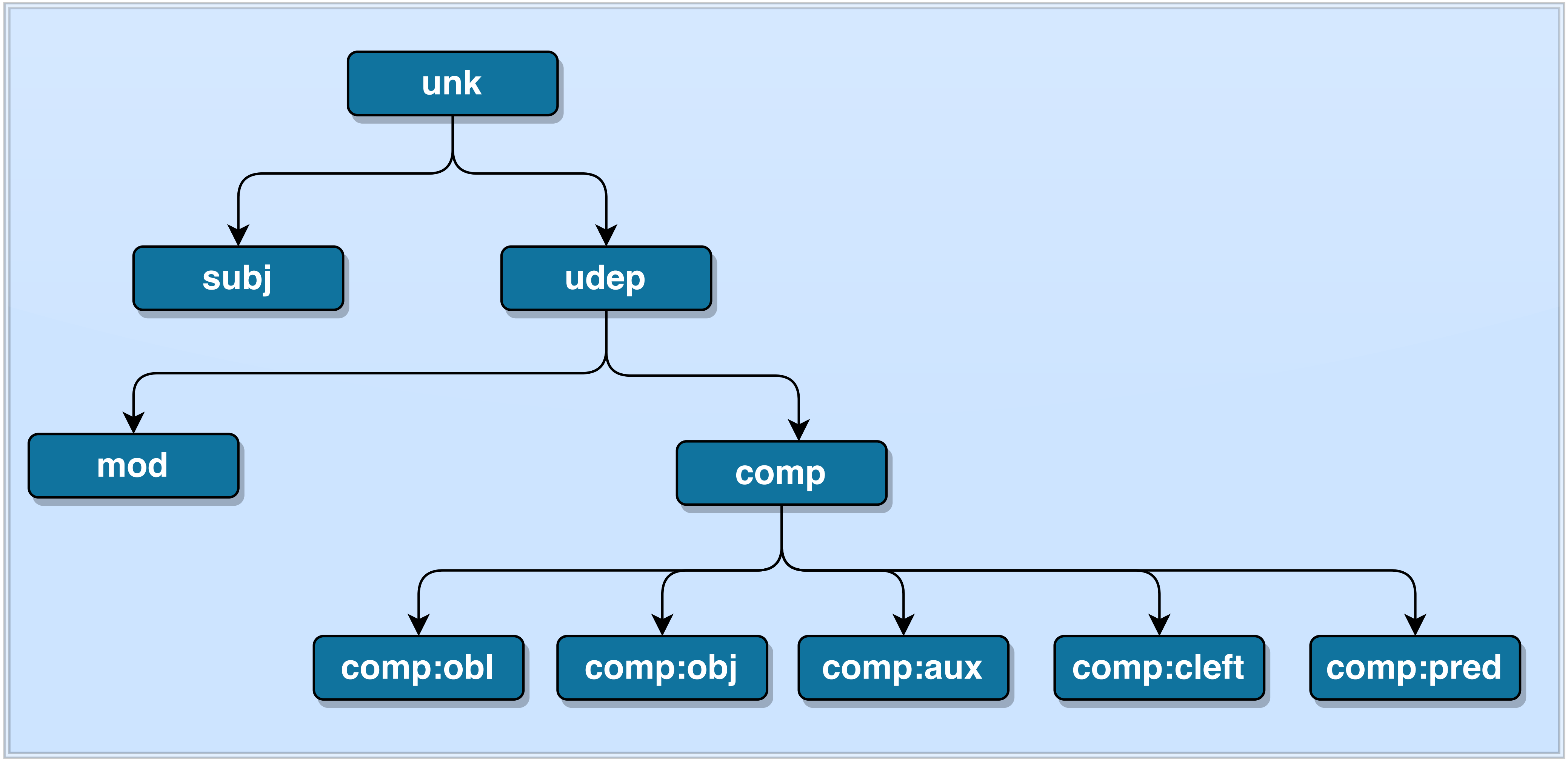
comp:obj

I imagine to dance

pronom VERB PART VERB

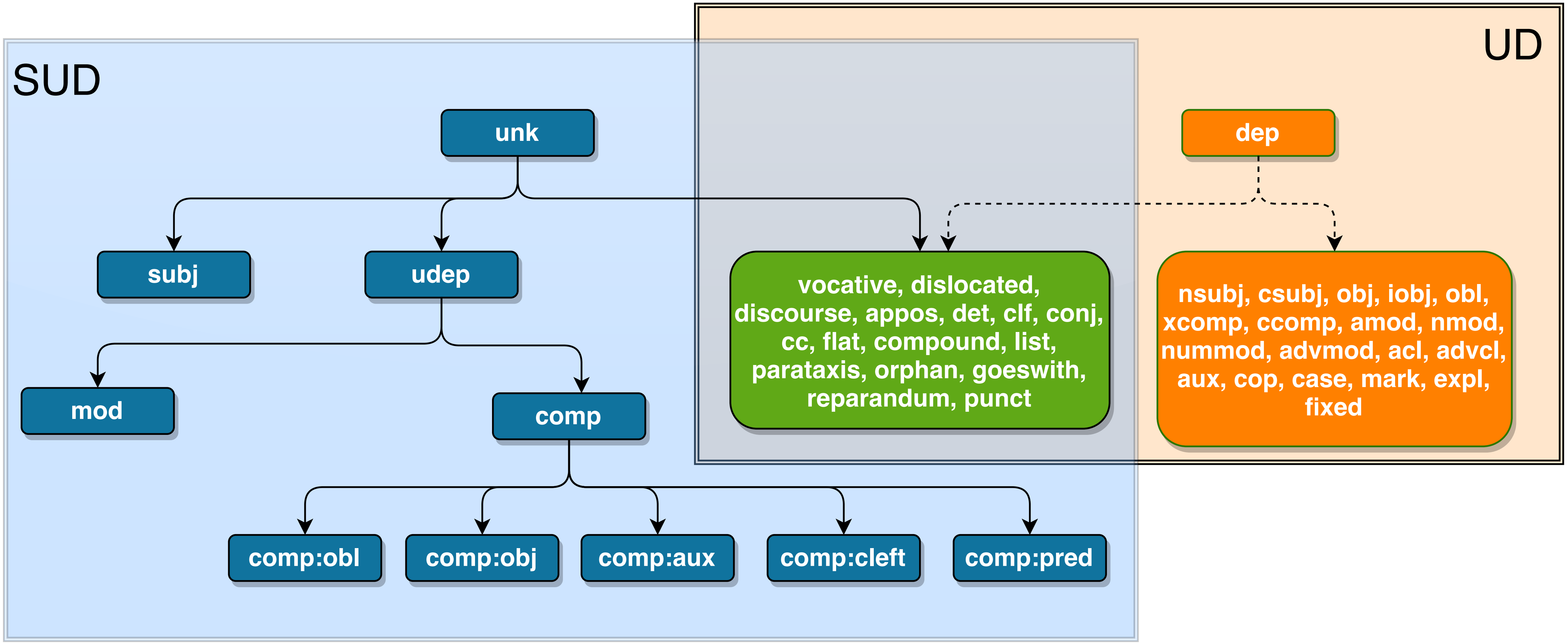
SUD relations

SUD relations are organised in a taxonomic hierarchy

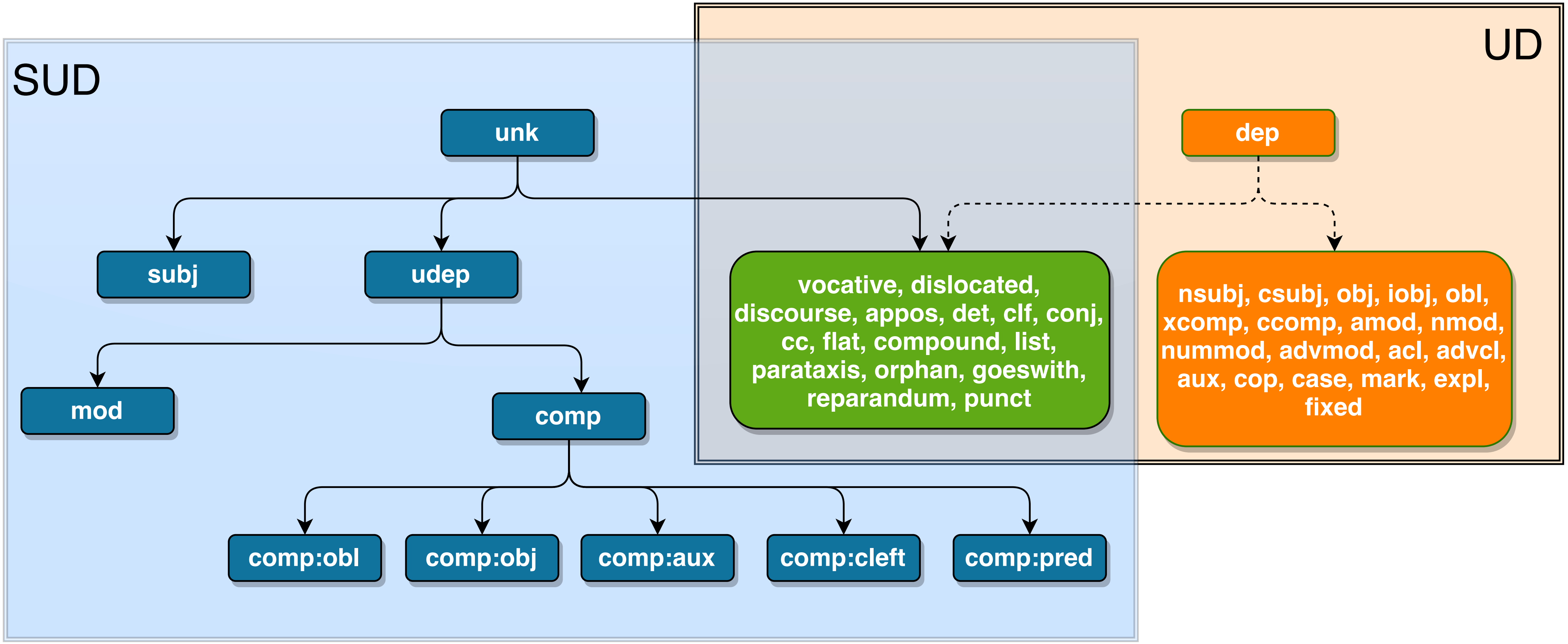


SUD relations

A subset of UD relations are used in SUD


SUD
UD
unk
subj
udep
mod
comp
comp:obl
comp:obj
comp:aux
comp:cleft
comp:pred
dep
vocative, dislocated, discourse, appos, det, clf, conj, cc, flat, compound, list, parataxis, orphan, goeswith, reparandum, punct
nsubj, csubj, obj, iobj, obl, xcomp, ccomp, amod, nmod, nummod, advmod, acl, advcl, aux, cop, case, mark, expl, fixed

SUD relations



SUD

UD

unk

subj

udep

mod

comp

comp:obl

comp:obj

comp:aux

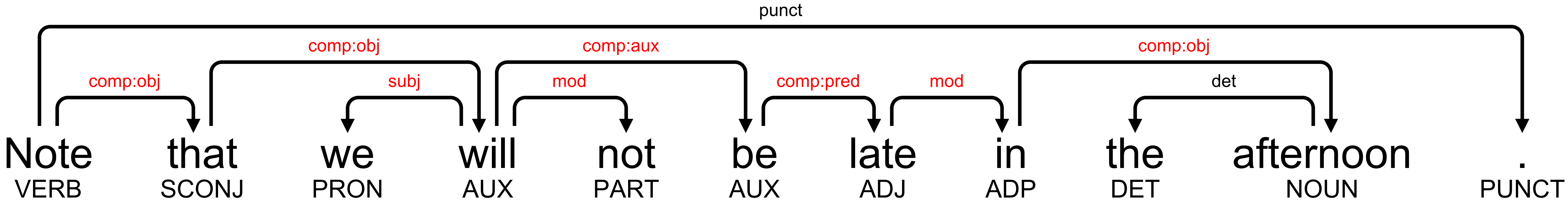
comp:cleft

comp:pred

dep

vocative, dislocated, discourse, appos, det, clf, conj, cc, flat, compound, list, parataxis, orphan, goeswith, reparandum, punct

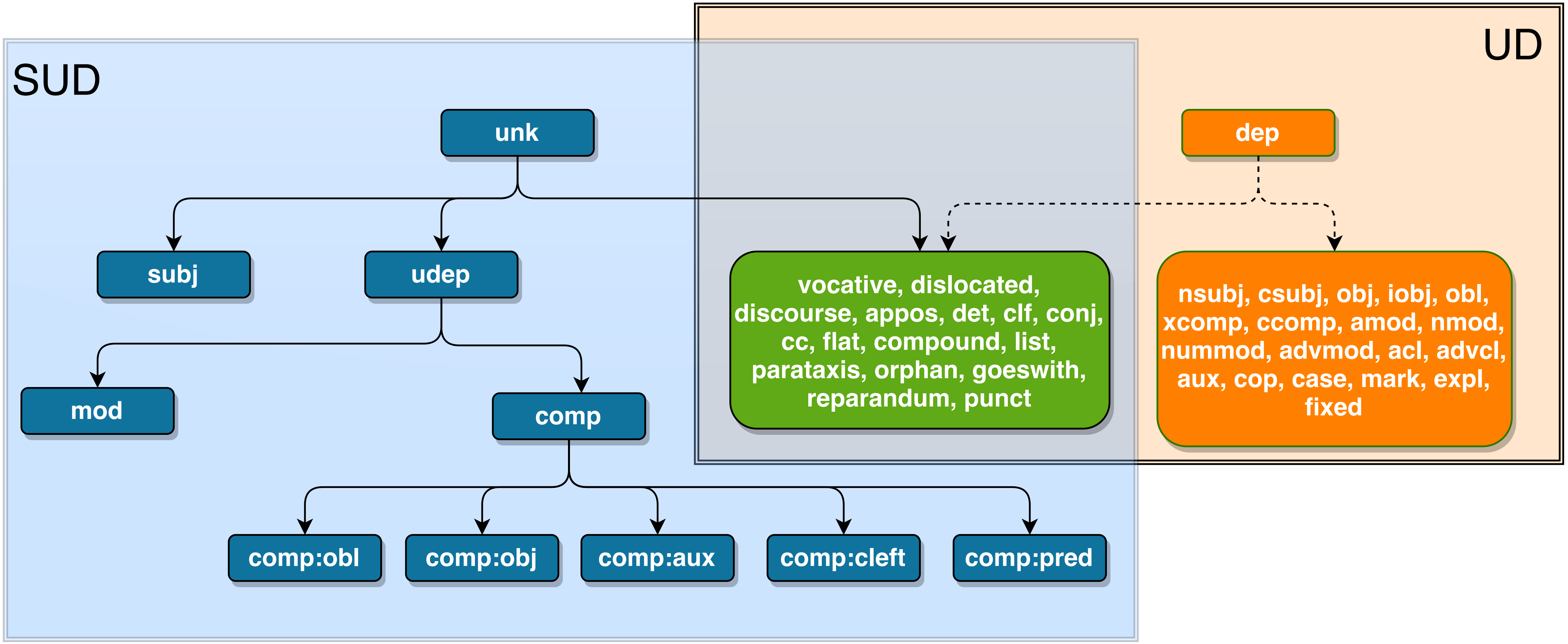
nsubj, csubj, obj, iobj, obl, xcomp, ccomp, amod, nmod, nummod, advmod, acl, advcl, aux, cop, case, mark, expl, fixed



Note that we will not be late in the afternoon .

VERB SCONJ PRON AUX PART AUX ADJ ADP DET NOUN PUNCT

SUD relations


SUD
UD
unk
subj
udep
mod
comp
comp:obl
comp:obj
comp:aux
comp:cleft
comp:pred
dep
vocative, dislocated, discourse, appos, det, clf, conj, cc, flat, compound, list, parataxis, orphan, goeswith, reparandum, punct
nsubj, csubj, obj, iobj, obl, xcomp, ccomp, amod, nmod, nummod, advmod, acl, advcl, aux, cop, case, mark, expl, fixed

All the rest is not surface syntax!

All the rest concerns syntax-semantics interface
⇒ deep syntax

Deep syntactic features

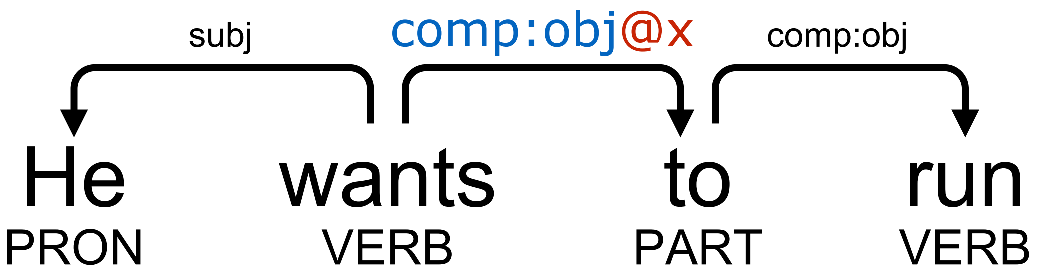
NEW

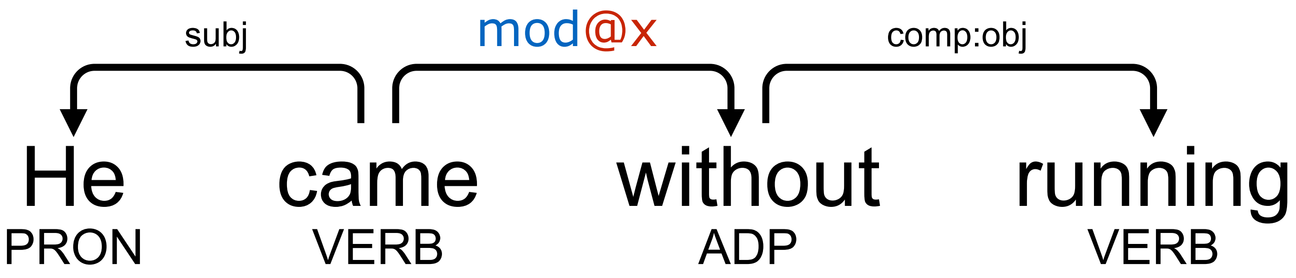
UD encodes both **surface-syntactic** relations and **deep-syntactic** features:

- `xcomp`, `aux:pass`, `aux:cause`, `obj:lvc`

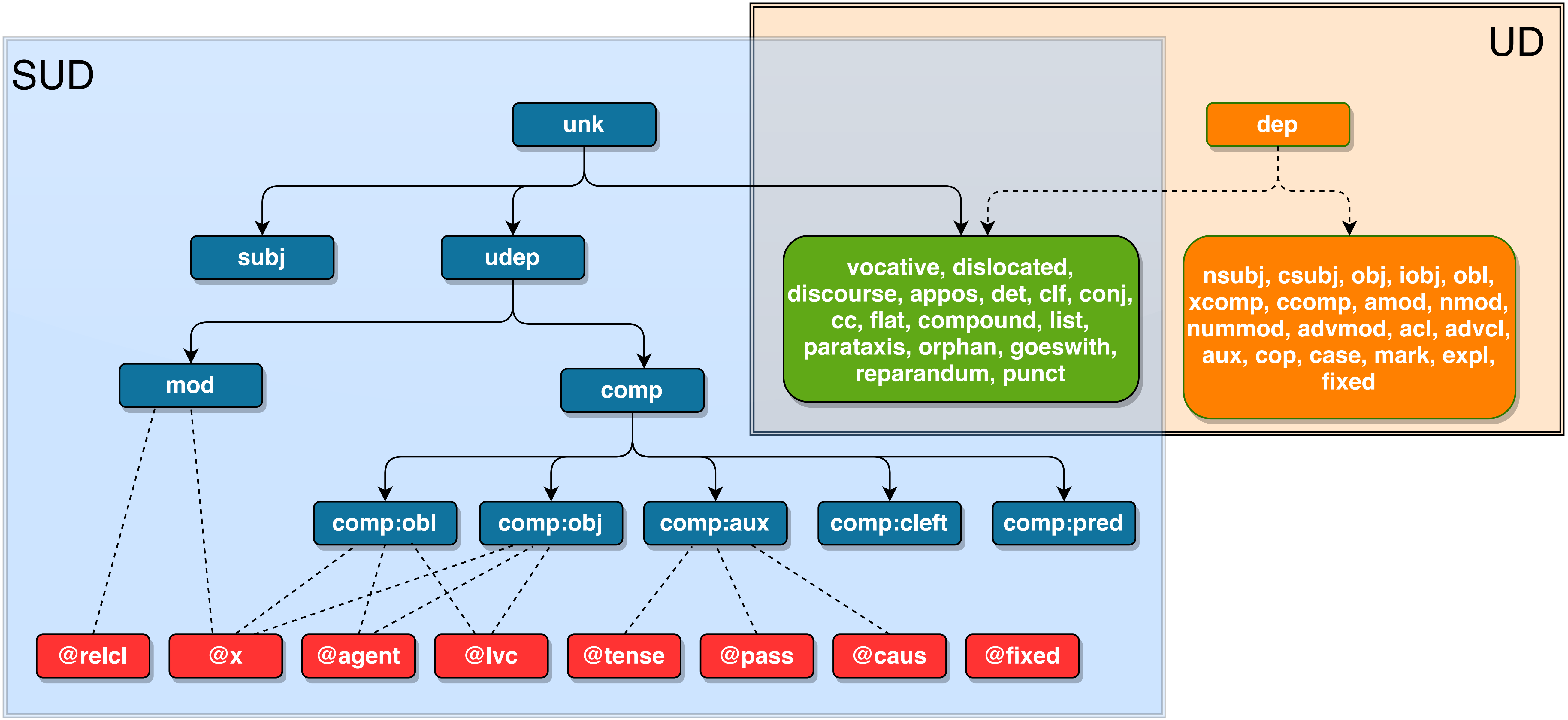
SUD proposes a strict separation between **surface-syntactic** relations and **deep-syntactic** features (written with @):

- `comp:obj@x`, `comp:aux@pass`, `comp:aux@cause`, `comp:obj@lvc`





SUD relations


SUD
UD
unk
dep
subj
udep
vocative, dislocated, discourse, appos, det, clf, conj, cc, flat, compound, list, parataxis, orphan, goeswith, reparandum, punct
nsubj, csubj, obj, iobj, obl, xcomp, ccomp, amod, nmod, nummod, advmod, acl, advcl, aux, cop, case, mark, expl, fixed
mod
comp
comp:obl
comp:obj
comp:aux
comp:cleft
comp:pred
@relcl
@x
@agent
@lvc
@tense
@pass
@caus
@fixed

Encoding **MWE**

NEW

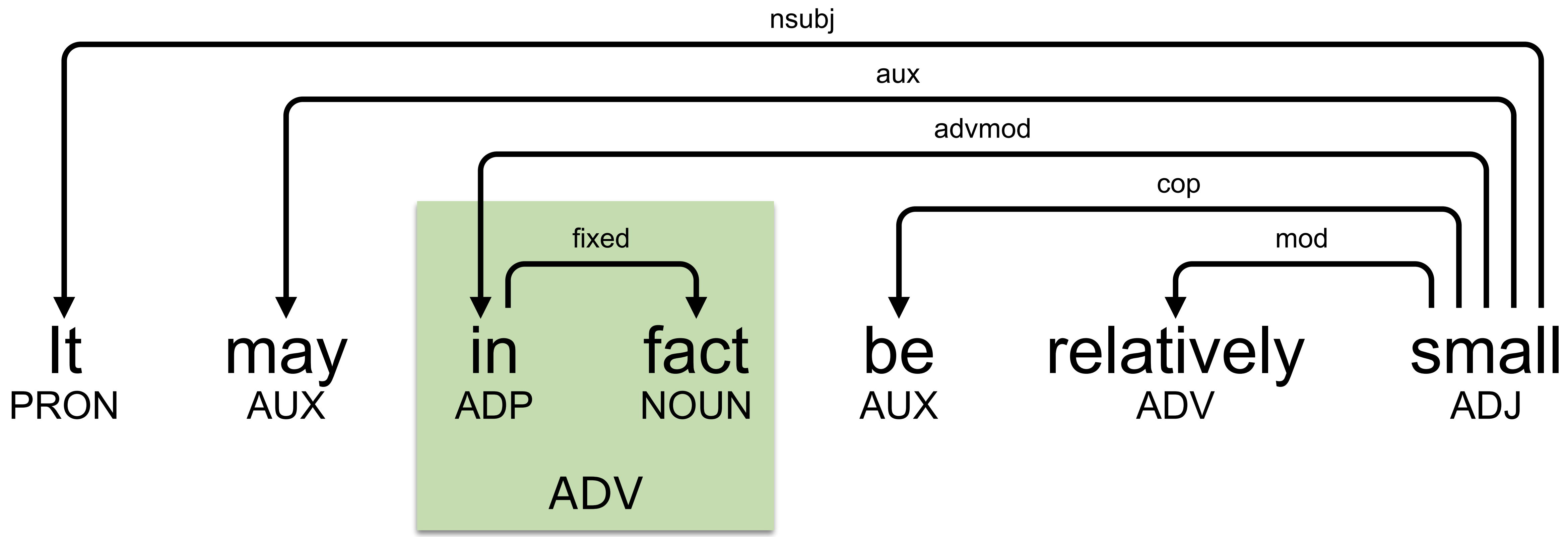
In UD, the `fixed` relation is used to annotate some MWEs:

“It is used for certain fixed grammaticized expressions that behave like function words or short adverbials.”

This relation encodes two different aspects:

- there is no clear internal syntactic structure
- whole expression may have a POS which is not predictable from the POS of the internal tokens (validation rules)

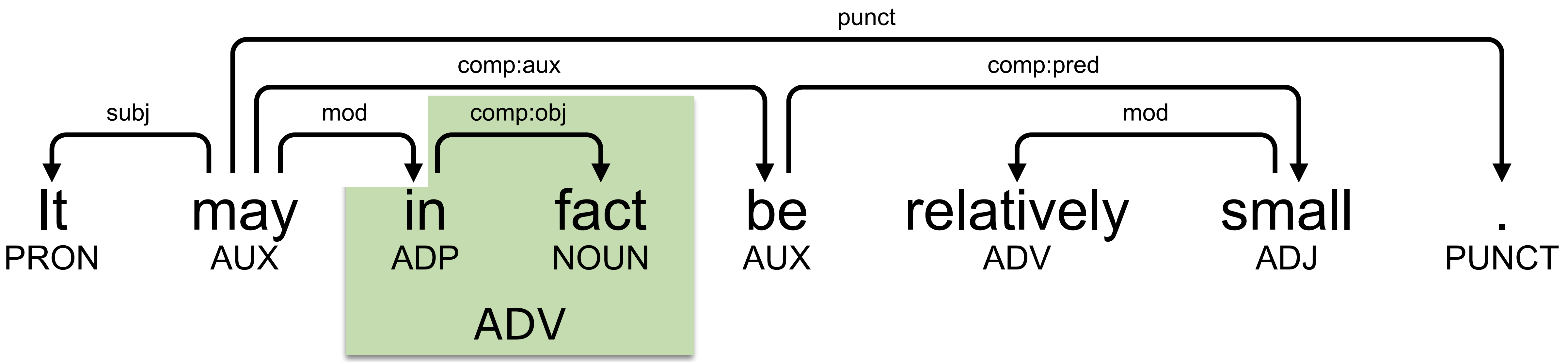
Encoding **MWE**

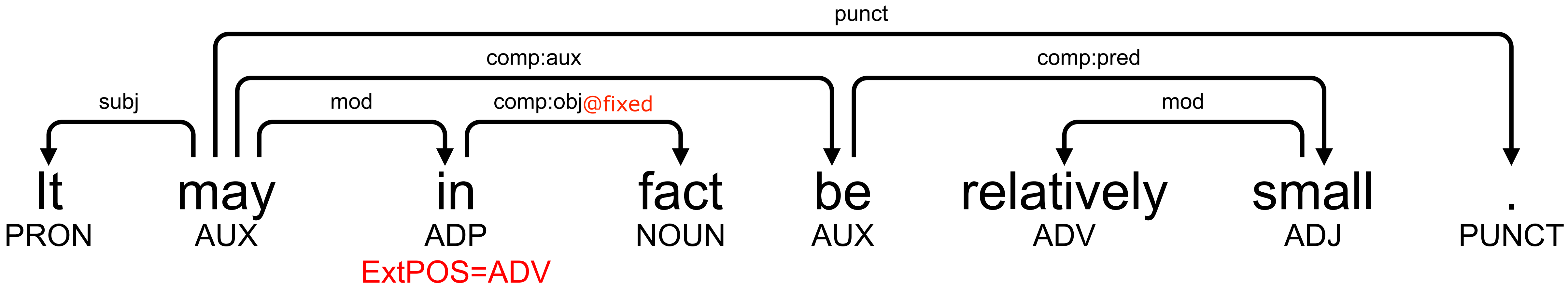


These two aspects are not necessarily linked:

- *in fact* can be analysed as ADP+NOUN with a case relation
- *in fact* can be used as a short adverbial

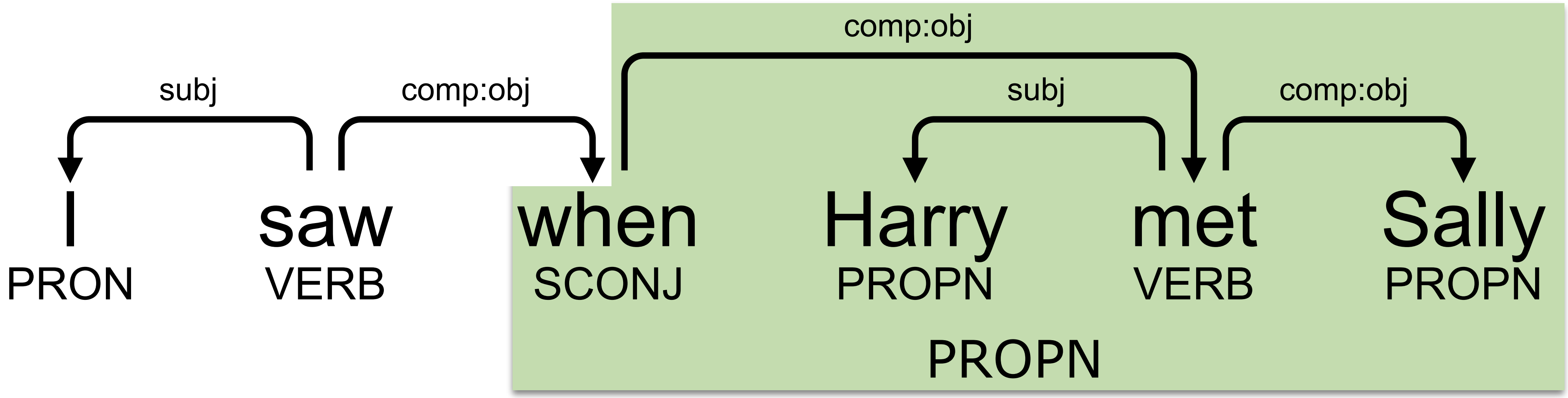
Encoding **MWE** in **SUD**

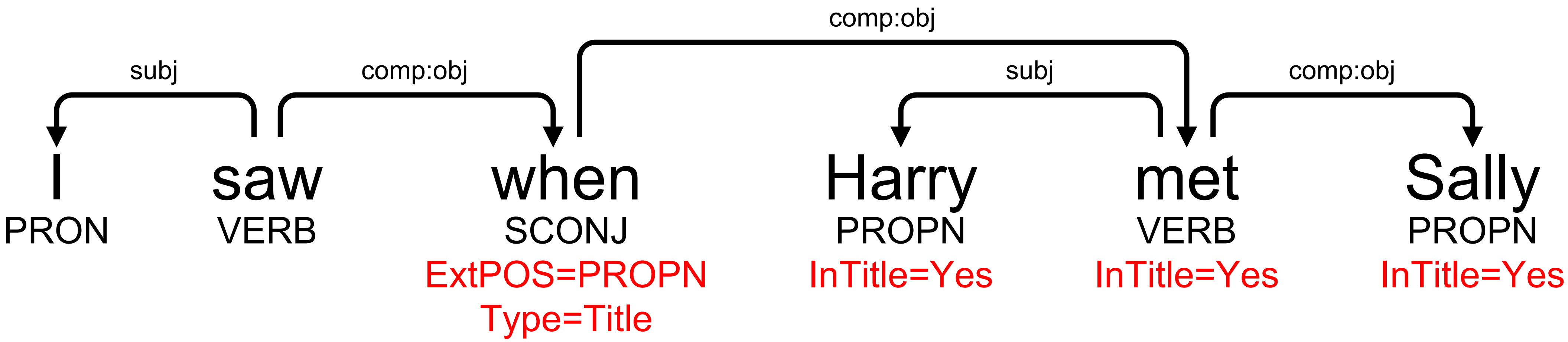




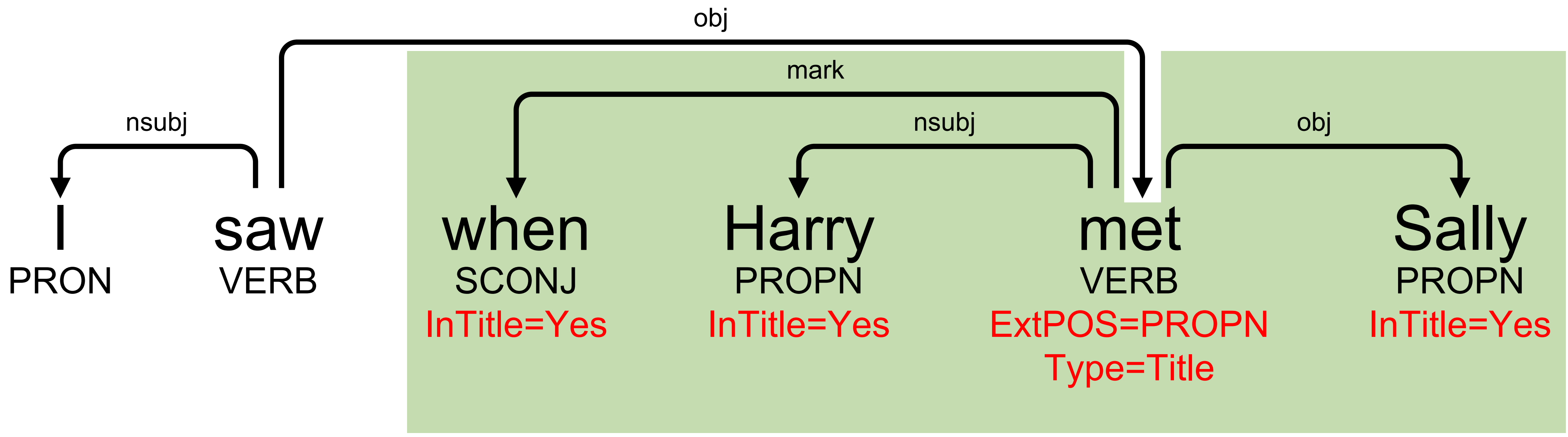
consistent with PARSEME project annotations

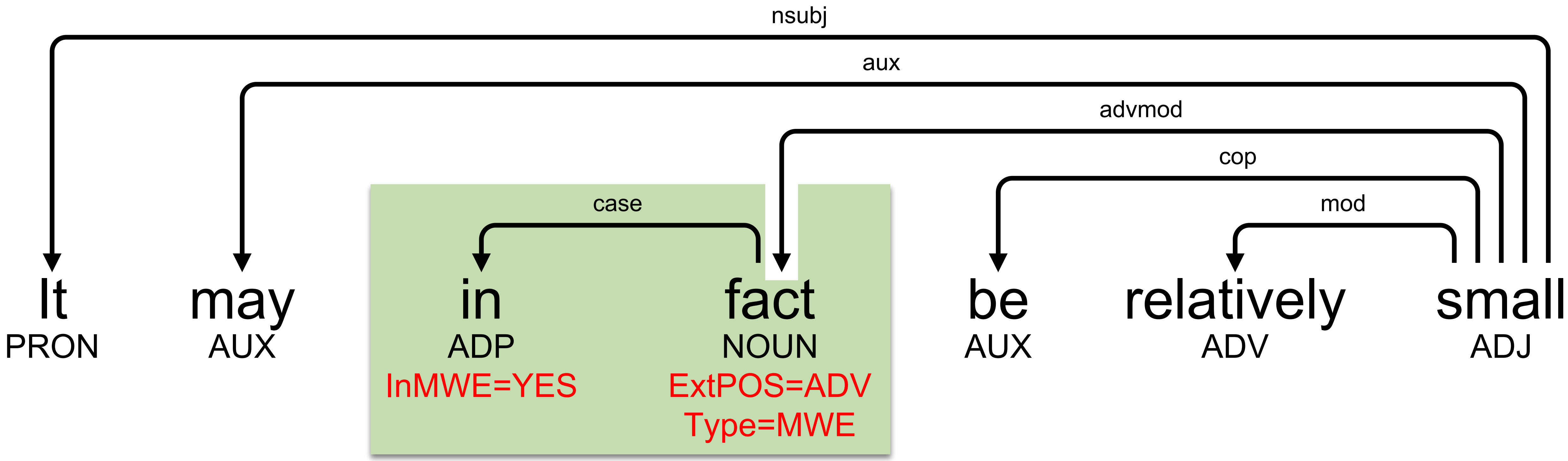
ExtPOS: Encoding titles in **SUD**





ExtPOS in UD?


obj
mark
nsubj
nsubj
obj
I
PRON
saw
VERB
when
SCONJ
InTitle=Yes
Harry
PROPN
InTitle=Yes
met
VERB
ExtPOS=PROPN
Type=Title
Sally
PROPN
InTitle=Yes


nsubj
aux
advmod
cop
case
mod
It
PRON
may
AUX
in
ADP
InMWE=YES
fact
NOUN
ExtPOS=ADV
Type=MWE
be
AUX
relatively
ADV
small
ADJ

Conclusion

The SUD principles have been refined:

- clear distinction between surface-syntax properties (based on distributional criteria), and deep-syntactic properties (concerning the syntax-semantics interface)
- encoding of the POS of MWEs and other irregularities

And also:

- We provide automatic transformation tools UD ⇒ SUD and SUD ⇒ UD
- We hope to bring cross-fertilisation to both projects, on ideas, resources, tools, …

surfacesyntacticud.github.io