Dependency syntax
SUD vs UD annotation schemes

Sylvain Kahane
(Modyco, Paris Nanterre & CNRS / IUF)

Chișinău, July 9, 2024

Definition of the syntactic structure

Content

- Principles of dependency syntax
  - connections
  - heads/dependencies
  - words and sentences
  - categories and relations
- (Surface-Syntactic) UD
  - annotation scheme *vs* theoretical principles
  - SUD tag set
  - conversion to UD

Word of the day

- *criterion, criterion, criterion, criterion*
- every decision must be based on criteria (and tests)
  - we don't all need to have the same criteria, but we do need to know each other's criteria
- criteria => annotation guidelines

Mel’čuk’s (1988) definition

- Criteria A: syntactic phrases (=> connection)
- Criteria B: head of a phrase (=> dependency)
- Criteria C: syntactic relations

  – Igor Mel’čuk (1988) *Dependency syntax: Theory and practice*, SUNY Press.
  – Richard Hudson (1984) *Word Grammar*, OUP.
  – Garde (1977) Ordre linéaire et dépendance syntaxique
  – Zwicky (1985) Heads

Criteria

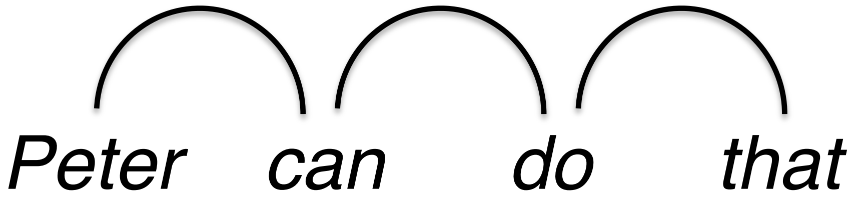
(Contrary to Mel’čuk we don’t presuppose we have words and sentences’ boundaries, as well as POS)

- Criteria A: syntactic phrases (=> connection)
- Criteria B: head of a phrase (=> dependency)
- Criteria C: syntactic relations
- Criteria for minimal and maximal units
  - lexeme, grammeme, word, sentence
- Criteria for POS (part of speech)
  - Kahane & Gerdes, 2022, *Syntaxe théorique et formelle*, Language Science Press.

Connections
(Criteria A)

Syntactic units and connections

- Syntactic units: any subpart of a sentence that has some kind of autonomy
  - especially subparts that can stand alone
  - Example: *Peter can do that*
    - *Peter can*
    - **Peter do*
    - *can do, do that*
- Things that go together must be connected



Criteria for syntactic units

- Syntactic units: any subpart of a sentence that has some kind of autonomy
  - especially subparts that can stand alone
  - subparts that **combine freely** with their context
  - Example:

Peter	*can do*	*that*
Mary	*does*	*such things*
She	*must do*	*everything*
My son	*knows*	*an interesting paper*
The girl	*is reading*	*it*
…	…	…

free combination

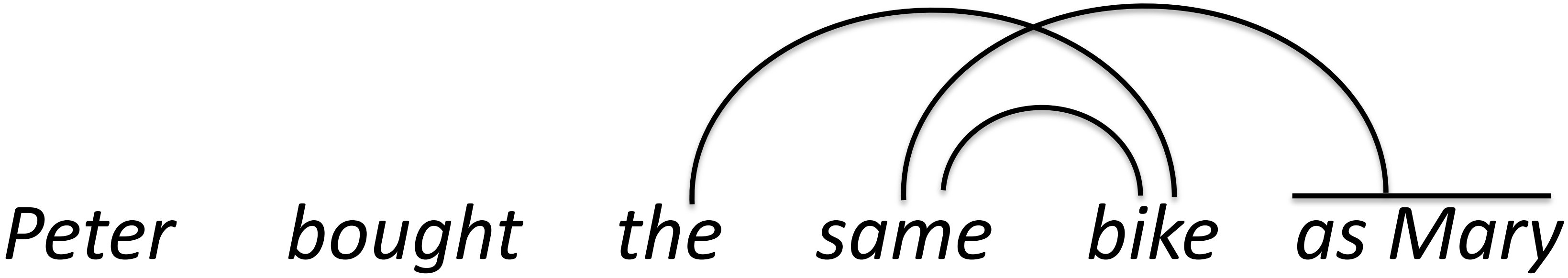
Connections

- If U, A, and B are syntactic units and U = AB, there is a **connection** between A and B
- Remarks
  - a connection has several instantiations
    - *the little boy — can do that*
      boy — can
  - the notion of connection does not presuppose a particular type of units: words, constituents, …

    even lexemes and grammemes (inflection)
    - *cat-Ø* vs *cat-s, stop-Ø* vs *stop-s* vs *stopp-ed* vs *stopp-ing*

Exercise 1

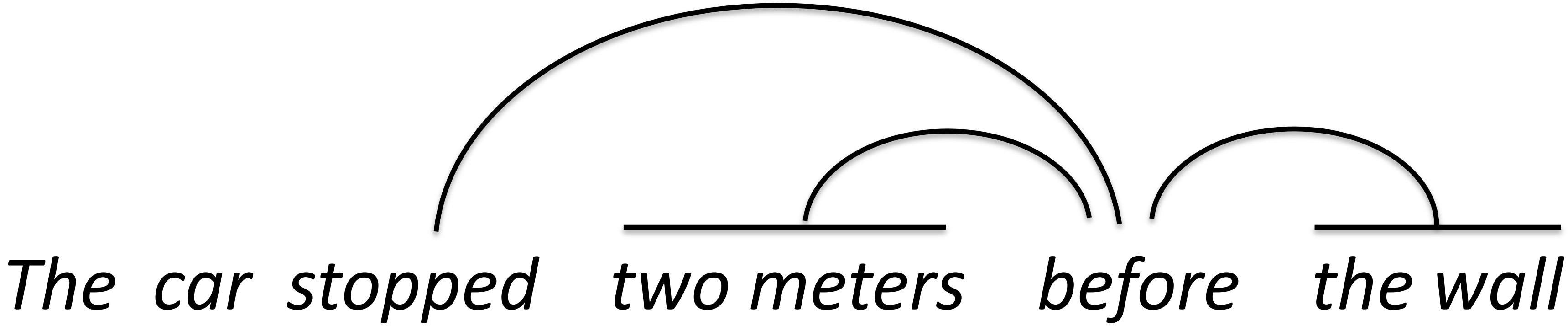
- What are the connections in:
  - *Peter bought the same bike as Mary*
    - $^{??}$the bike as Mary
    - same as Mary
    - the bike
    - same bike



Peter bought the same bike as Mary

Exercise 2

- What are the connections in:
  - *The car stopped two meters before the wall*
    - the car stopped
    - stopped before
    - two meters before
    - before the wall

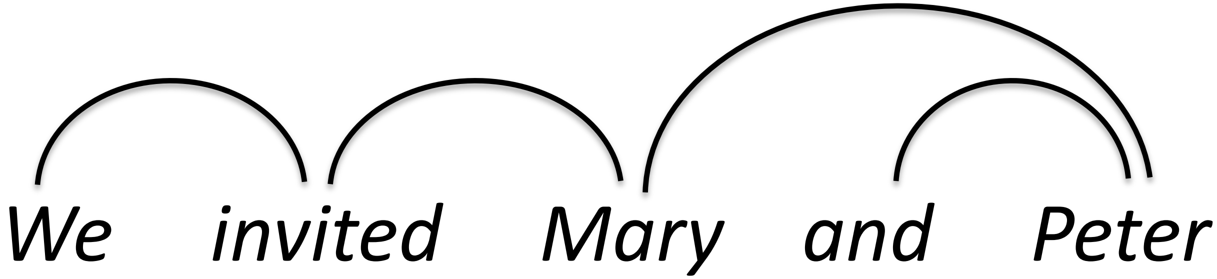


The car stopped two meters before the wall

Exercise 3

- What are the connections in:
  - *We invited Mary and Peter*
    - *Mary and
    - and Peter
    - Mary, Peter …

=> asymmetrical analysis of coordination



We invited Mary and Peter

Connections

- Just by looking at what goes together
  => graph structure
  (Gerdes & Kahane 2011)
- We don’t need the notion of word or sentence to define the notion of connection.
- Some problems remain:
  - determiner-noun: *The dog slept*
    - *the slept
    - ?*dog slept

Heads and dependencies (Criteria B)

Syntactic head

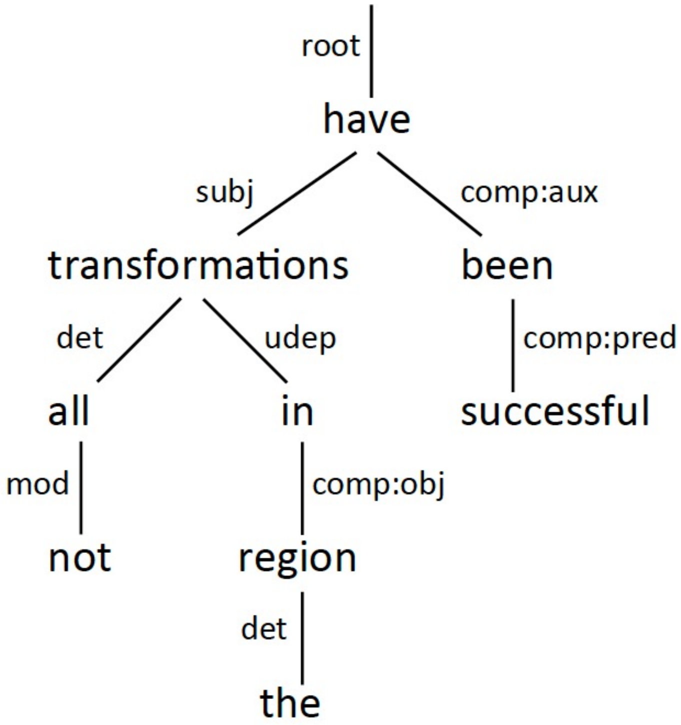
- Most connections are asymmetric:
  - governor/head
  - dependent
- The **head** of a unit is the word that **controls** its distribution, that is, the position that the unit can occupy
  - Example: *We think that …*

here we need a finite verb

the head of a clause is the finite form

Dependencies

- The head of a unit is the word that controls its distribution, that is, the position that the unit can occupy
  - Example: *We think that …*

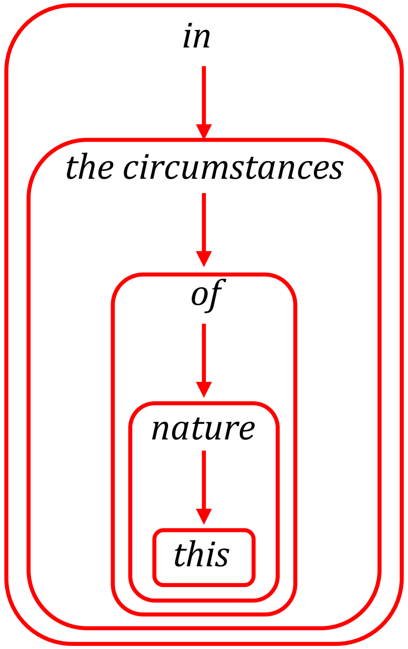


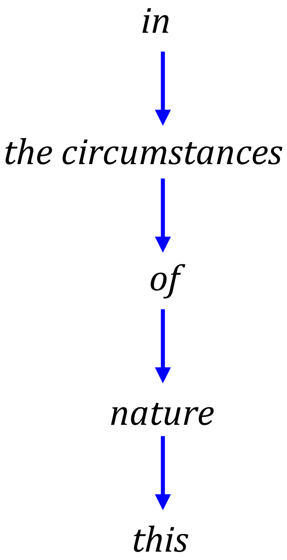
here we need a finite verb

the head of a clause is the finite form

Beauzée 1765

"For instance in the sentence *with the care requested in the circumstances of this nature*; the word *nature* is the grammatical complement of the preposition *of*; *this nature* is its logical complement; the preposition *of* is the initial complement of the appellative noun *the circumstances*; and *of this nature* is its total complement; *the circumstances* is the grammatical complement of the preposition *in*; and *the circumstances of this nature* is its logical complement."





- grammatical complement = initial complement = dependent
- logical complement = total complement = constituent

Bloomfield 1933

- Immediate Constituent Analysis: Leonard Bloomfield (1933), *Language*
  - immediate constituents:
    - *poor John* and *ran away* are the **immediate constituents** of *poor John ran away*
  - endocentric constructions:
    - *John* is the **head** of *poor John* because *John* and *poor John* have the same “function” (= distribution)
- the development of ICA is inseparable from the development of the notion of head, until the break of Chomsky (1957)

Criteria for heads

- Why is the adposition the head of an adpositional phrase?
- Example: *Peter talk* ***to Mary***
  - First criterion: the units *to Mary* and *Mary* have very different distributions
    - ***Mary*** *slept* vs *****To Mary*** *slept*
    - *I like* ***Mary*** vs **I like* ***to Mary***
    - *I talk* ***to Mary*** vs **I talk* ***Mary***
  - Second criterion: the preposition controls the possible positions:
    - *Peter depends* ***on****/****to*** *Mary*
    - *Peter talk* ***to****/****on*** *Mary*

Exercise 4

- What is the head in ***and Peter*** (*I invited Mary and Peter*)?
- Criteria for choosing *and*
  - *Peter* and *and Peter* do not have the same distribution
    - ?**I invited and Peter*
- Criteria for choosing *Peter*
  - *and Peter, and blue, and went* do not have the same distribution
    - **Mary and blue, *Mary and went, *red and Peter …*

Syntactic relation (Criteria C)

Relations

- Two units that occupy the same position (mutual exclusion) have the same syntactic function
  - *I understand **your problem***
  - *I understand **that you have a problem***
  - *I hope **that you may come***
  - *I hope **to see you***

=> object complement

Relations

- Two positions that are occupied by the same paradigm of elements and have the same properties bear the same function
  - I talk **to Mary**
  - I gave **a book** **to Mary**
  - I read **a book**
- Exercise:
  - Have the two positions of *to Mary* exactly the same properties?

Exercise 5

- Have the two positions the same properties?
  - I talk **to Mary**
  - I go **to Chisinau**
- Answer: No, but both *obl* in UD
- Two remarks about relations in UD

Comparative concepts

- Two remarks about relations in UD
  - Relations can be apprehended at different levels of granularity
    - UD (and SUD) have a coarse-grained categorization of constructions
  - UD tags are comparative concepts

    ≠ descriptives categories,
    which are language-specific

Haspelmath 2010, 2018 Comparative concepts and descriptive categories in crosslinguistic studies

Part of speech

Exercise 6

- The lexeme *easy* can appear in three constructions:
  1) I have an **easy** solution to this question.
  2) The solution seems **easy**.
  3) Breathe **easy**!
- Why do we have the same POS in (1) and (2) and not in (3)?
- What criteria do we use?

Solution

- Same distributional class in (1) and (2), not in (3)
  - most adjectives can appear in constructions 1 and 2, not in 3
- POS = homogeneous distributional classes of lexemes
- distribution class = set of units that can appear in the same constructions

Exercise 7

- English has a distributional class whose elements have properties similar to auxiliaries and not verbs. Explain.
  - *I took it*
  - *I can take it*
  - *I am taking it*
- Answer:
  - *can I take it?, I cannot take it, I can easily take it*
- AUX is a distributional class in English, but AUX shouldn’t have the same extension in other languages

POS

- part of speech = lexical category
  = distributional class of lexemes
  (not words!)
- Example:
  - *drives, drove, driven* have very different distributions!
  - *drives* = DRIVE-ind.pres.3sg
    drove = DRIVE.ind.past

Words and sentences?

Syntax

- Traditional definition:
  “Syntax is the study of the organization of words inside the sentence.”
- Problem
  - Can we define the notion of words and sentences before introducing the principles of syntax?
- A better definition of syntax:
  - syntax is the study of free combinations (which includes inflection)

Minimal units?

- minimal syntactic units (syntaxemes)
  - lexeme = minimal lexical unit
  - grammeme = minimal grammatical unit
- word = a particular level of cohesion between lexemes and grammemes
  - Problem: Where is the boundary between words and MWEs?

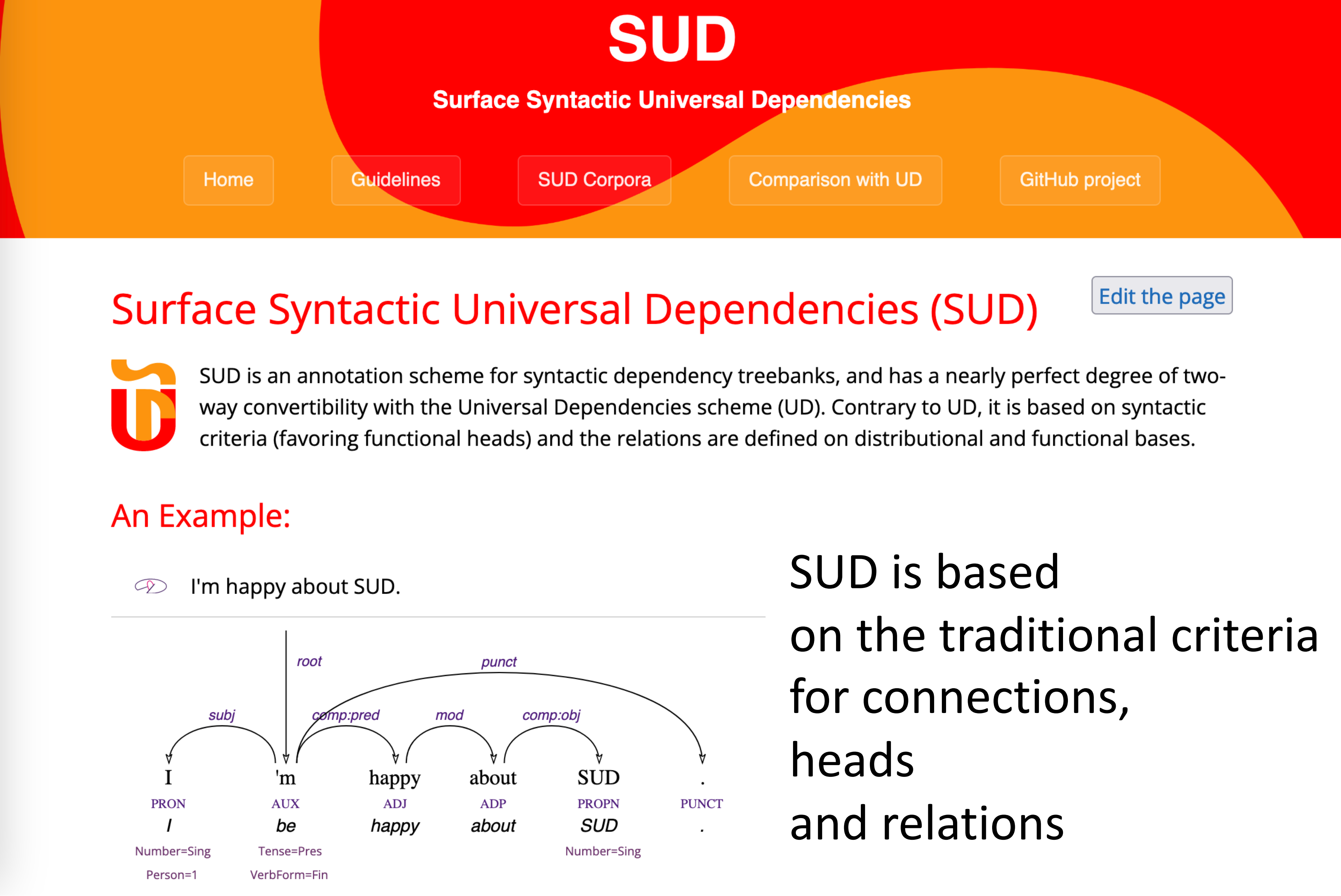
Exercise 7

- Why *a_way* is one word in (1) and two in (2)?
  - *Go away!*
  - *I took a way I hadn't known before.*
- Answer:
  - no commutation on *a* (not a free combination)
    - *around, ahead, aside, across, atop; along, abroad*
  - not the distribution of DET+NOUN
    - *a-* is the only morpheme with such a distribution
  - inseparability: *a long way*
    - not sufficient: *syntax book, *syntax good book*

Maximal units?

- where does the syntax stop? what is a sentence?
  - punctuation?
    - *Mary said: “I will stay here. The place is nice.”*
  - what about corpora without punctuation?
- cf. Unidive WG 1.5 on spoken languages, where we will discuss criteria

Annotation schemes



SUD

Surface Syntactic Universal Dependencies

Home | Guidelines | SUD Corpora | Comparison with UD | GitHub project

Edit the page

Surface Syntactic Universal Dependencies (SUD)

SUD is an annotation scheme for syntactic dependency treebanks, and has a nearly perfect degree of two-way convertibility with the Universal Dependencies scheme (UD). Contrary to UD, it is based on syntactic criteria (favoring functional heads) and the relations are defined on distributional and functional bases.

An Example:

I'm happy about SUD.

SUD is based

on the traditional criteria

for connections,

heads

and relations

https://surfacesyntacticud.github.io/

Annotation scheme issues

- An annotation scheme is a compromise between:
  - theoretical principles
  - practical issues
    - annotator-oriented issues: simplicity, reproducibility
    - user-oriented issues: what is the treebank for?
  - political issues
    - we must be compatible with the standards

(Gerdes & Kahane 2016)

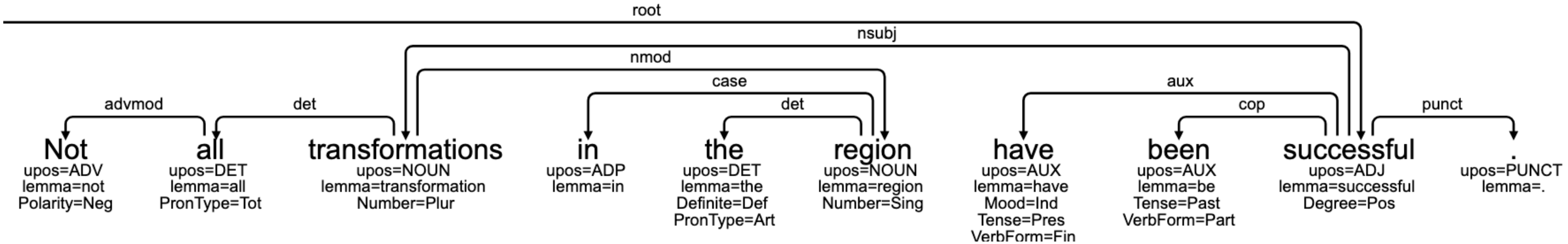
Surface-Syntactic UD

- (Gerdes, Guillaume, Kahane, †Perrier, 2018, 2019, 2021, 2023)
- SUD is based on distributional criteria
- SUD must be converted into UD
  (because UD is the standard)
  - same POS
  - same morphosyntactic features

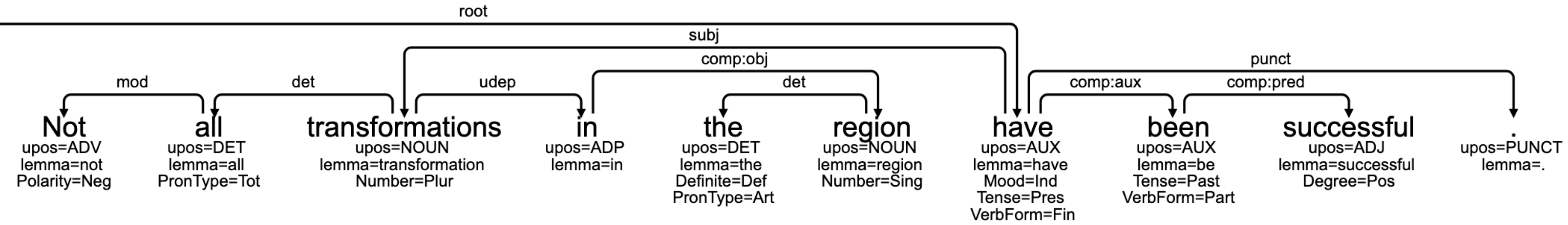
but

  - different structures
  - different relations

UD



SUD

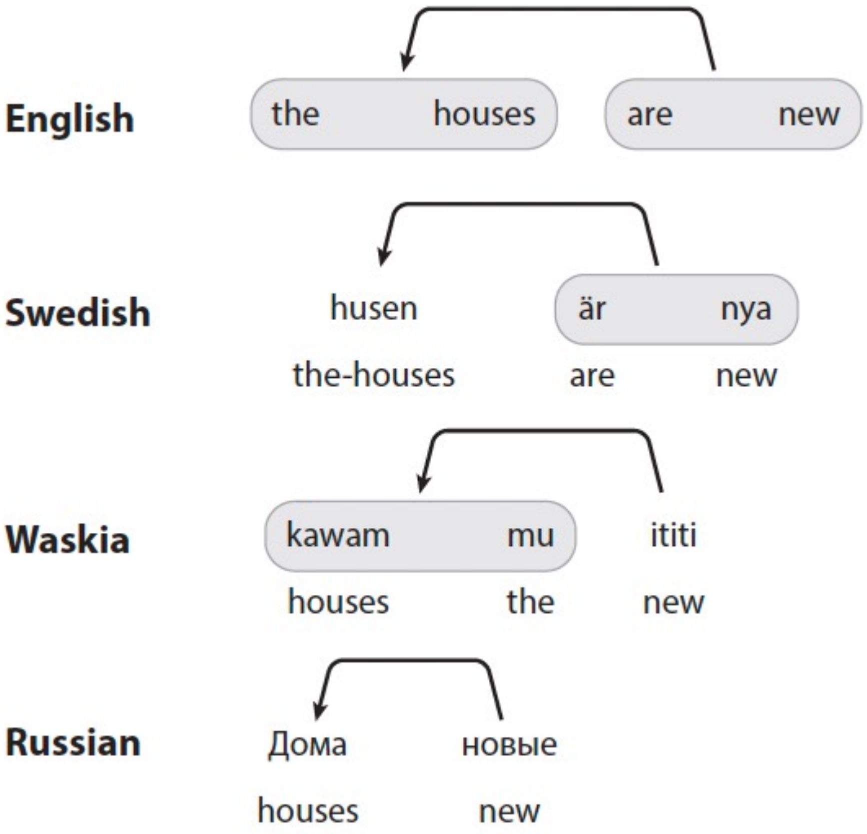


- Automatic conversion UD => SUD
    - – Grew (Graph Rewriting Grammars) (Guillaume 2021)

UD and nuclei

- UD principles:
  - connections between content words
  - function words are leaves of the tree
- nucleus = a content word with functions words
  - Tesnière 1959
    - connections are between nuclei
  - de Marneffe, Nivre 2019, Dependency grammar
    - UD and SUD have the connections between nuclei
    - languages tend to have the same connections between nuclei (but different structures inside nuclei)

- languages tend to have the same connections between nuclei (but different structures inside nuclei)



English

the houses are new

Swedish

husen är nya

the-houses are new

Waskia

kawam mu ititi

houses the new

Russian

Дома новые

houses new

Figure 5

Strategies for expressing nonverbal predication (and definiteness).

- UD = nucleus-level annotation

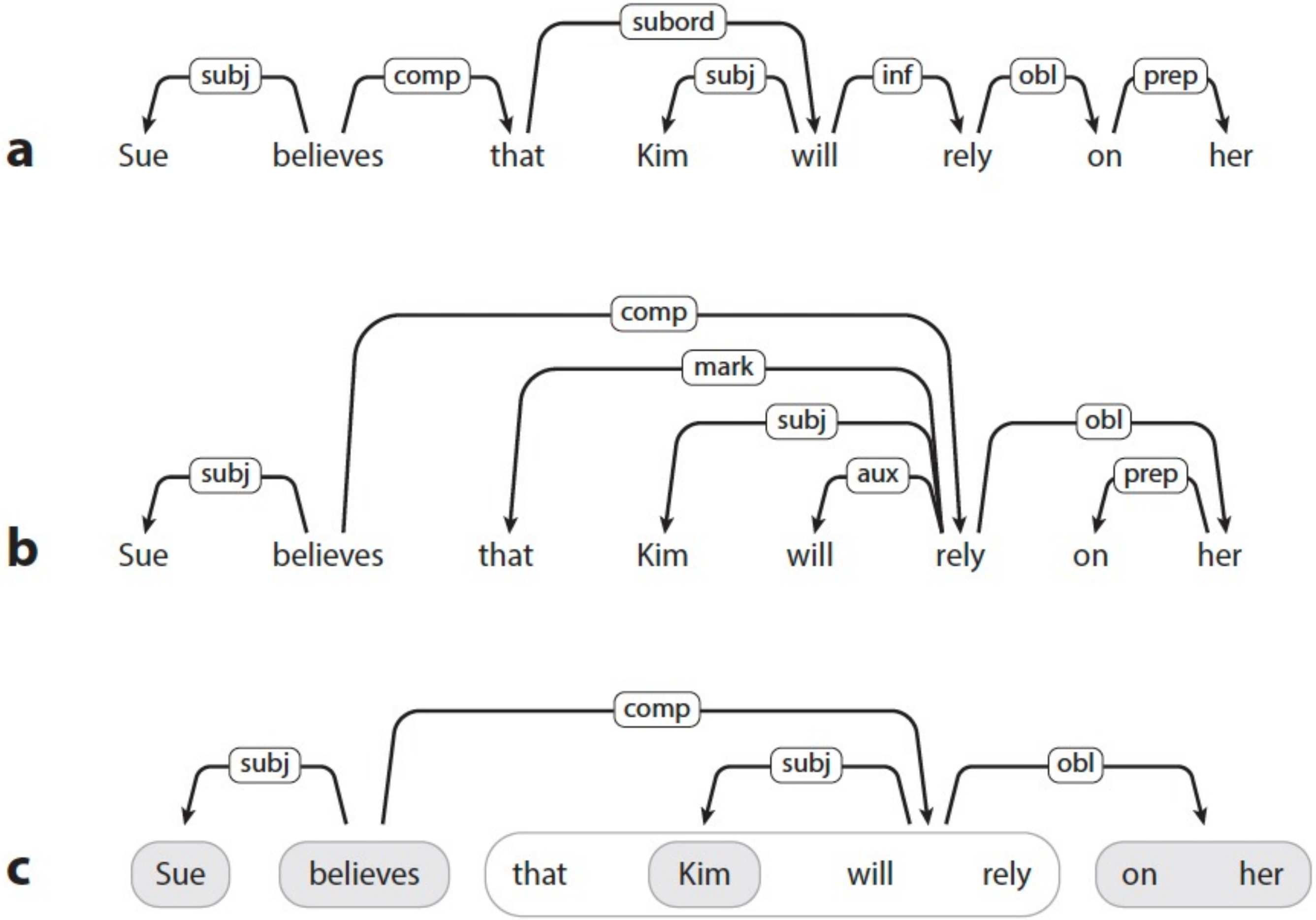


Figure 4

Dependency trees with different heads. (*a*) Function words. (*b*) Content words. (*c*) Nuclei à la Tesnière.

UD vs SUD

- Three problems with UD:

1. First problem: what is a content word?

- Example: *The car stopped two meters* ***before*** *the wall*
  - is *before* a content word?
- No, in UD terms
  - content words: NOUN, VERB, ADJ, ADV
  - function words: ADP, SCONJ, CCONJ, AUX, DET

2. Second problem: UD does not keep the connectedness of syntactic units

- Examples: ***Peter can*** *do that*

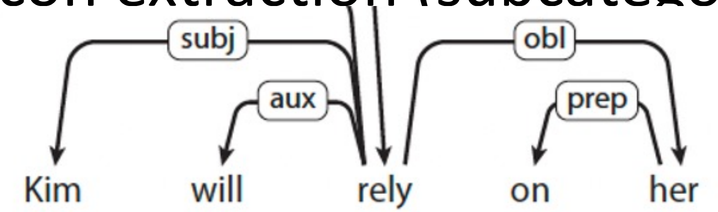
The car stopped ***two meters*** ***before*** ***the wall***

UD vs SUD

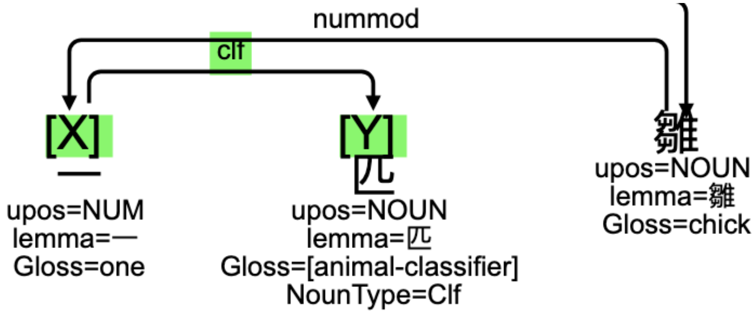
- Third problem: head-marking languages
  - markers of a relation can be on the dependent, alone, or on the head
    - Latin: *Petr***i** *canis* ‘Peter’s dog’ vs *Petrus* ‘Peter’
    - French: *le chien* ***de*** *Pierre ‘*the dog of Peter’
    - Wolof: *xaj***u** *Peer bi* ‘dog-u Peter the’ vs *xaj bi* ‘dog the’
  - English preposition:
    - in V-Prep-N, Preps tend to form a nucleus with V
      - *the problem I am* ***talking about*** (preposition stranding)
      - *Peter* ***is talking about*** *syntax and Mary semantics*
    - prepositions tend to becomes verbal particules and to freeze with V: *go on, take off, figure out …*

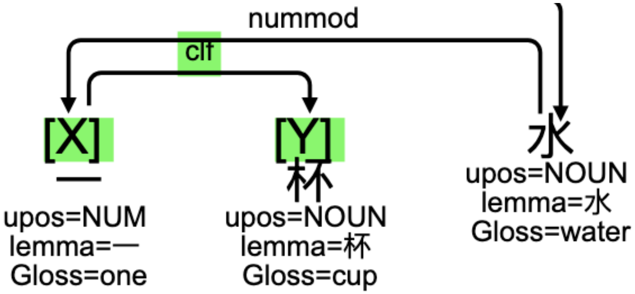
UD vs SUD

- Advantages of UD
  - better parallelism between languages
    - “superficial” differences are flattened
  - better for lexicon extraction (subcategorization frame)



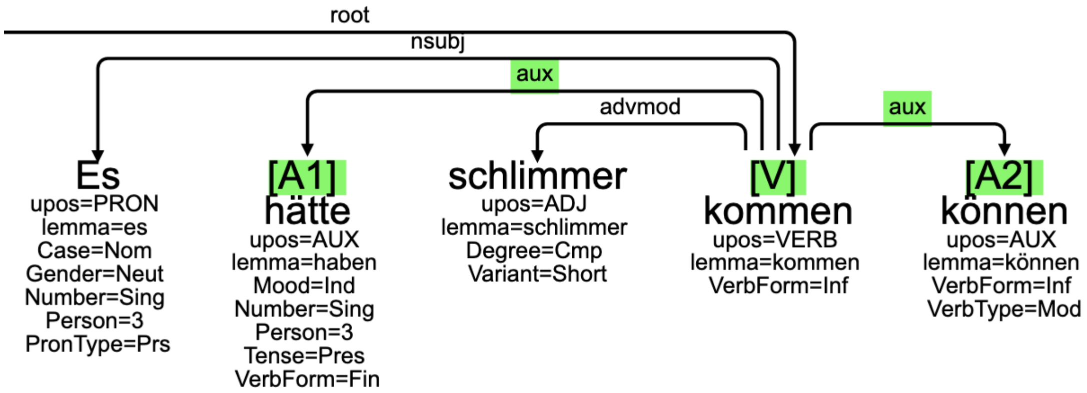
  - simpler annotation (if you know what a content word is)





UD vs SUD

- Advantages of SUD
  - distributional criteria
  - richer

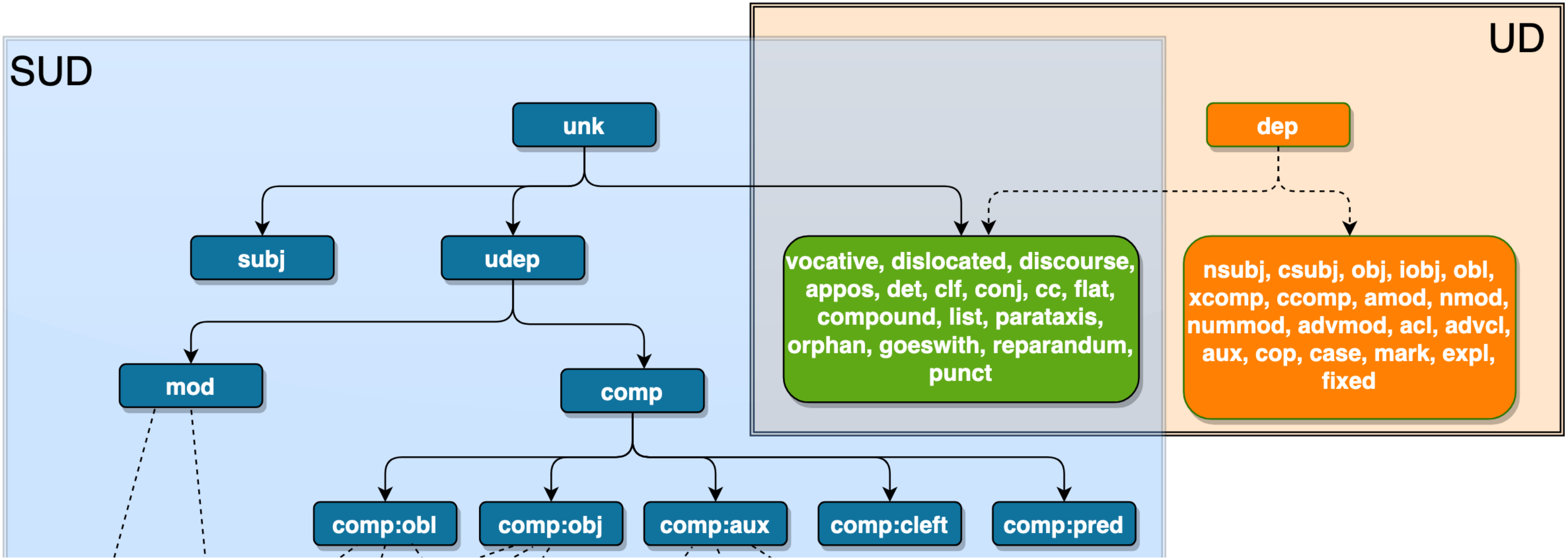


‘It could have been worse’

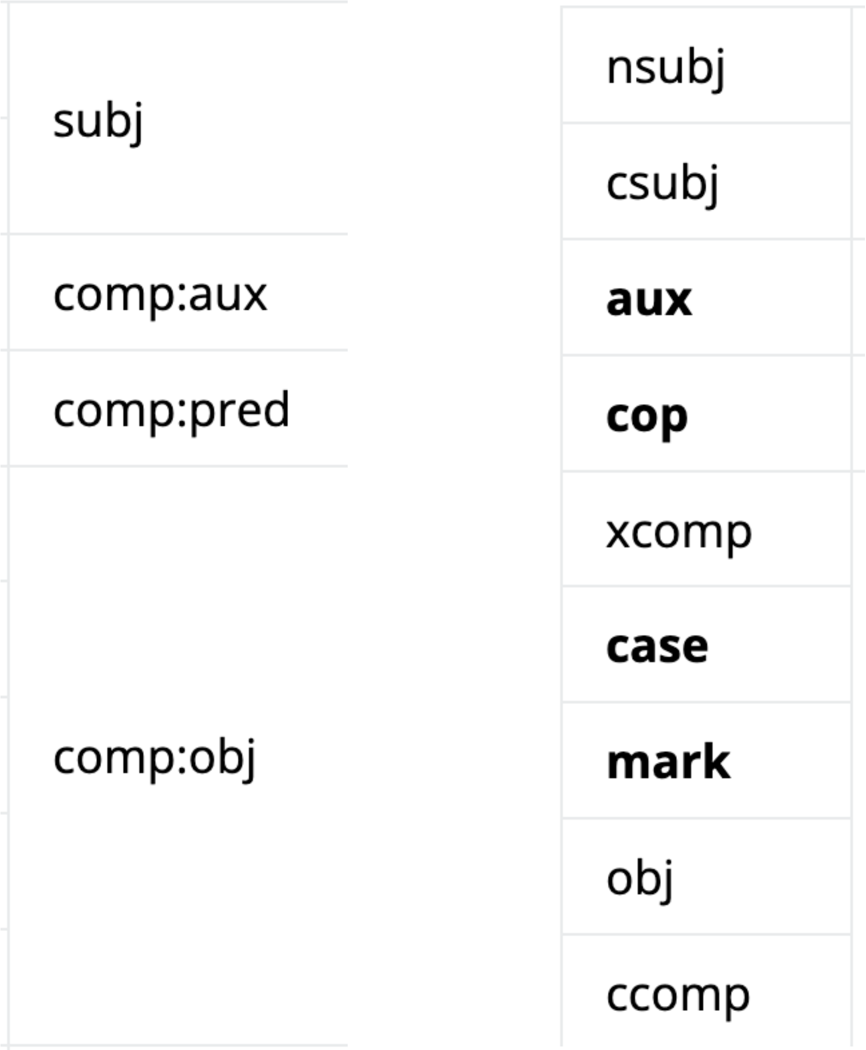
  - better for word order studies (VO ⬄ ADP-N)

SUD

SUD relations


UD
SUD
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

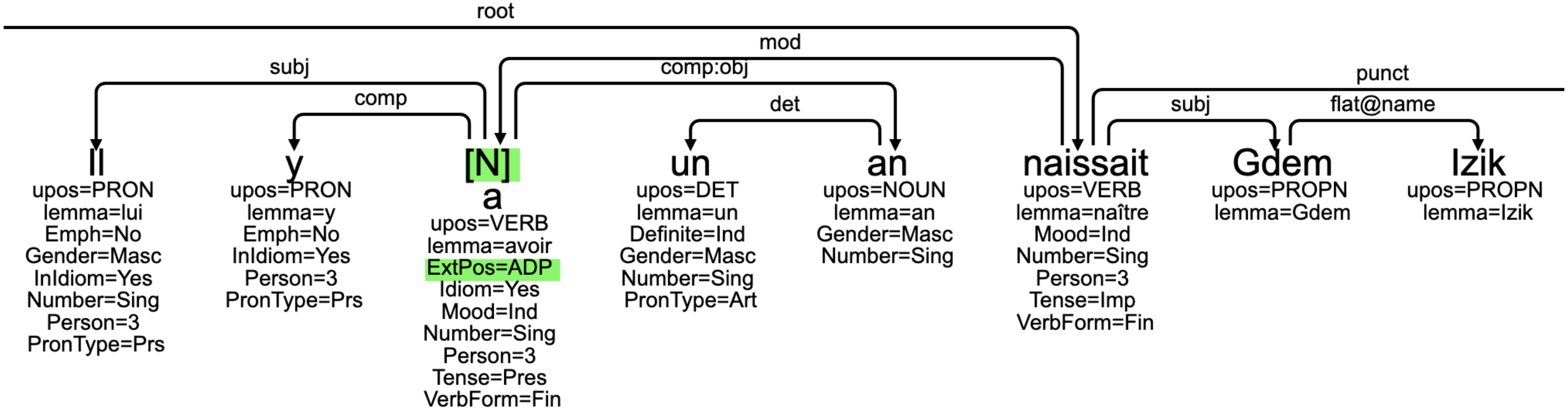
SUD => UD conversion



subj	nsubj	NOUN	
	csubj	VERB	
comp:aux	**aux**		*reversed*
comp:pred	**cop**		*reversed*
comp:obj	xcomp	VERB	
	case	NOUN	*reversed*
	mark	VERB	*reversed*
	obj	NOUN	
	ccomp		

SUD extensions (towards UD)

- Distinction between modifier and argument for oblique dependent
  - obl:mod => mod
  - obl:arg => comp:obl
- Annotation of the internal structure of MWEs

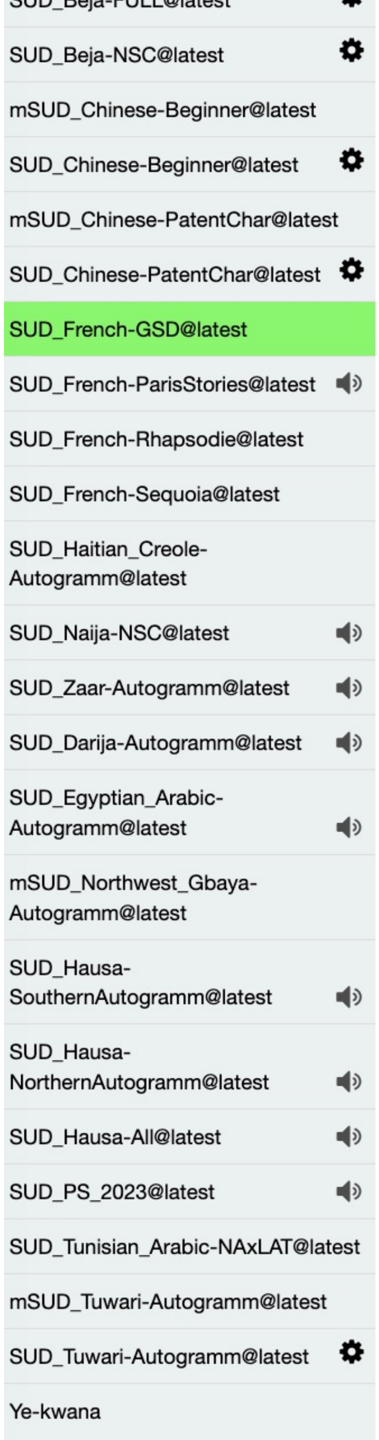


‘Gdem Izik was born one year **ago**.’, lit. **there is one year**

Conclusion

- about 20 native SUD treebanks
- SUD guidelines:
  - https://surfacesyntacticud.github.io/
- Automatic conversion SUD => UD
  - Grew (Graph rewriting grammar)

    (Bruno Guillaume 2023)
  - Possibility to mix SUD and UD
  - Possibility to add your own tags and to convert them in UD and SUD afterward

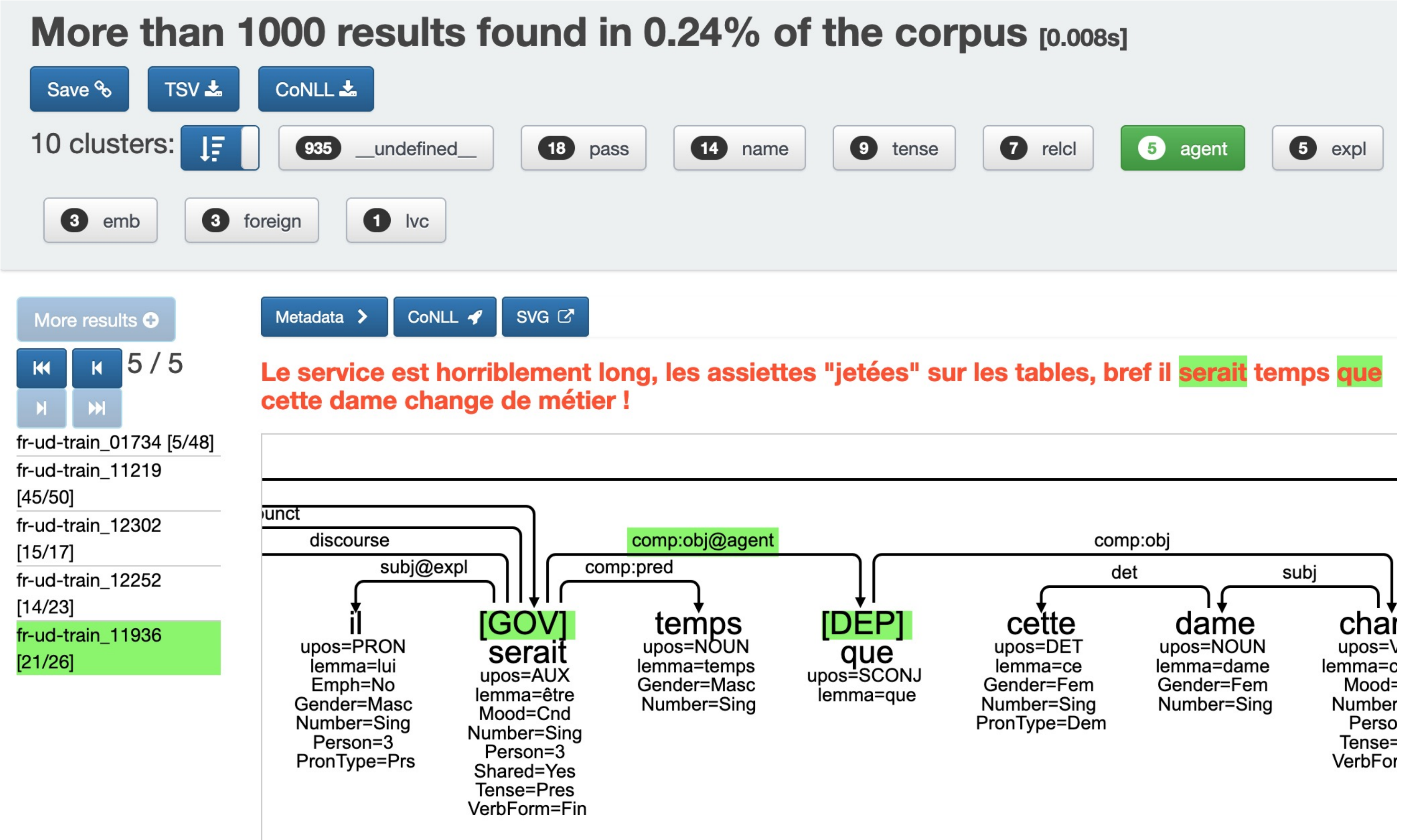


SUD_Beja-FULL@latest ⚙
SUD_Beja-NSC@latest ⚙
mSUD_Chinese-Beginner@latest
SUD_Chinese-Beginner@latest ⚙
mSUD_Chinese-PatentChar@latest
SUD_Chinese-PatentChar@latest ⚙
SUD_French-GSD@latest
SUD_French-ParisStories@latest 🔊
SUD_French-Rhapsodie@latest
SUD_French-Sequoia@latest
SUD_Haitian_Creole-Autogramm@latest
SUD_Naija-NSC@latest 🔊
SUD_Zaar-Autogramm@latest 🔊
SUD_Darija-Autogramm@latest 🔊
SUD_Egyptian_Arabic-Autogramm@latest 🔊
mSUD_Northwest_Gbaya-Autogramm@latest
SUD_Hausa-SouthernAutogramm@latest 🔊
SUD_Hausa-NorthernAutogramm@latest 🔊
SUD_Hausa-All@latest 🔊
SUD_PS_2023@latest 🔊
SUD_Tunisian_Arabic-NAxLAT@latest
mSUD_Tuwari-Autogramm@latest
SUD_Tuwari-Autogramm@latest ⚙
Ye-kwana

Thanks

Other differences between SUD and UD

- Semantic features on syntactic dependencies
  - deprel:subrel@deep



More than 1000 results found in 0.24% of the corpus [0.008s]

10 clusters: 935 __undefined__ · 18 pass · 14 name · 9 tense · 7 relcl · 5 agent · 5 expl · 3 emb · 3 foreign · 1 lvc

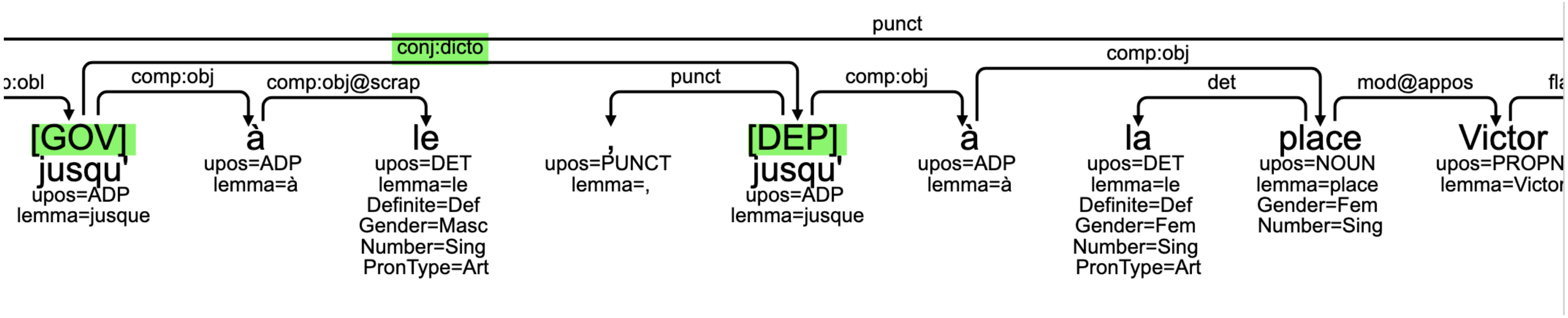
5 / 5

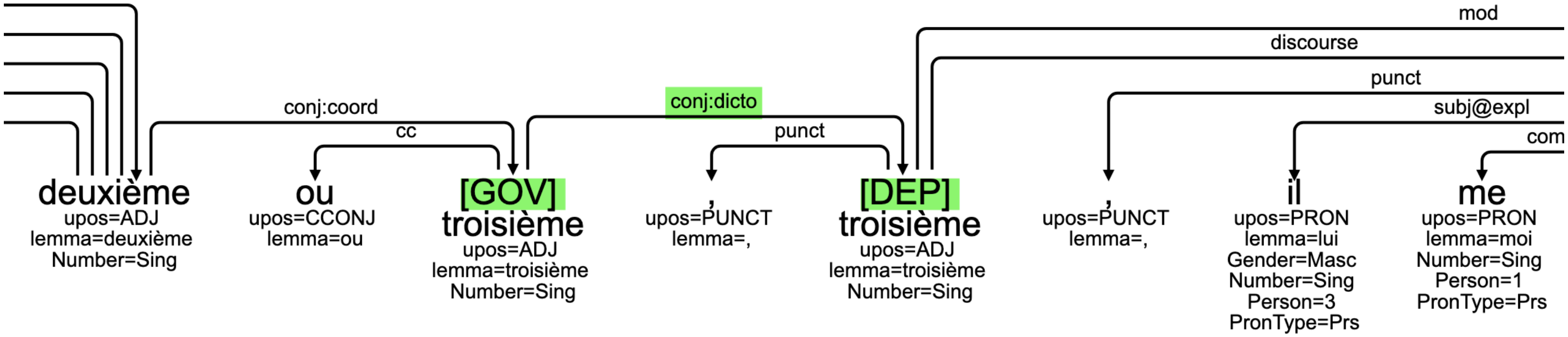
fr-ud-train_01734 [5/48]
fr-ud-train_11219 [45/50]
fr-ud-train_12302 [15/17]
fr-ud-train_12252 [14/23]
fr-ud-train_11936 [21/26]

Le service est horriblement long, les assiettes "jetées" sur les tables, bref il serait temps que cette dame change de métier !

Other differences between SUD and UD

- *conj:dicto* replaces *reparandum*





Annotation schema

- An annotation scheme is a **compromise** between:
  - – conception-oriented considerations:
    - what do authors want to do?
  - – annotator-oriented considerations:
    - how complicated will it be to annotate?
  - – end user-oriented considerations:
    - how will the resource be used?
  - – political considerations

(Gerdes & Kahane, 2016,

Gerdes & Kahane, 2016, LAW

Politics

P1. Visibility
P2. Availability of tools and guidelines
P3. Perspectives of richer collaborations

Conception-oriented consideration

A1. Adequacy
A2. Uniformity
A3. Level coverage
A4. Text coverage

Annotator-oriented considerations

B1. Formalization
B2. Simplicity
B3. Minimality
B4. Concision
B5. Naturalness
B6. Separability
B7. Independence
B8. Intuitiveness

End User-oriented considerations

- Theory
- NLP
- Pedagogy

C1. Quality
C2. Precision
C3. Learnability
C4. Readability
C5. Universality
C6. Transformability

UD schema

- “**What is needed for UD to be successful?** The secret to understanding the design and current success of UD is to realize that the design is a very subtle compromise between approximately 6 things:
  - UD needs to be satisfactory on linguistic analysis grounds for individual languages.
  - UD needs to be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
  - UD must be suitable for rapid, consistent annotation by a human annotator.
  - UD must be suitable for computer parsing with high accuracy.
  - UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. We refer to this as seeking a *habitable* design, and it leads us to favor traditional grammar notions and terminology.
  - UD must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, …).

It’s easy to come up with a proposal that improves UD on one of these dimensions. The interesting and difficult part is to improve UD while remaining sensitive to all these dimensions.”

(Chris Manning, UD web site, 2017)

UD dependencies

- “The Primacy of Content Words
  - Dependency relations hold primarily between content words
  - Function words (preposition, determiner, auxiliary …) attach as direct dependents of the most closely related content word
- Preferring content words as heads maximizes parallelism between languages because content words vary less than function words between languages. In particular, one commonly finds the same grammatical relation being expressed by morphology in some languages or constructions and by function words in other languages or constructions, while some languages may not mark the information at all (such as not marking tense or definiteness).”

UD relations

- microsyntax (syntactic function + POS)
  - nsubj vs csubj
  - obj vs. xcomp vs. ccomp
- piles
  - conj (coordination)
  - appos (double formulation)
  - reparandum (disfluency + reformulation)
- macrosyntax
  - parataxis (reported speech, parenthesis, verbal DM …)
  - discourse (non verbal discourse marker)
  - dislocated