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

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 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
Heads:

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

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

- String analysis of coordination
Heads:

- Distributional criteria (Bloomfield, Hudson, Mel’čuk) favours functional heads \(\Rightarrow\) ADP, AUX, SCONJ are heads

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

![Diagram](https://via.placeholder.com/150)

- **SA**
  - `comp:obj`
  - `I pronom imagine a dance`
  - `I pronom imagine a dance`
  - `I pronom imagine a dance`
  - `I pronom imagine a dance`
  - `I pronom imagine a dance`

- **SB**
  - `comp:obj`
  - `I pronom imagine a dance`
  - `I pronom imagine a dance`
  - `I pronom imagine a dance`
  - `I pronom imagine a dance`
  - `I pronom imagine a dance`
SUD relations are organised in a taxonomic hierarchy
A subset of UD relations are used in SUD

SUD relations

SUD

- unk
- subj
- udep
- mod
- comp
- comp:obl
- comp:obj
- comp:aux
- comp:cleft
- comp:pred

UD

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

Note that we will not be late in the afternoon.

SUD relations

- **subj**
- **udep**
- **mod**
- **comp**
  - **comp:obl**
  - **comp:obj**
  - **comp:aux**
  - **comp:cleft**
  - **comp:pred**

UD

- **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**

- **comp:obj**
- **subj**
- **mod**
- **comp:aux**
- **comp:pred**
- **mod**
- **det**
- **comp:obj**

Note that we will not be late in the afternoon.
SUD relations

All the rest is not surface syntax!

All the rest concerns syntax-semantics interface

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

SUD relations

**SUD**

- unk
- subj
- udep
- mod
- comp
  - comp:obl
  - comp:obj
  - comp:aux
  - comp:clef
  - comp:pred
  - @relcl
  - @x
  - @agent
  - @lvc
  - @tense
  - @pass
  - @caus
  - @fixed

**UD**

- 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
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)
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

It may in fact be relatively small.

ExtPOS=ADV

consistent with PARSEME project annotations
**ExtPOS: Encoding titles in SUD**

```
I PRON saw VERB when SCONJ Harry PROPN met VERB Sally PROPN

subj comp:obj subj comp:obj subj comp:obj

I PRON saw VERB when SCONJ Harry PROPN
ExtPOS=PROPn
InTitle=Yes
Type=Title

met VERB Sally PROPN
InTitle=Yes

Harry PROPN
InTitle=Yes

Sally PROPN
InTitle=Yes
```
ExtPOS in UD?

I saw when Harry met Sally

It may be relatively small

ExtPOS in UD

ExtPOS=PROPN
Type=Title

ExtPOS=ADV
Type=MWE

ExtPOS=ADP
InMWE=YES

ExtPOS=ADV
Type=MWE
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 $\Rightarrow$ SUD and SUD $\Rightarrow$ UD
- We hope to bring cross-fertilisation to both projects, on ideas, resources, tools, …

surfacesyntacticud.github.io