Universal Dependencies for English

Nathan Schneider

July 31, 2017

https://static.pexels.com/photos/20974/pexels-photo.jpg

Why Dependencies?

- Dependency Grammar theories are based on the observation that many syntactic relationships can be characterized as *asymmetric*, binary relations between **head** and **modifier** words. (Tesnière 1959, Sgall et al. 1986, …)
  - If you learned sentence diagramming in grade school (Reed & Kellogg 1877), that is a form of dependency grammar!
  - Not all constructions fit cleanly (coordination, relative clauses, …); different theories have different solutions. **Labeling** the dependencies can clarify the nature of the relationship.
- While constituency grammars work well for “well-behaved” languages like English, Turkish and other languages introduce complications.
- Because dependency parses are structurally simpler, they are computationally easier to produce. (Faster parsers!)
- Syntactic dependencies are not too far from **semantic** dependencies, useful for many applications.

Universal Dependencies

- PTB is a *de facto* standard for constituency syntax, at least for English.
- But despite the popularity of dependencies, conventions/label sets abound.
  - Different sets of head rules for converting from PTB trees
  - Different edge labels for dependency treebanks
- **Universal Dependencies** (UD) are a recent (≈2014–2016) attempt to agree on cross-linguistic conventions.
  - Evolved from Stanford Typed Dependencies → Universal Stanford Dependencies
  - Headedness conventions and types designed for uniformity across languages
  - Also conventions for annotating morphology & POS, not discussed here
  - Guidelines and corpora from dozens of languages freely available at http://universaldependencies.org/

UD Treebanks

Ancient Greek	182K
Ancient Greek-PROIEL	198K
Arabic	217K
Arabic-NYUAD	629K
Basque	97K
Belarusian	6K
Bulgarian	140K
Catalan	472K
Chinese	111K
Coptic	3K
Croatian	183K
Czech	1,330K
Czech-CAC	482K
Czech-CLTT	26K
Danish	94K
Dutch	197K
Dutch-LassySmall	93K
English	229K
English-ESL	88K
English-LinES	67K
English-ParTUT	38K
Estonian	34K
Finnish	181K
Finnish-FTB	143K
French	381K
French-ParTUT	17K
French-Sequoia	58K
Galician	109K
Galician-TreeGal	14K
German	277K
Gothic	45K
Greek	51K
Hebrew	106K
Hindi	316K
Hungarian	37K
Indonesian	110K
Irish	13K
Italian	195K
Italian-ParTUT	39K
Japanese	173K
Japanese-KTC	189K
Kazakh	<1K
Korean	63K
Korean-Sejong	89K
Latin	18K
Latin-ITTB	280K
Latin-PROIEL	159K
Latvian	44K
Lithuanian	40K
Norwegian-Bokmaal	280K
Norwegian-Nynorsk	276K
Old Church Slavonic	47K
Persian	135K
Polish	72K
Portuguese	201K
Portuguese-BR	268K
Romanian	202K
Russian	87K
Russian-SynTagRus	988K
Sanskrit	1K
Slovak	93K
Slovenian	126K
Slovenian-SST	19K
Spanish	411K
Spanish-AnCora	495K
Swedish	76K
Swedish-LinES	64K
Swedish Sign Language	<1K
Tamil	8K
Turkish	46K
Ukrainian	12K
Urdu	123K
Uyghur	1K
Vietnamese	31K

Upcoming UD Treebanks

Amharic	-
Buryat	-
Cantonese	-
Chinese-HK	-
Faroese	-
Kurmanji	-
Marathi	-
Serbian	-
Somali	-
Sorani	-

as of March 2017

Manning's Law


Chris Manning

From http://universaldependencies.org/introduction.html:

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:

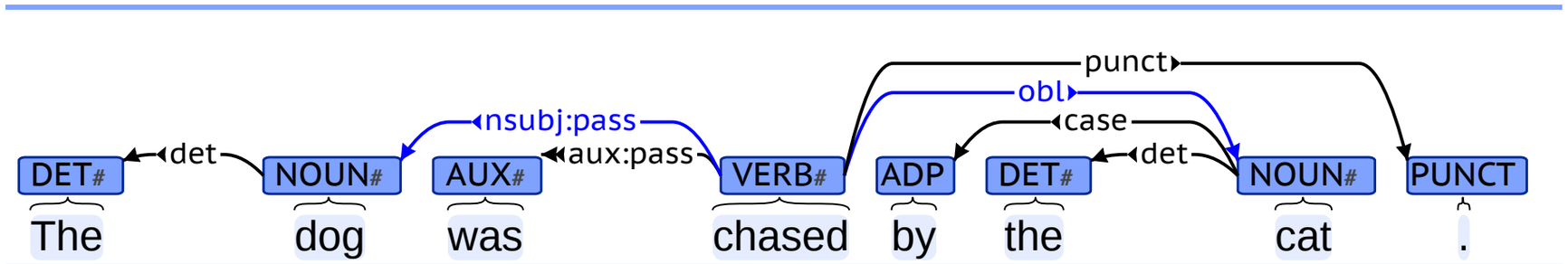
1. UD needs to be satisfactory on linguistic analysis grounds for **individual** languages.
2. UD needs to be good for linguistic **typology**, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
3. UD must be suitable for rapid, consistent **annotation** by a human annotator.
4. UD must be suitable for computer **parsing** with high accuracy.
5. 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.
6. 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.

Cross-linguistic Parallelism

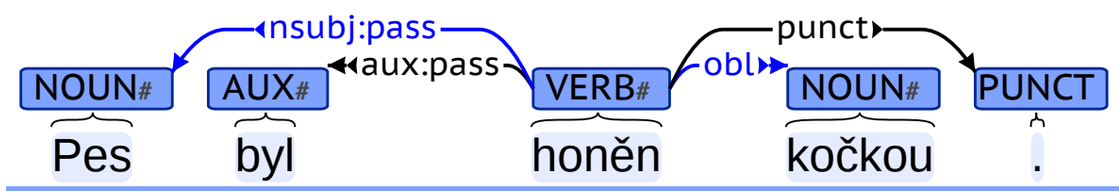
Examples from http://universaldependencies.org/introduction.html:

English

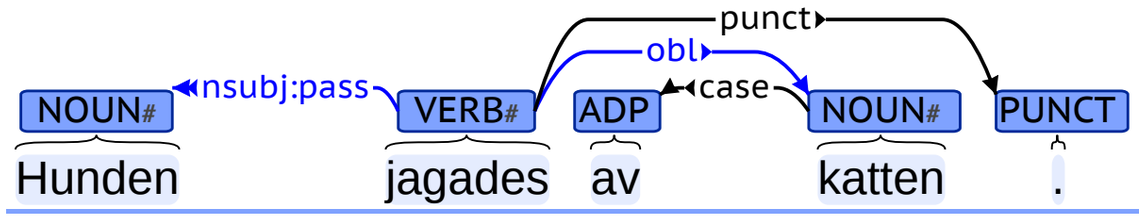


Bulgarian

Czech



Swedish

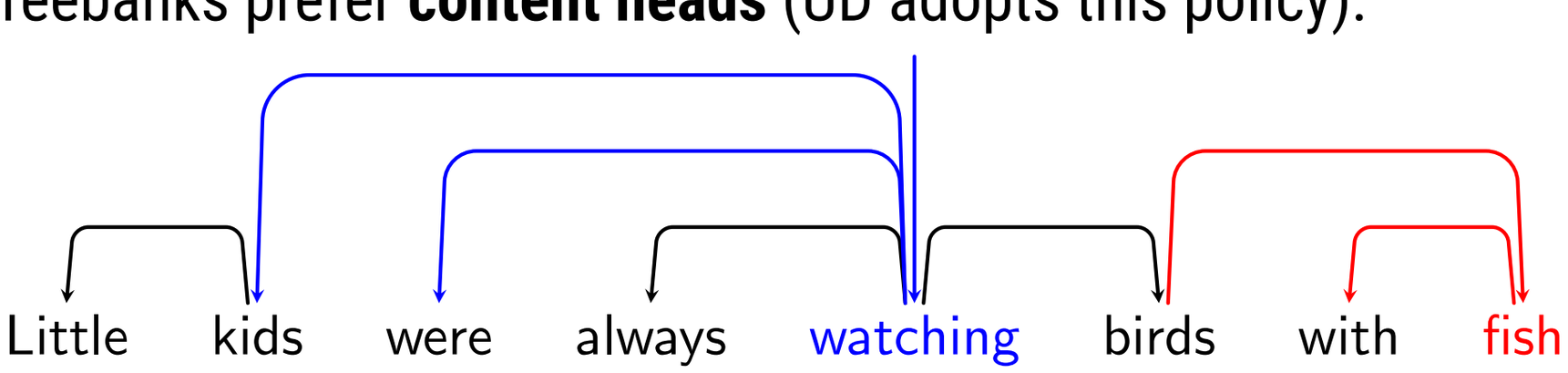


Content vs. Functional Heads

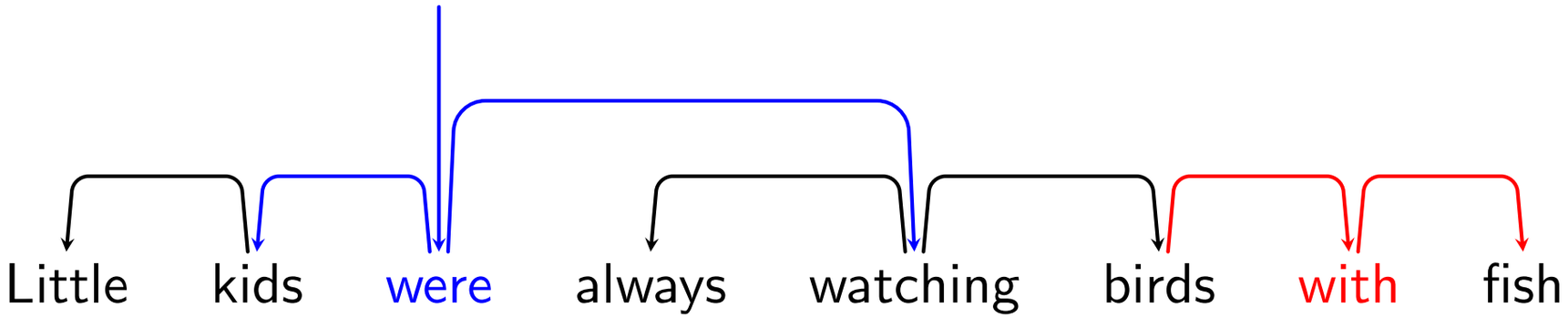
- Between two related **content** words, deciding which is the head (the direction of the arrow) is usually easy: e.g., *catch* → *fish* and *cute* ← *puppies*.

- **Function** words like auxiliaries, copulas, and adpositions are trickier.

- Some treebanks prefer **content heads** (UD adopts this policy):



- Others prefer **functional heads**:



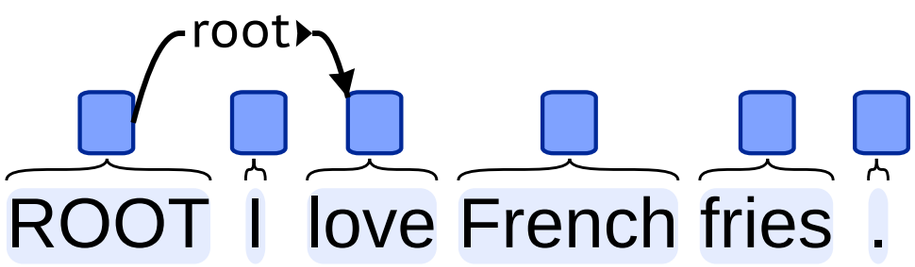
UD Annotation for English: A Crash Course

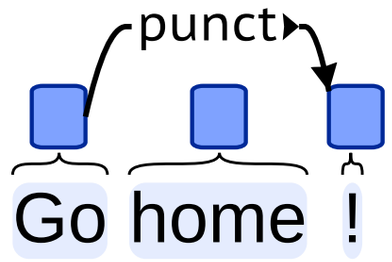
Adapted from the v2 Universal guidelines at http://universaldependencies.org/ with additional examples from the main English UD treebank; refer to the website for many, many additional details

Root, Punctuation

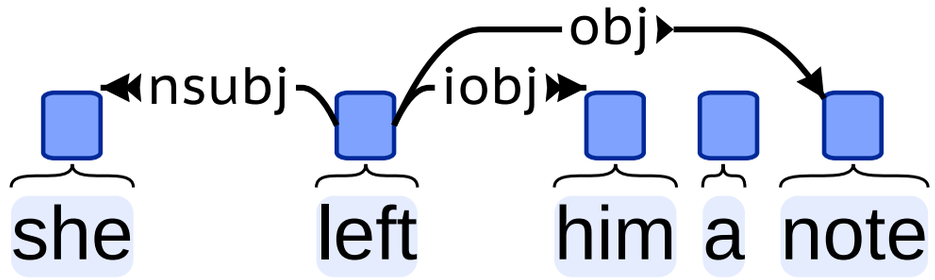
root the only word not headed by any other; usually the main predicate
Can be drawn as an unlabeled edge coming from above the sentence, or coming from a dummy ROOT node.

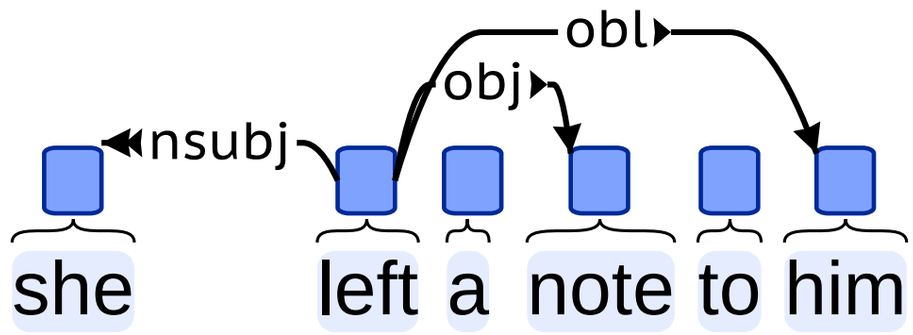
punct any punctuation token, attached to the head of its nearest containing phrase (often the head of the clause)

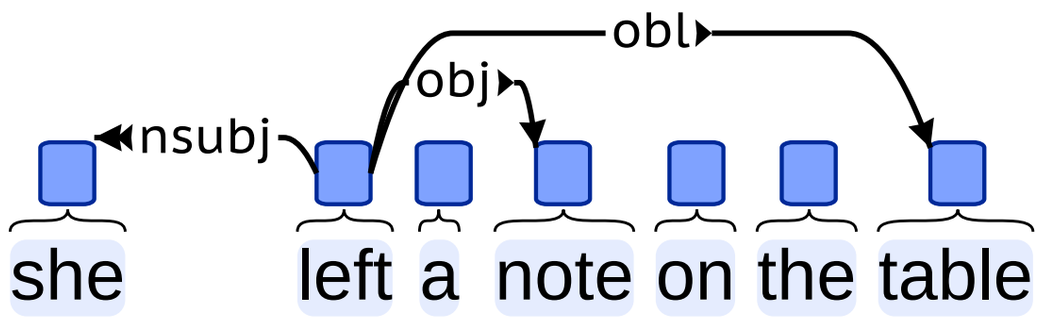




Subject, Object, Oblique

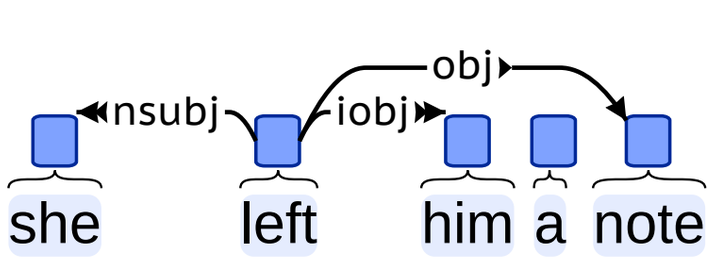


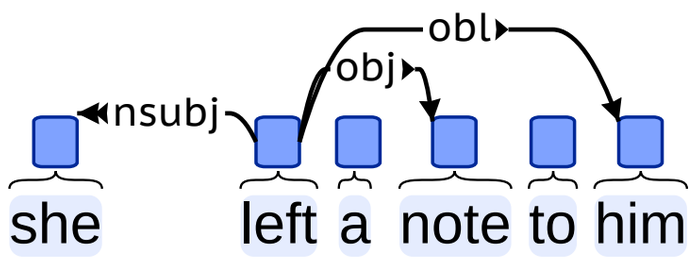


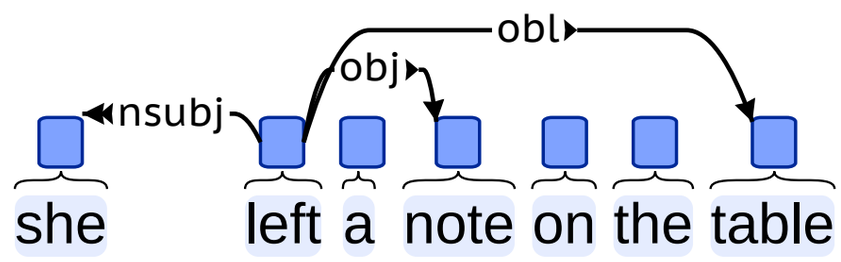


Subject, Object, Oblique

subjects	objects	obliques
nsubj nominal subject	**obj** direct object	**obl** case-marked noun
	iobj indirect object	

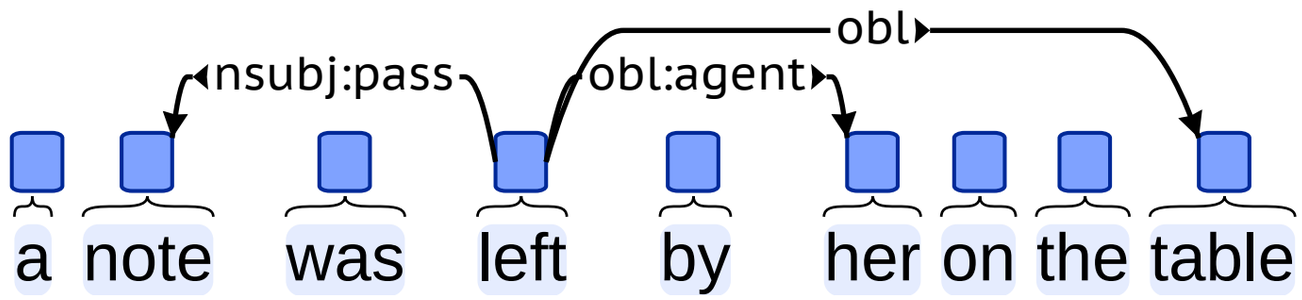


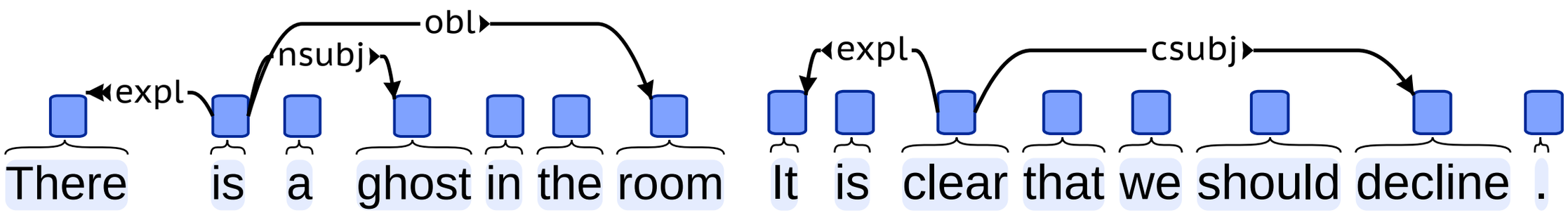




Subject, Object, Oblique

subjects	objects	obliques
nsubj nominal subject	**obj** direct object	**obl** case-marked noun
nsubj:pass nominal subject of passive	**iobj** indirect object	**obl:agent** passive *by* argument*
csubj clausal subject	**advmod** modifying adverb	**obl:tmod** temporal noun (adverbial or case-marked)
csubj:pass clausal subject of passive		
expl expletive subject		





* Not distinguished from **obl** in the English UD treebank.

Auxiliaries, Copulas, Case

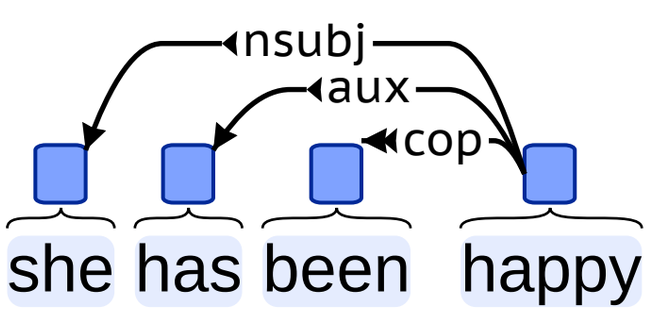
Remember: these are **function** words, so they modify content words!

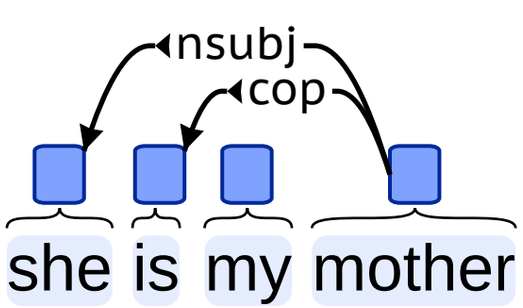
aux auxiliary

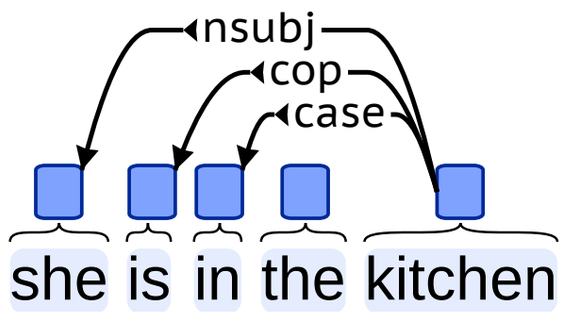
cop copula

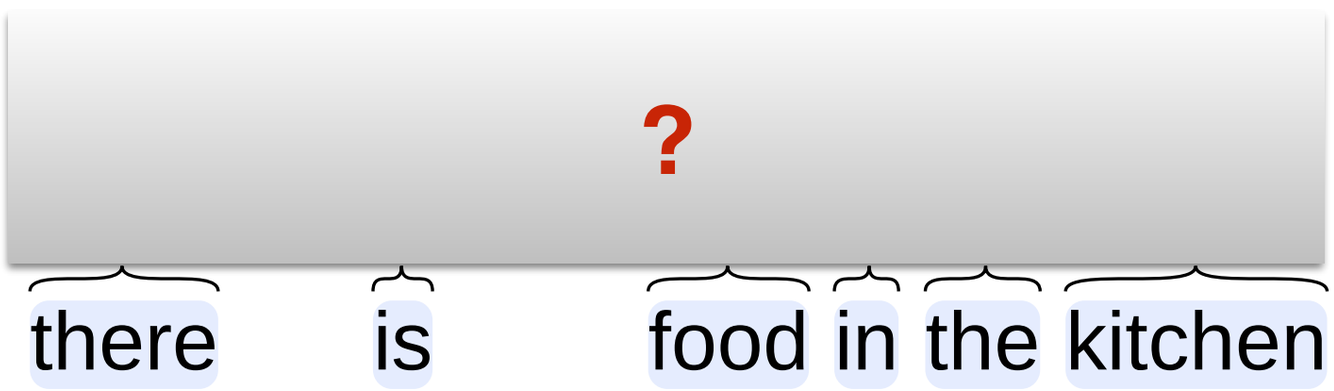
case preposition or case clitic modifying a nominal

aux:pass passive auxiliary (form of *be* or *get*)









Auxiliaries, Copulas, Case

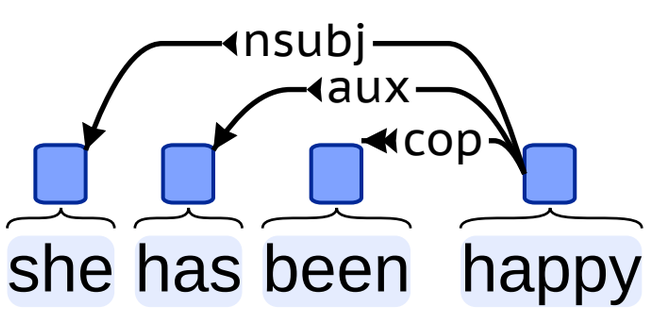
Remember: these are **function** words, so they modify content words!

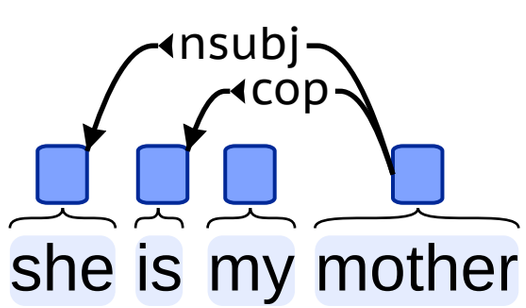
aux auxiliary

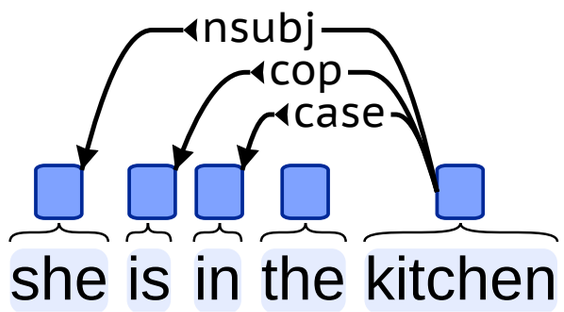
cop copula

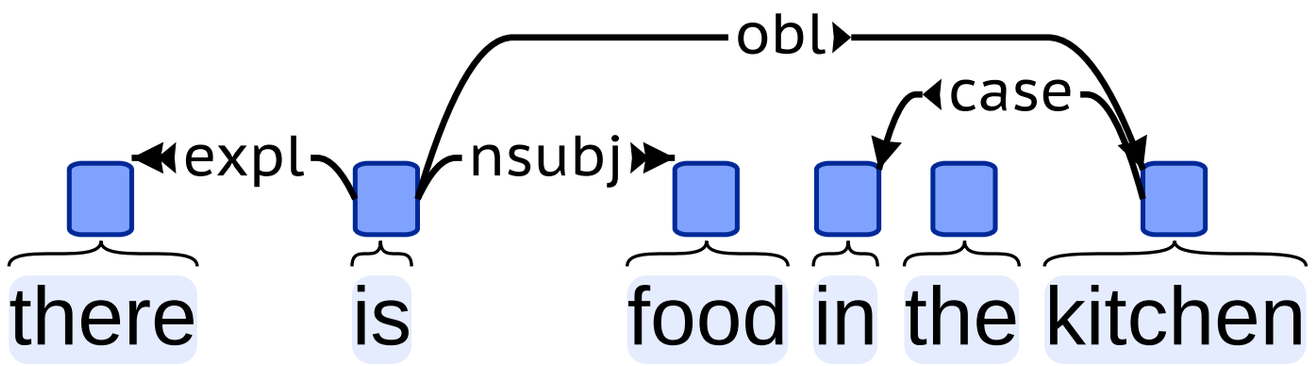
case preposition or case clitic modifying a nominal

aux:pass passive auxiliary (form of *be* or *get*)









Adjectives, Determiners, Nominal modifiers

amod modifying adjective

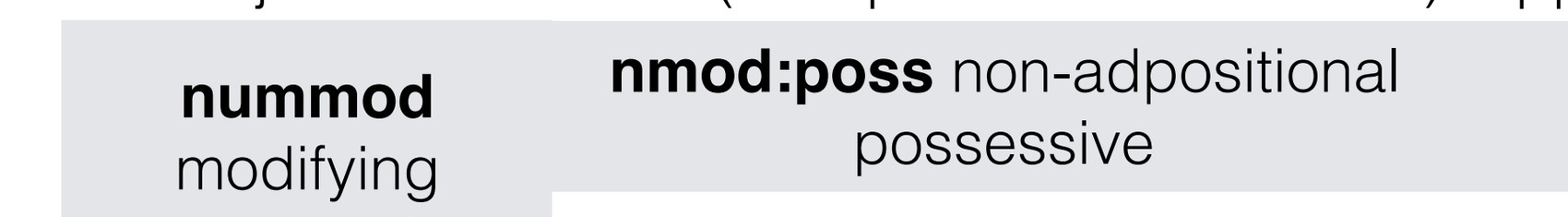
nmod modifying nominal with case (except at the clause level)

appos appositive

nummod modifying numeral

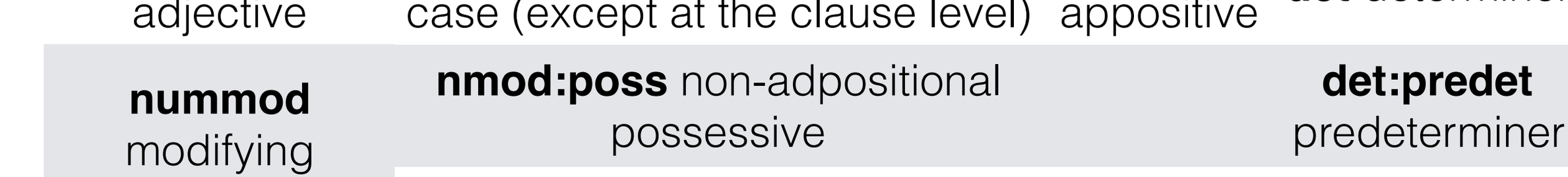
nmod:poss non-adpositional possessive

nmod:tmod modifying temporal nominal in an NP



Adjectives, Determiners, Nominal modifiers

amod modifying adjective	**nmod** modifying nominal with case (except at the clause level)	**appos** appositive	**det** determiner
nummod modifying numeral	**nmod:poss** non-adpositional possessive		**det:predet** predeterminer
	nmod:tmod modifying temporal nominal in an NP		



Compounds, Flat names, Fixed expressions

compound

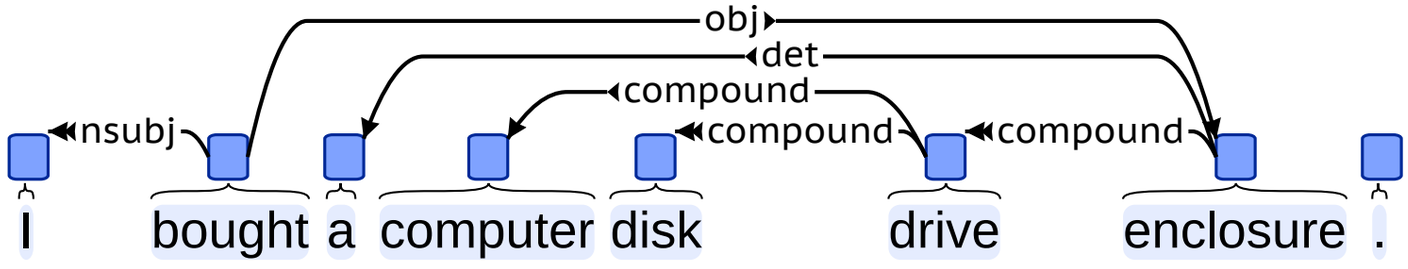
compound:prt verb particle

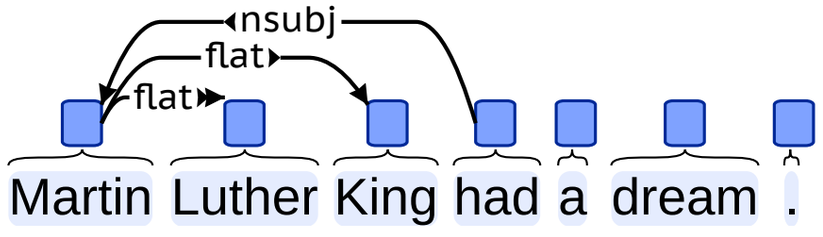
compound:svc serial verb construction

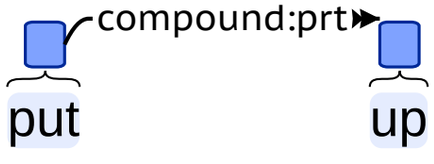
flat names without head-modifier structure

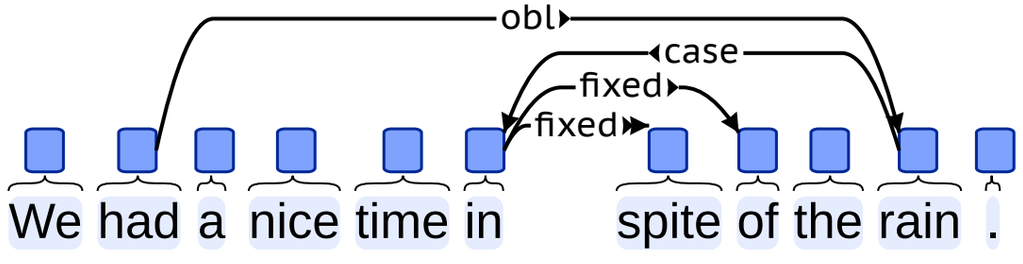
fixed fixed grammatical expressions

*With **fixed** and **flat**, the first word heads all other words in the expression.*





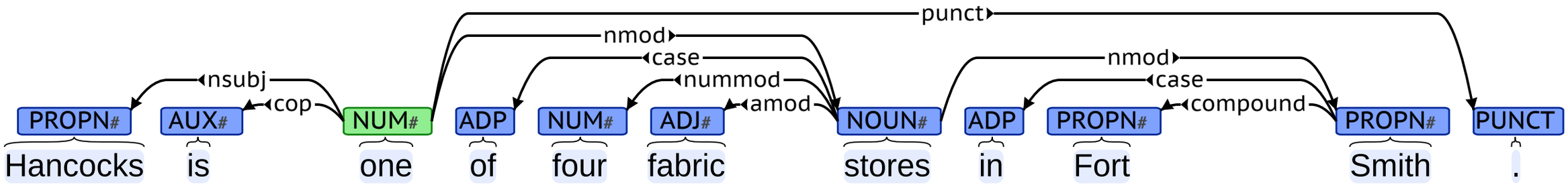




Example

								?					
Hancocks	is		one	of	four	fabric		stores	in	Fort		Smith	.

Example

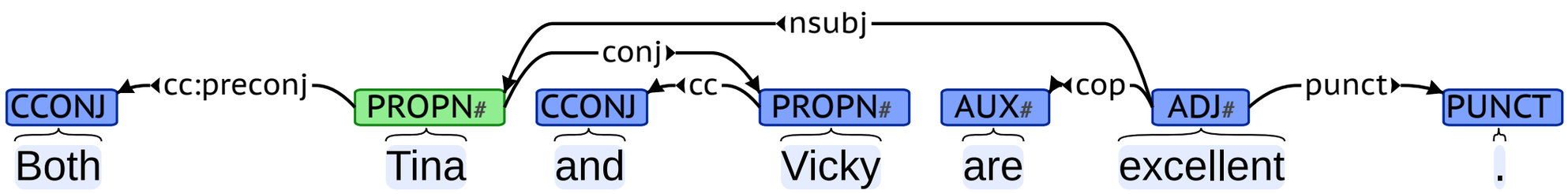

punct
nmod
case
nummod
amod
nsubj
cop
nmod
case
compound
PROPN# AUX# NUM# ADP NUM# ADJ# NOUN# ADP PROPN# PROPN# PUNCT
Hancocks is one of four fabric stores in Fort Smith .

Coordination

conj non-initial conjunct

cc coordinating conjunction
(attaches to successive conjunct)

cc:preconj preconjunction

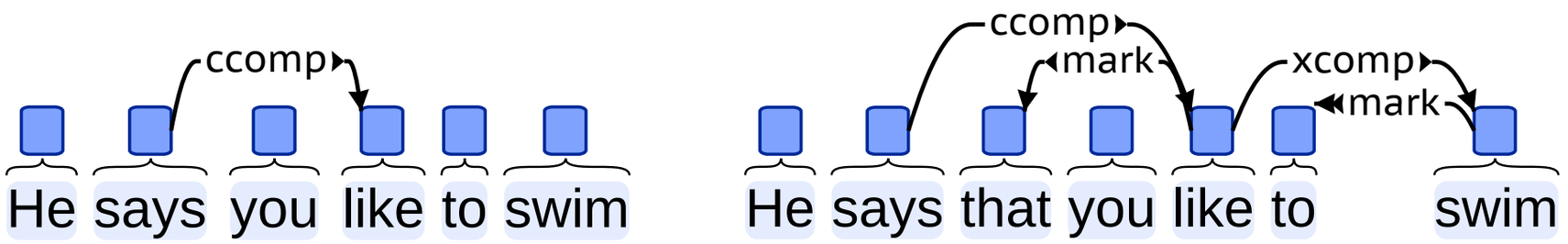


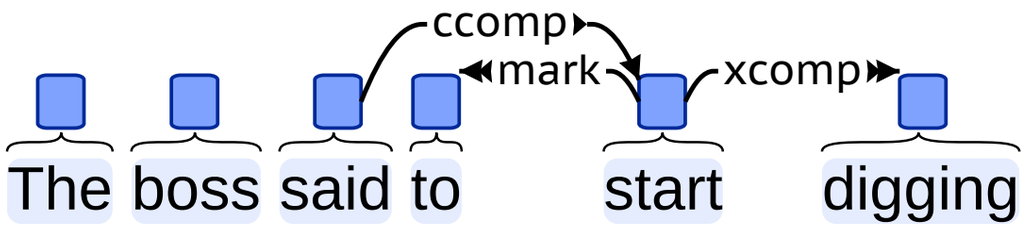
Complement Clauses

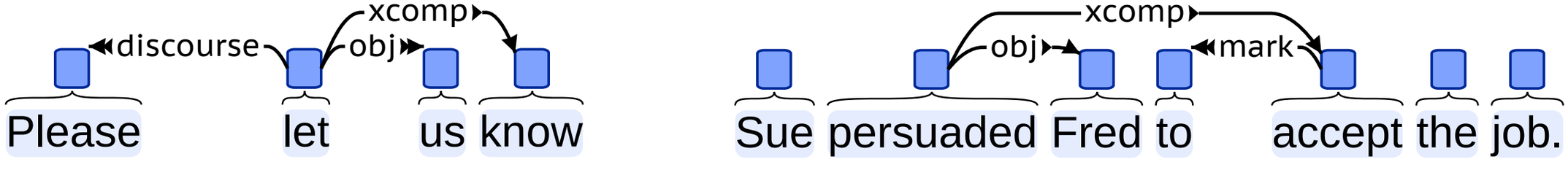
ccomp clausal complement

mark subordinator, complementizer, or infinitive marker

xcomp a predicate's clausal (or predicate A/N) complement that shares an argument with the matrix predicate





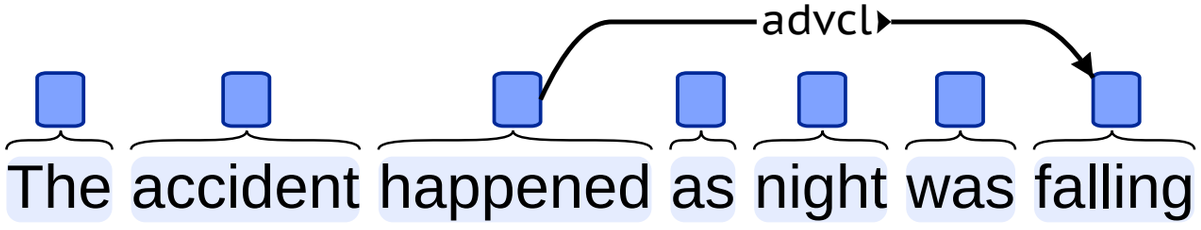


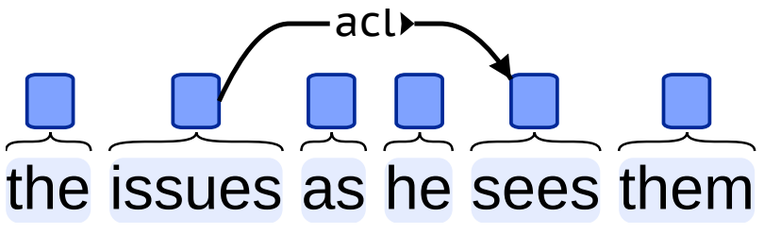
Modifier Clauses

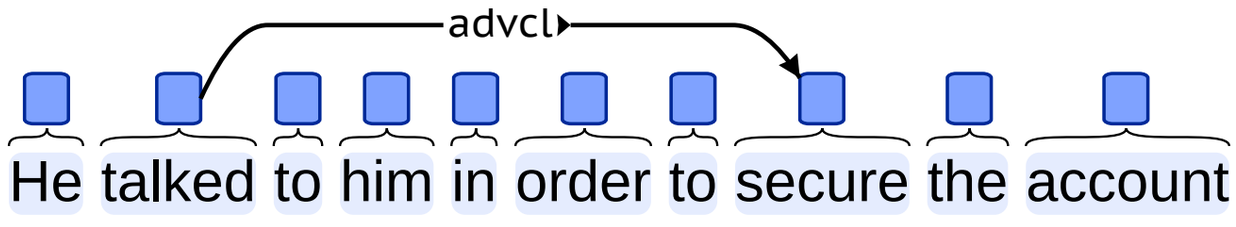
advcl adverbial clause (e.g. expressing time, purpose, reason, condition…)

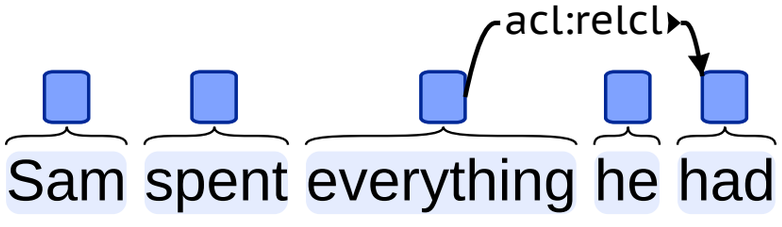
acl adjectival clause

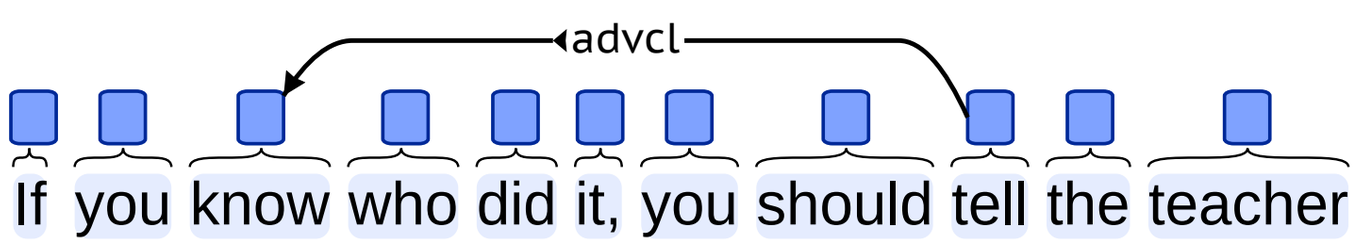
acl:relcl relative clause

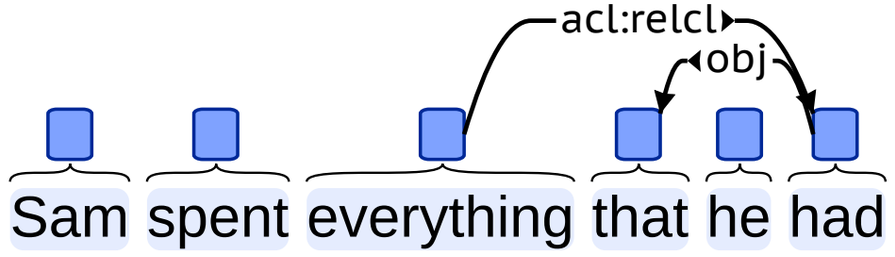






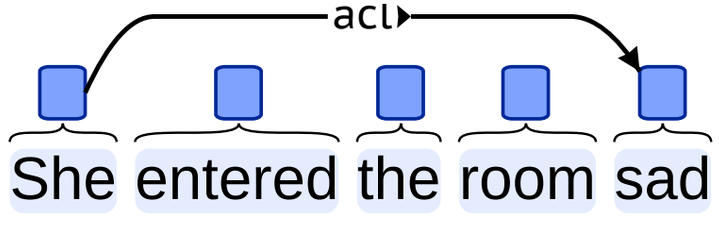




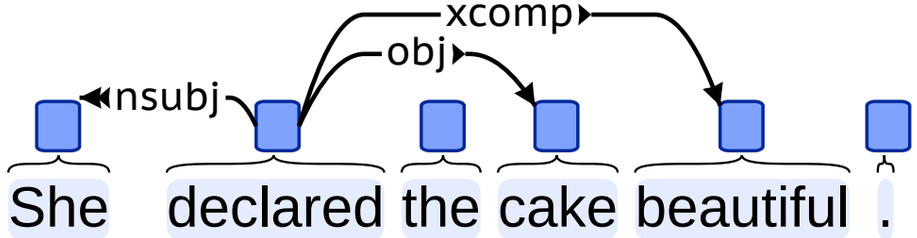


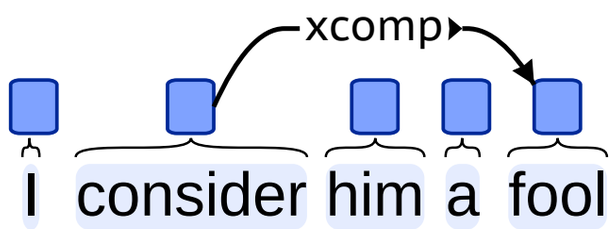
Depictives, Resultatives, Secondary Predicates

Depictive, not a dependent of verb

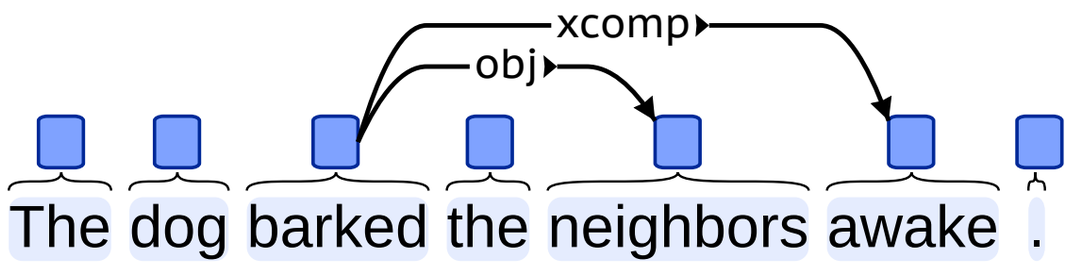


Obligatory argument of verb which is understood as **predicating** one of the verb's nominal arguments



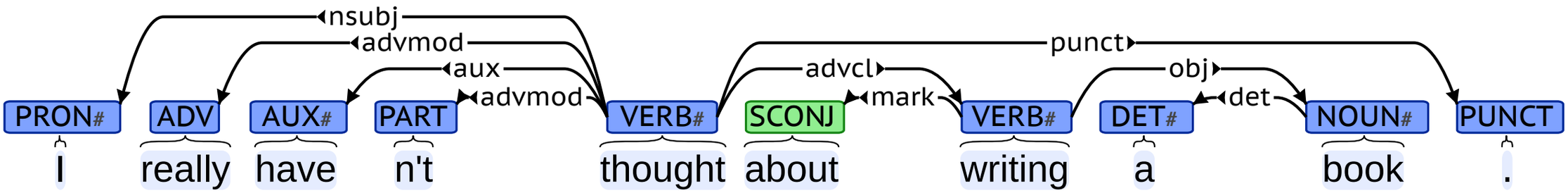


Resultative, predicate indicating an outcome of the verbal event on one of its nominal arguments



Example

Example



*N.B. This is an example from the English treebank, but it is debatable whether **advcl** is correct.*

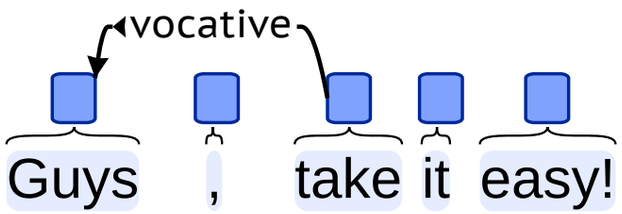
https://static.pexels.com/photos/20974/pexels-photo.jpg

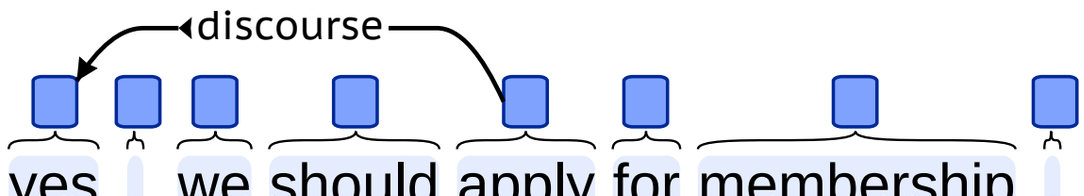
Discourse Stuff™ 1

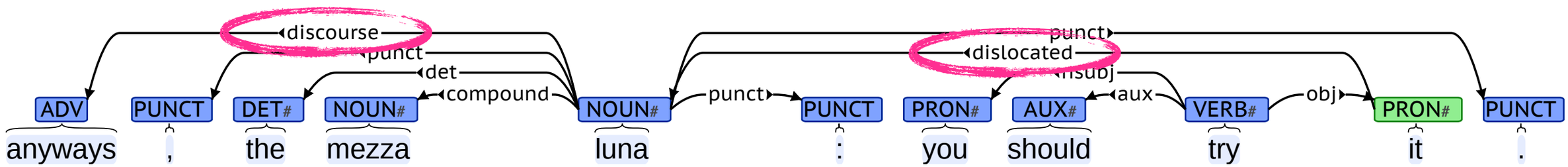
vocative addressee

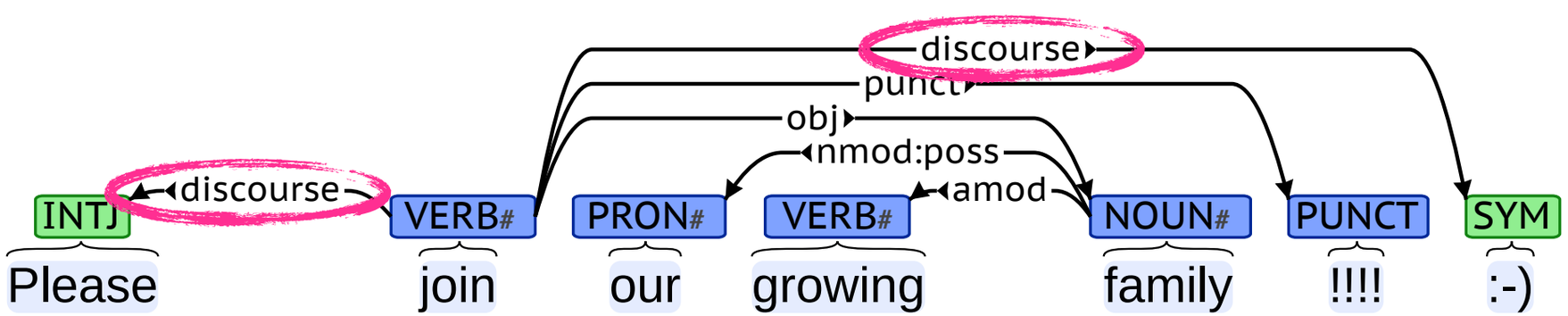
dislocated topicalized noun phrase

discourse expression functioning as an interjection, filler, or similar conversational marker





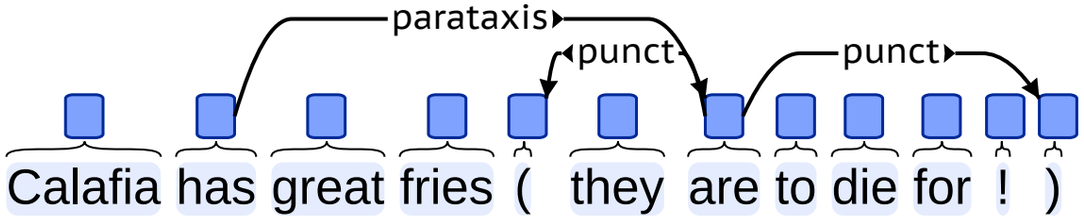


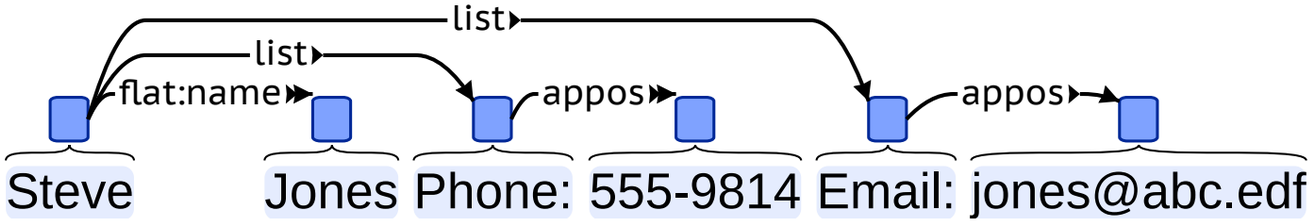


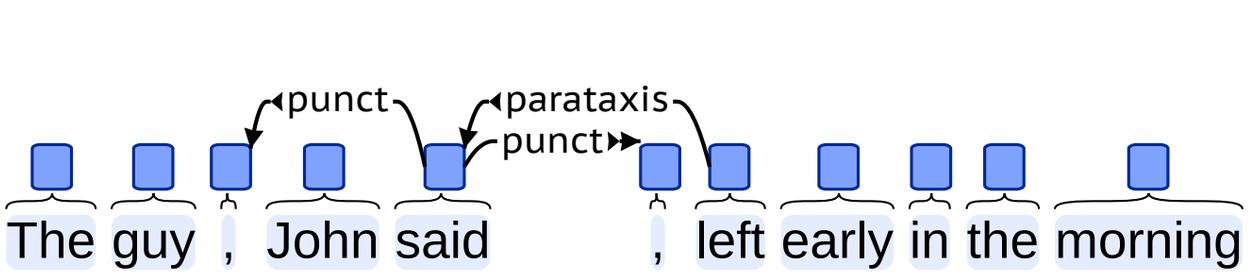
Discourse Stuff™ 2

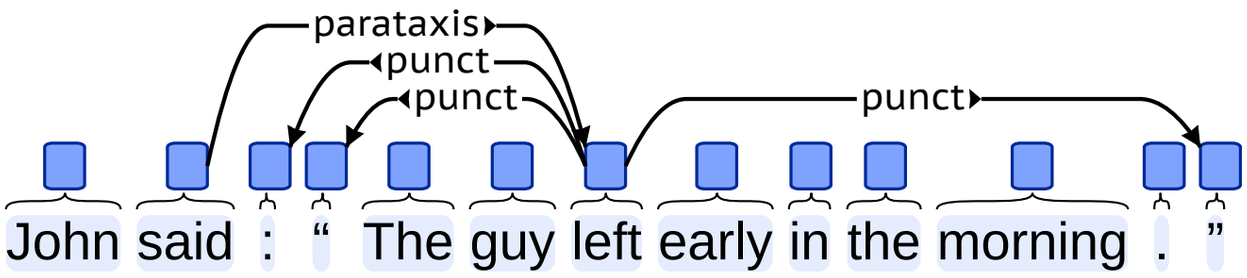
parataxis independent clauses/fragments forming a larger sentence, ideally separated with punctuation (but no conjunction); includes parentheticals, reported speech, tag questions

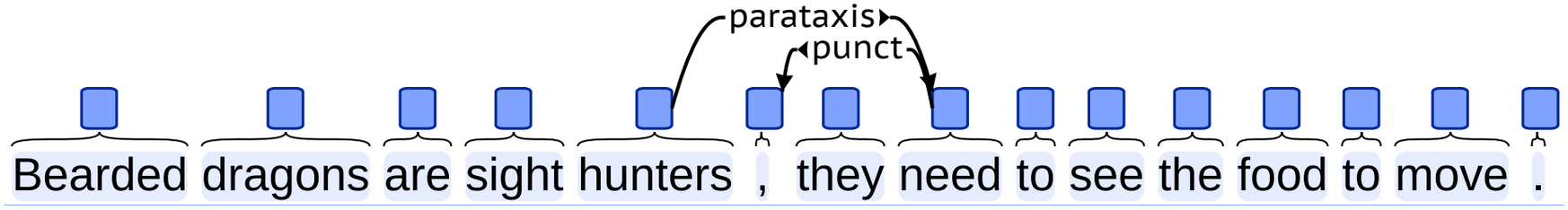
list items that do not form a syntactic sentence







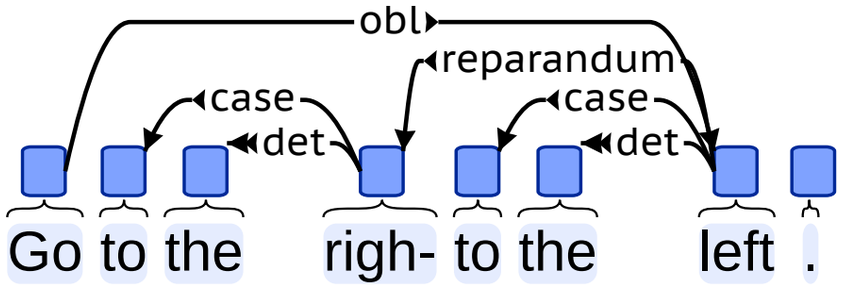


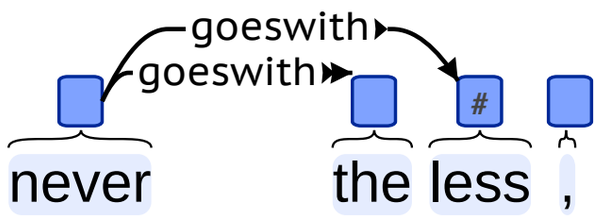


Speech Errors and Overtokenization

reparandum superfluous word or phrase, such as a speech error

goeswith superfluous space between words (would normally be written as a single word). *As with **fixed** and **flat**, the first word heads all other words in the expression.*



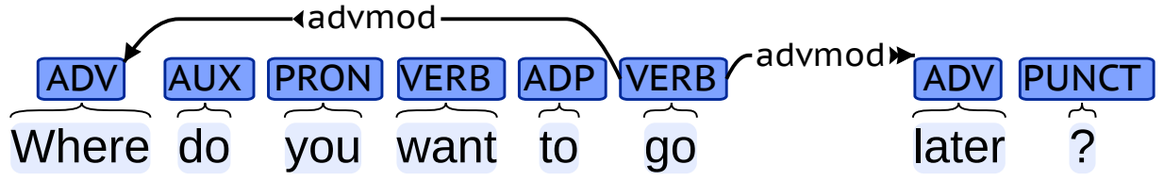


Questions

- There are no special dependency types for questions (or, for that matter, imperatives, which simply lack an overt subject).

- For yes/no questions, try rephrasing as a confirmation question. The dependencies will be the same.
  - Do you like my hat? ⇒ You *do* like my hat?
  - Is this a hat? ⇒ This *is* a hat?

- **For WH-questions, rephrase with an *in situ* WH-word.**
  - Why do you like my hat? ⇒ You do like my hat *why*?
  - What did you eat? ⇒ You did eat *what*?
  - Who do you think wants my hat? ⇒ You do think (that) *who* wants my hat?

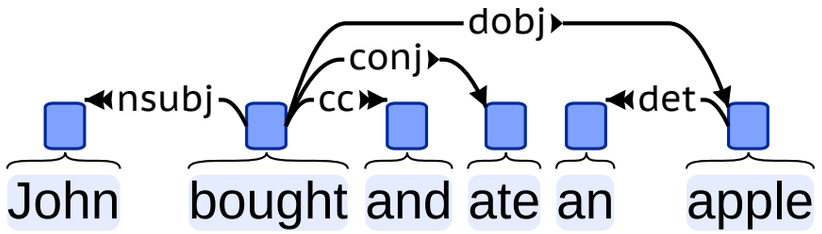


Ellipsis

From http://universaldependencies.org/u/overview/specific-syntax.html#ellipsis:

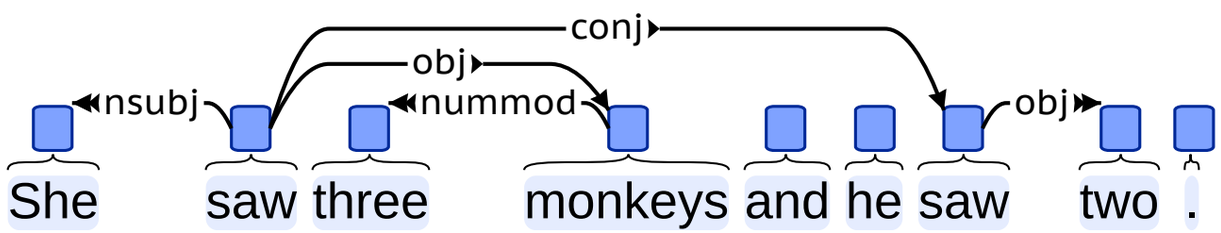
The UD approach to ellipsis can be summarized as follows:

1. If the elided element has no overt dependents, we do nothing.



elided: ~~an~~ ←det ~~apple~~

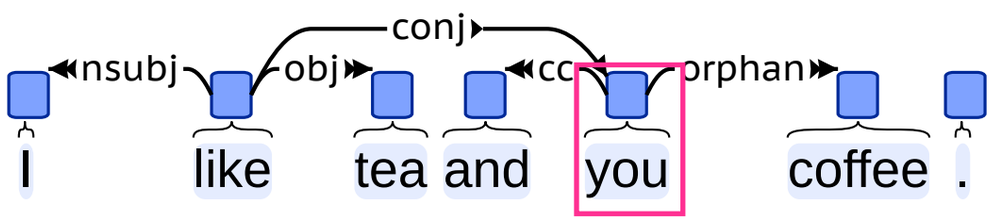
2. If the elided element has overt dependents, we **promote** one of these to take the role of the head.



elided: **two** ←nummod ~~monkeys~~
promoted

3. If the elided element is a predicate and the promoted element a core argument, we use the **orphan** relation when attaching other non-functional dependents to the promoted head.

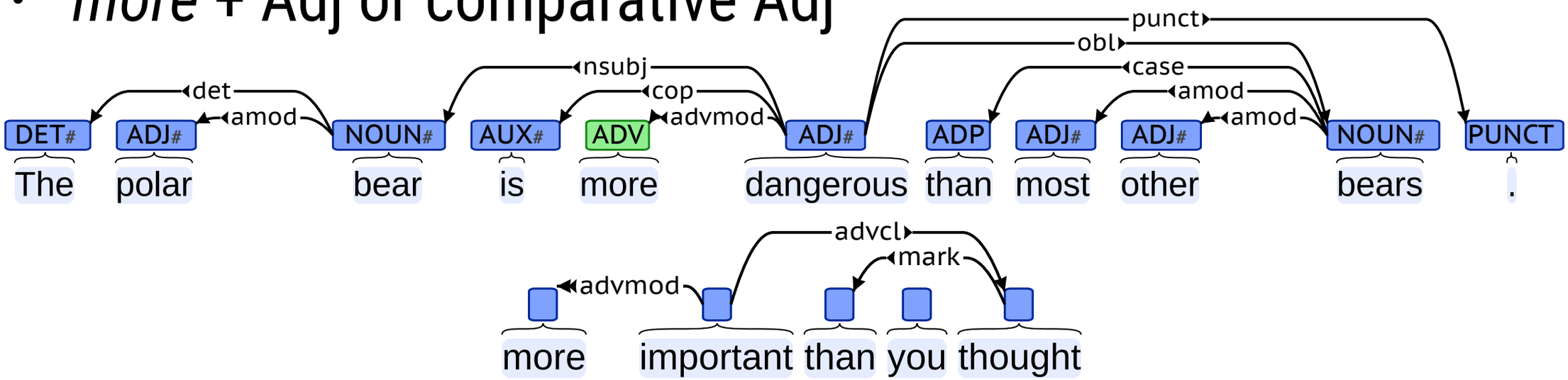
orphan dependent of elided material

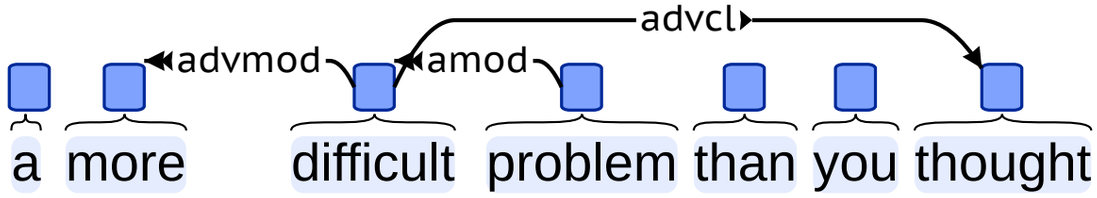


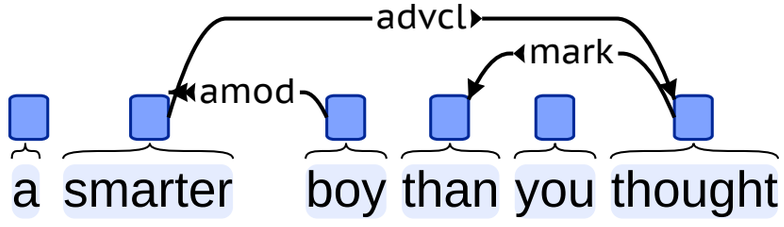
elided: **you** ←nsubj ~~like~~ obj→ **coffee**
promoted ***orphaned***

Comparatives 1

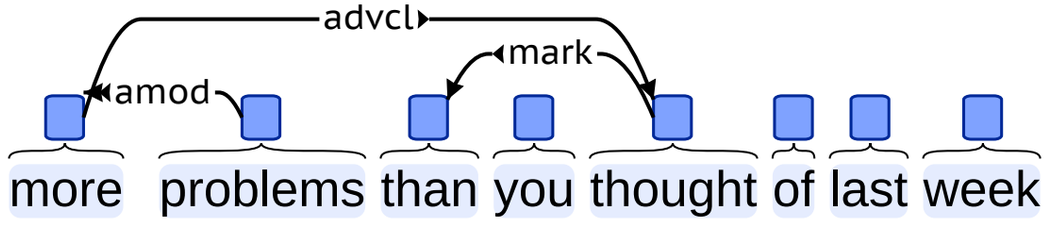
- *more* + Adj or comparative Adj

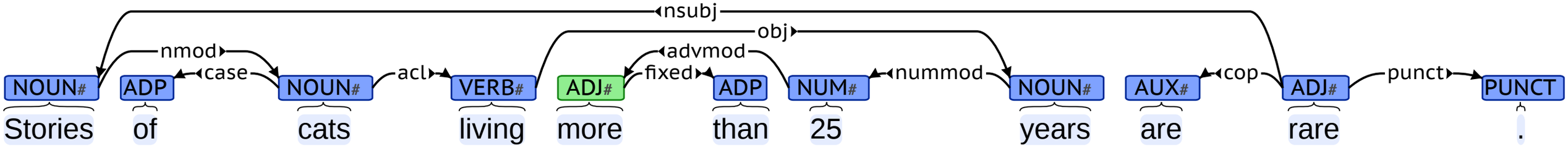






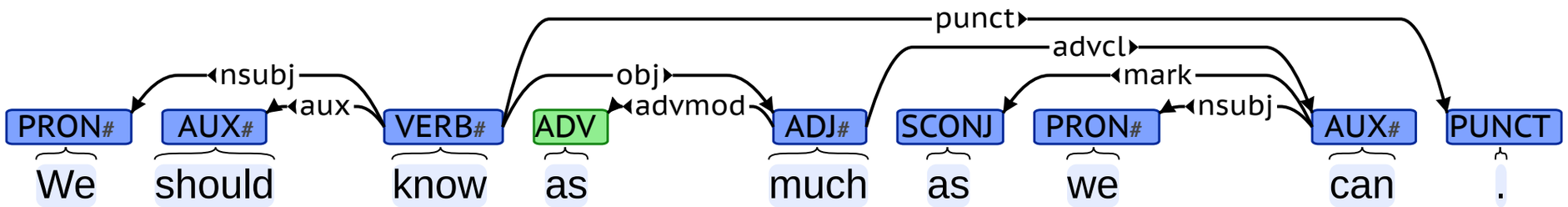
- *more* as a quantity



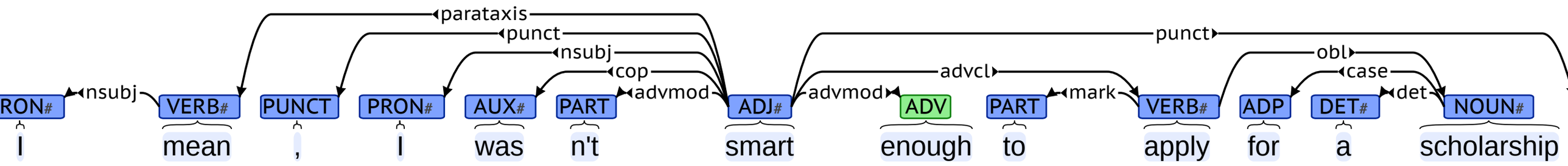


Comparatives 2

- *as-as*



- *enough*



flat:foreign

- E.g., *ad hoc*

nmod:npmod and obl:npmod

- These subtypes are (IMO, unintuitively) applied to **rates**, **compounds where only one of the words is a noun**, and a few other postmodifier-of-a-noun constructions.

- Details: https://github.com/UniversalDependencies/docs/issues/478

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

* The advmod relation is used for modifiers not only of predicates but also of other modifier words.

http://universaldependencies.org/u/dep/

The 37 “universal” relations (omits subtypes; **clf** – *classifier* not used for English)

http://bionlp-www.utu.fi/dep_search

UD Treebank Search

If you see problems in the online guidelines/data

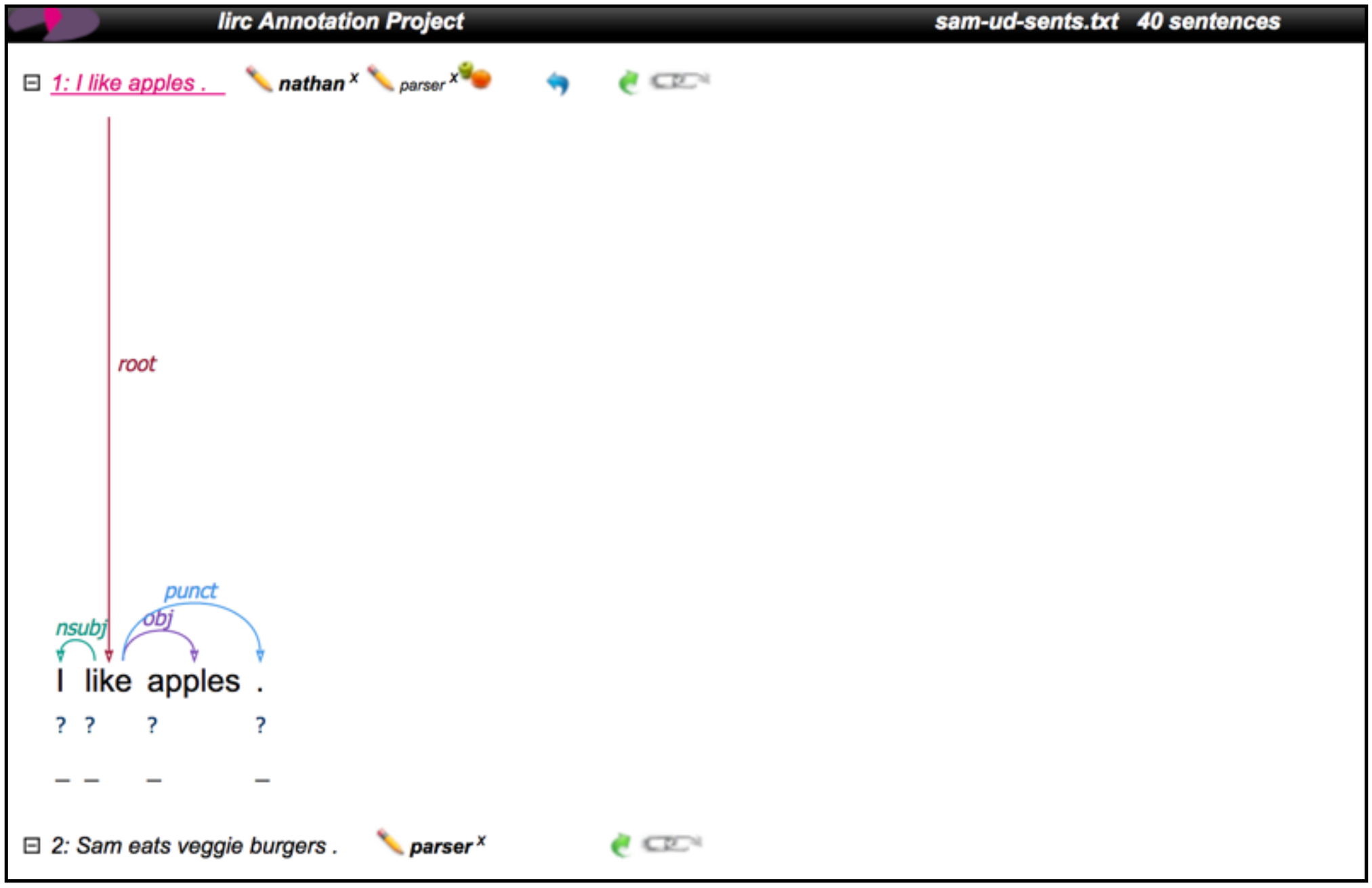
https://github.com/UniversalDependencies/docs/issues

Parsers

- The development of parsing algorithms is a major topic of NLP research.
  - Tradeoffs: **accuracy**, **speed**, **complexity** (constituency parse more complex than dependency parse)
  - For *Wall Street Journal* news, state-of-the-art accuracies are in the low-to-mid 90% range!
  - But HUGE variation in accuracy for other genres and languages
- Many parsers are open source. E.g. Stanford Parser, TurboParser, spaCy
  - May require you to use a command line interface or a programming language
- Web demos that sometimes work: **Stanford** (http://corenlp.run/—currently UDv1), **TurboParser** (http://demo.ark.cs.cmu.edu/parse—Stanford Dependencies, not quite UD!)

https://github.com/kimgerdes/arborator/

Arborator



Thanks

Marie Catherine de Marneffe

Chris Manning

Sebastian Schuster

Amir Zeldes

Yi Zhu

Students in the Corpus Linguistics course at the
2017 Linguistic Institute, Lexington, KY