



Universal Dependencies (vs. PDT): Syntactic Annotation

Markéta Lopatková

Institute of Formal and Applied Linguistics, MFF UK

lopatkova@ufal.mff.cuni.cz

Universal Dependencies (UDs)



idea:

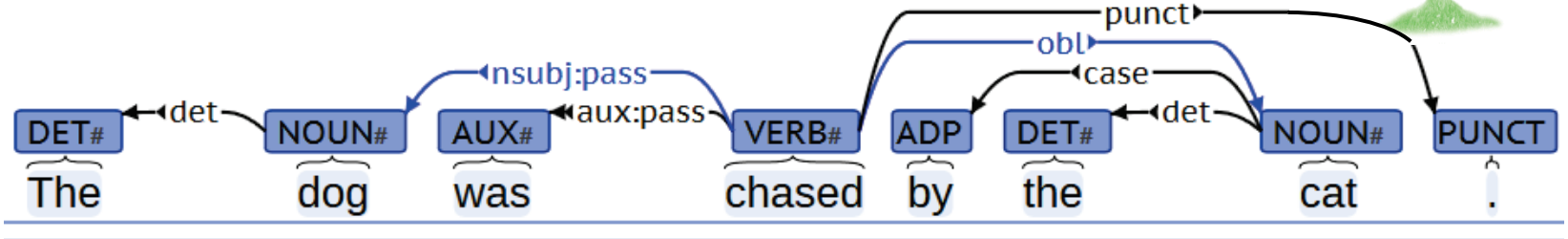
- cross-linguistically consistent treebank annotation
- annotation for different languages as similar as possible
- support multilingual NLP (parser development, cross-lingual learning, ...) and linguistic research

<http://universaldependencies.org/>

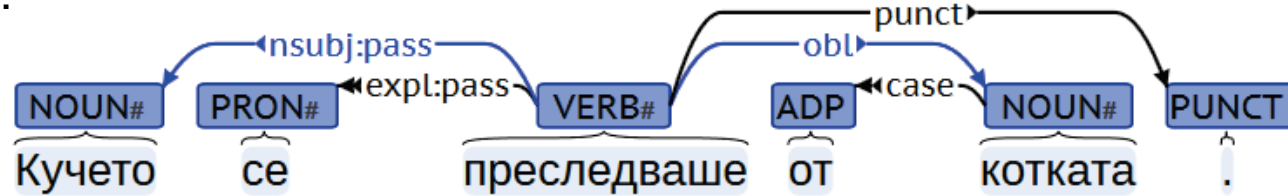
Universal Dependencies (UDs)



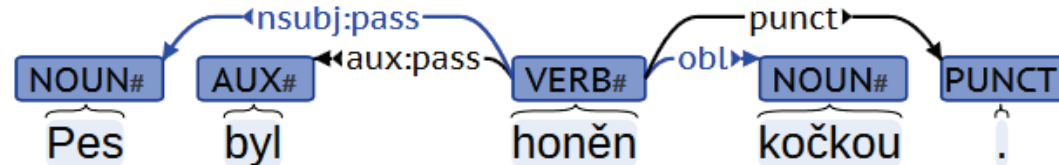
English:



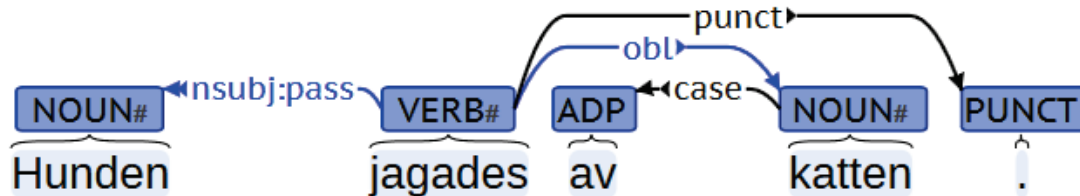
Bulgarian:



Czech:

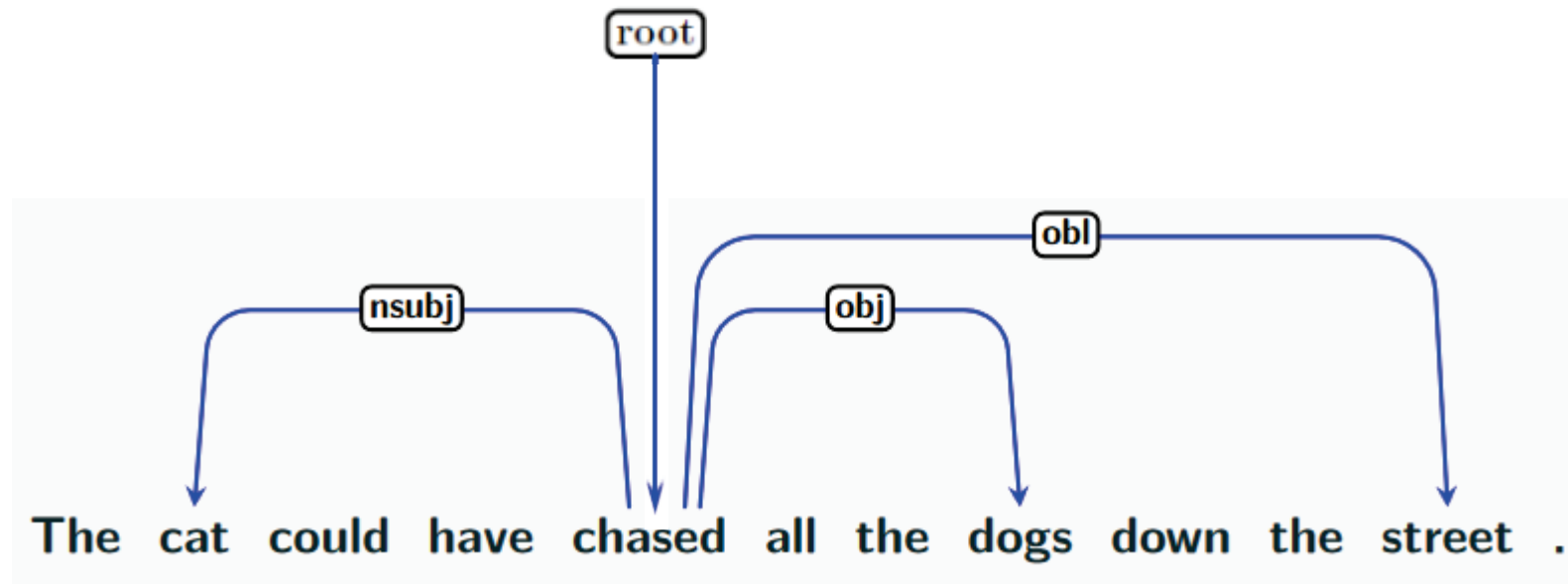


Swedish:



Syntactic annotation – main principles

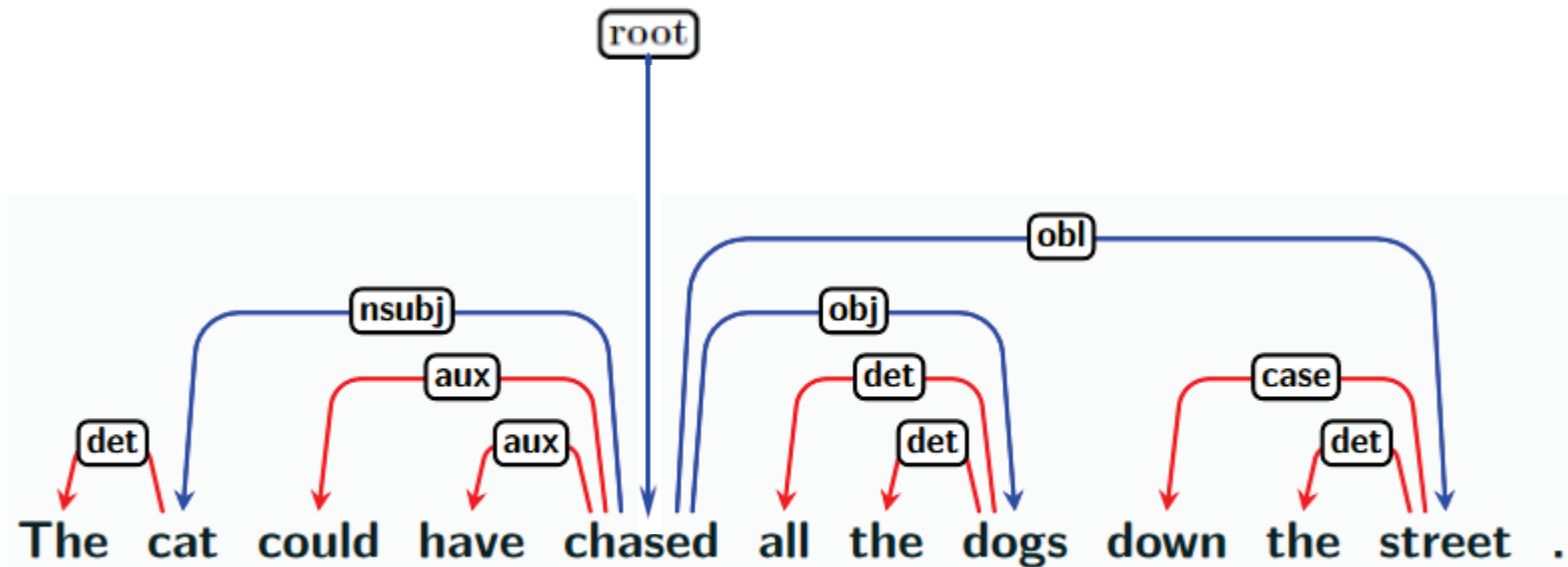
- content words are related by dependency relations





Syntactic annotation – overview

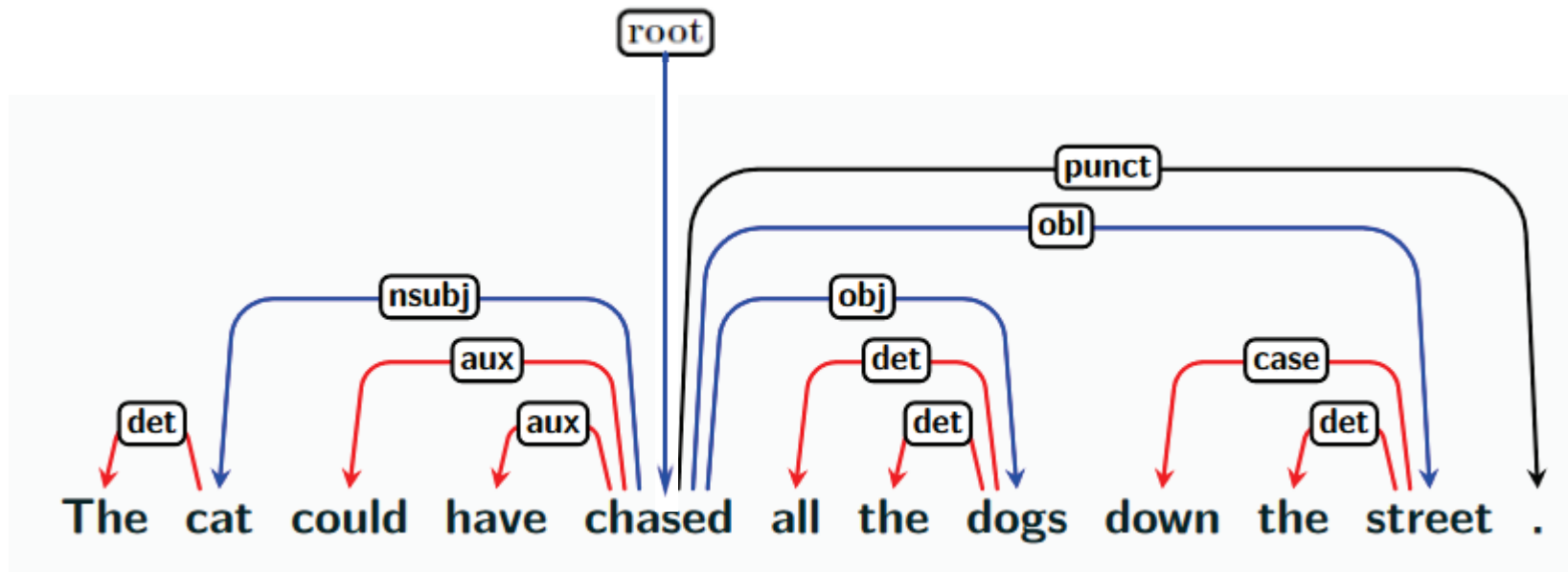
- content words are related by dependency relations
- function words attach to the content word they modify





Syntactic annotation – overview

- content words are related by dependency relations
- function words attach to the content word they modify
- punctuation attach to head of phrase or clause





Main Characteristics of UDs

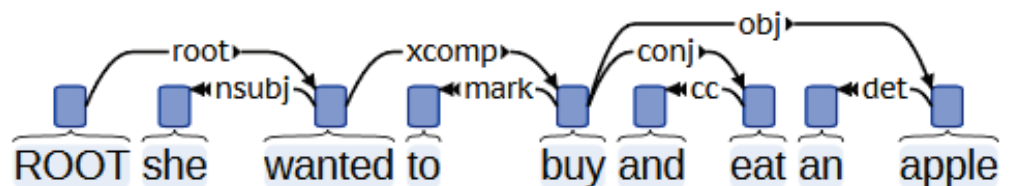
- *dependency annotation*
 - lexicalization
- *maximize parallelism* – but don't overdo it:
 - don't annotate the same thing in different ways
 - don't make different things look the same
 - don't annotate things that are not there
- universal taxonomy with possible language-specific features
 - languages select from a universal pool of categories
 - allow language-specific extensions

!!! NOT a new linguistic theory (but linguistically informed and relevant)



Syntactic annotation – main principles

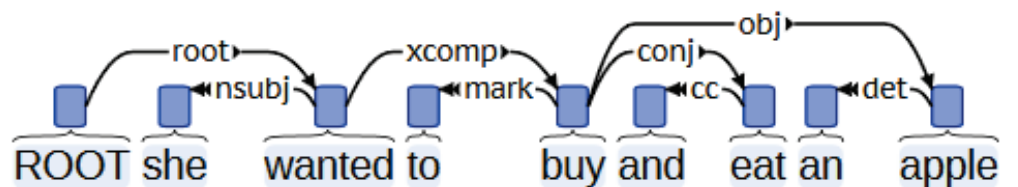
- basic dependency representation: rooted tree
 - (surface) syntax
 - a notional ROOT
 - all other words are dependent on another word in the sentence
 - obligatory for all UD treebanks



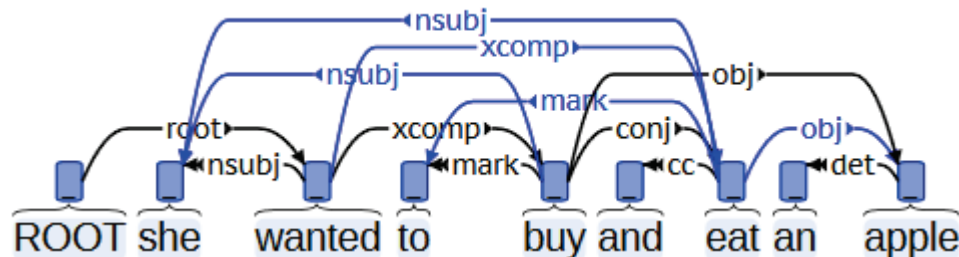


Syntactic annotation – main principles

- basic dependency representation: rooted tree
 - (surface) syntax
 - a notional ROOT
 - all other words are dependent on another word in the sentence
 - obligatory for all UD treebanks



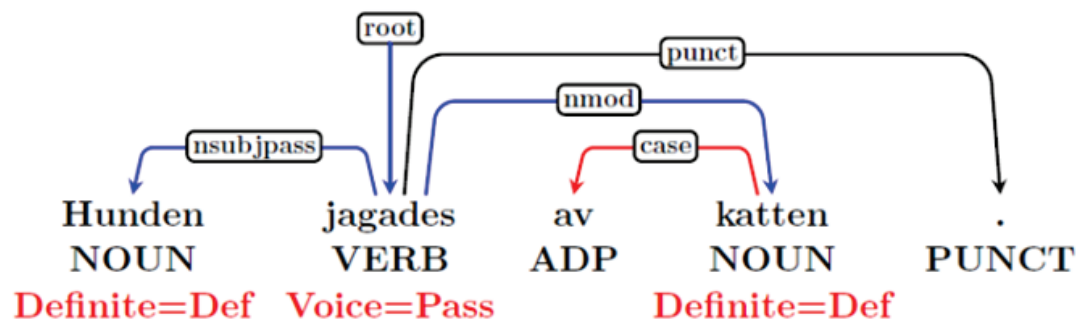
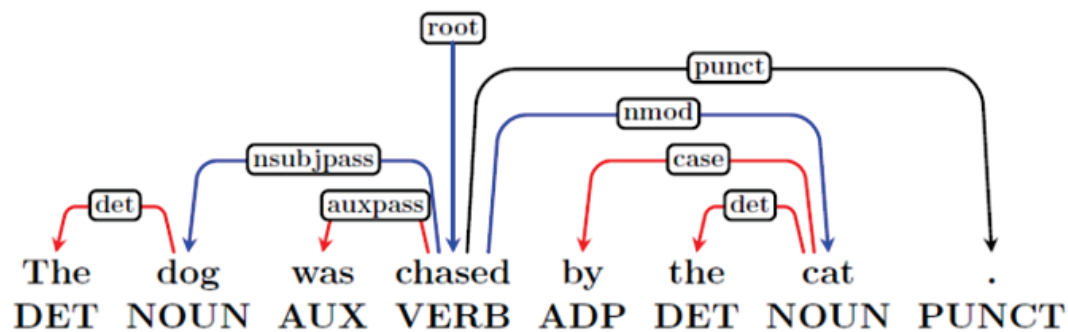
- enhanced dependency representation
 - in general not a tree
 - adds (change in a few cases) relations ... for semantic interpretation





Syntactic annotation – overview

- content words as heads
→ maximizes parallelism between languages
typed dependencies



Syntactic annotation – overview



- content words as heads
 - maximizes parallelism between languages*typed dependencies*

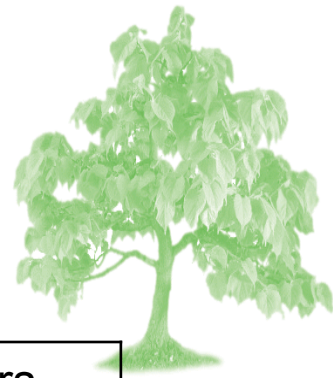
- relations between function w. and content words
 - operations modifying the grammatical category of content w.*functional relations* or *function word relations*
 - function words (normally) have no dependents

Overview of UD Syntactic Relations



- taxonomy of 37 universal grammatical relations
 - broadly attested in language typology
 - language specific subtypes may be added
 - (incl. functional relations)
- organizing principles
 - 3 types of structures: *nominals, clauses, modifiers*
 - core arguments / oblique modifiers
 - coordination
 - multiword expressions
 - function words
 - other relations

Nominals, Clauses, Modifiers



		nominal	clauses	modifiers
dependents of clausal predicates	core	<u>nsubj</u>	<u>csbj</u>	
		<u>obj</u>	<u>ccomp</u>	
		<u>iobj</u>	<u>xcomp</u>	
	non-core	<u>obl</u>	<u>advcl</u>	<u>advmod</u>
		<u>vocative</u>		<u>discourse</u>
		<u>expl</u>		
		<u>dislocated</u>		
nominal dependents		<u>nmod</u>	<u>acl</u>	<u>amod</u>
		<u>appos</u>		
		<u>nummod</u>		

Core Arguments / Oblique Modifiers



- **core arguments**
 - verbs usually only *agree* with core arguments
 - core args typically appear *as bare nominals* (= without pre/postposition)
 - certain cases (nominative, accusative, and absolutive) typically mark core args
 - core args in many languages occupy *special positions* in the clause
 - some syntactic phenomena are limited to core arguments in some languages.

~ more-or-less Obj in PDT
- **oblique modifiers**
 - oblique args may usually or always appear marked by *an adposition*

~ more-or-less Adv in PDT

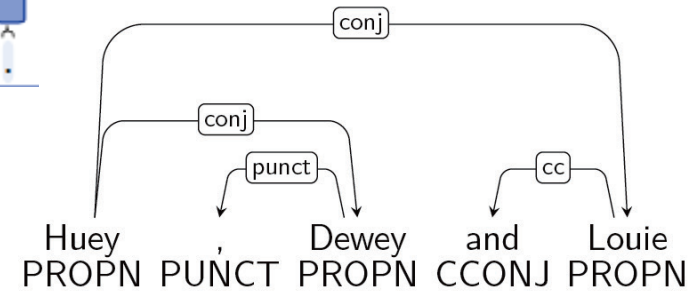
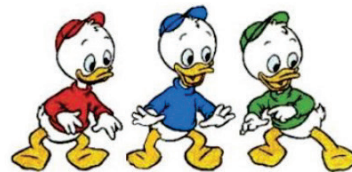
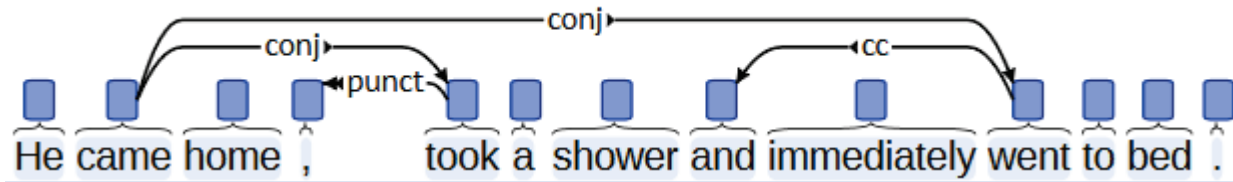
NOT argument / adjunct distinction !!

Coordination



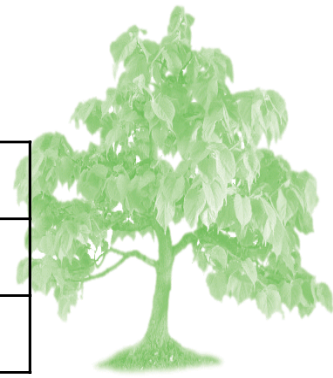
coordination	<u>conj</u>
	<u>cc</u>

- head ... the first conjunct
- all other conjuncts ... depend on the head via the conj relation
- coordinating conjunction and punctuation
 - ... attached to the immediately **following** conjunct
 - via the cc and punct relations
 - (different from version 1 where attached to the first one)

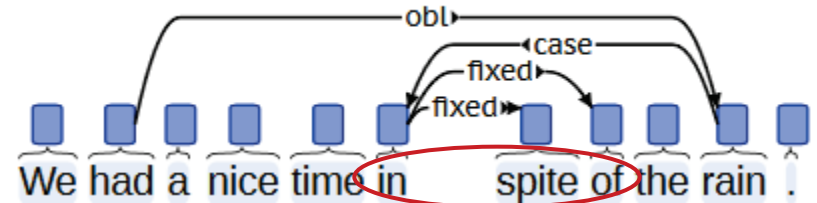


Multiword Expressions

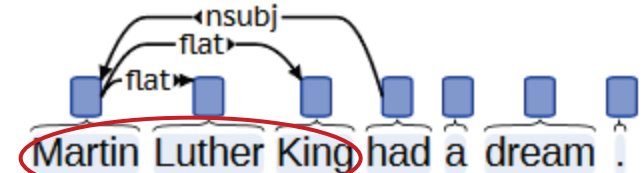
mwe	<u>fixed</u>
	<u>flat</u>
	<u>compound</u>



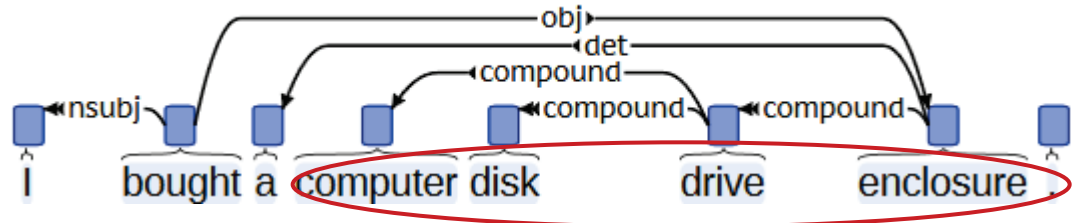
- fixed ... for fixed grammaticalized function-word MWEs
e.g., *in spite of*

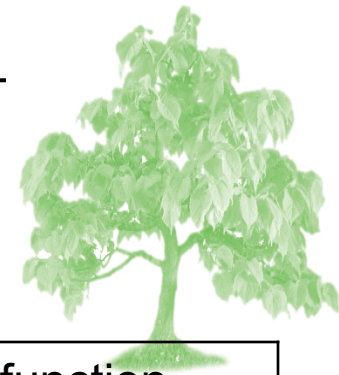


- flat ... for exocentric semi-fixed MWEs
e.g., *Barack Obama* (with no clear head)



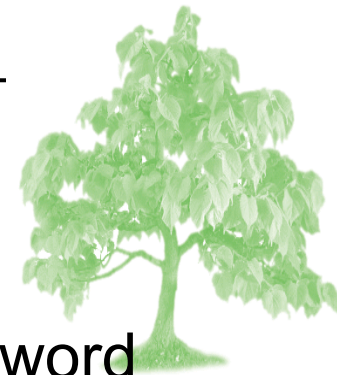
- compound ... for (headed or endocentric) compounds
i.e., with *clear inner structure*





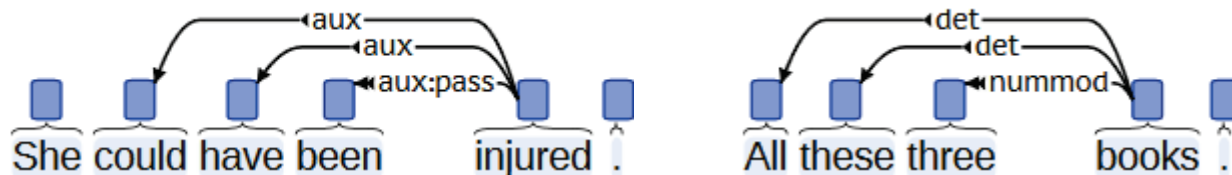
Functional Words

		nominal	clauses	modifiers	function words
dependents of clausal predicates	core	<u>nsubj</u>	<u>csubj</u>		
		<u>obj</u>	<u>ccomp</u>		
		<u>iobj</u>	<u>xcomp</u>		
	non-core	<u>obl</u>	<u>advcl</u>	<u>advmod</u>	<u>aux</u>
		<u>vocative</u>		<u>discourse</u>	<u>cop</u>
		<u>expl</u>			<u>mark</u>
		<u>dislocated</u>			
nominal dependents		<u>nmod</u>	<u>acl</u>	<u>amod</u>	<u>det</u>
		<u>appos</u>			<u>clf</u>
		<u>nummod</u>			<u>case</u>



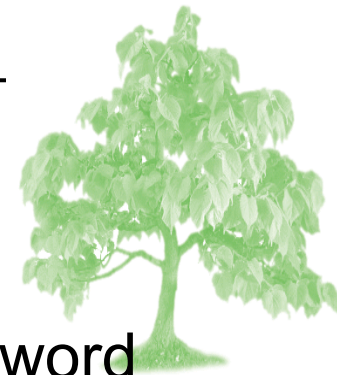
Status of Function Words

- *multiple function* words related to the same content word → as siblings



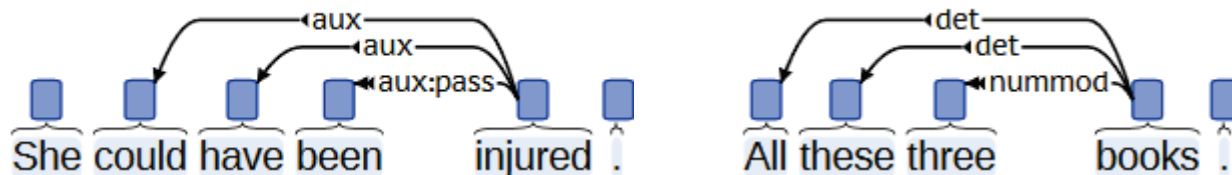
- *copula* as a function word *but not for clauses as non-verbal predicates*



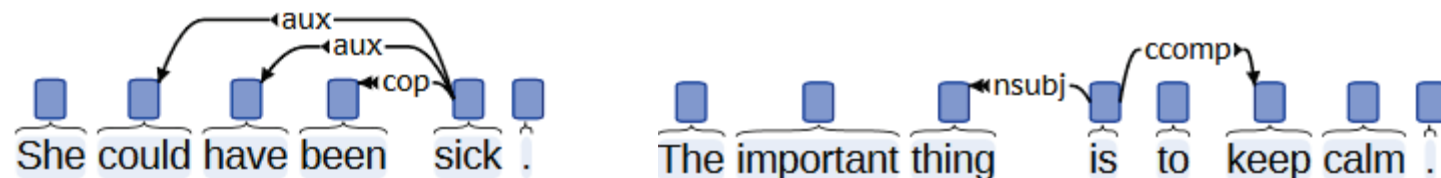


Status of Function Words

- *multiple function* words related to the same content word → as siblings



- *copula* as a function word *but not for clauses as non-verbal predicates*

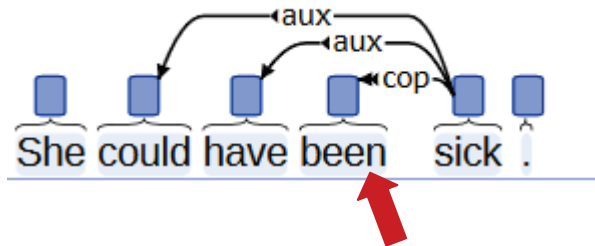


- 4 exceptions:
 - multiword function words
 - coordinated function words
 - function word modifiers
 - promotion by head elision

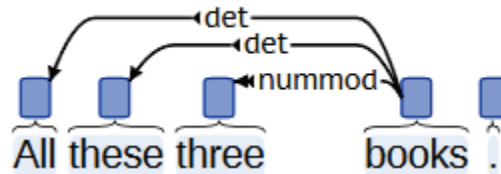
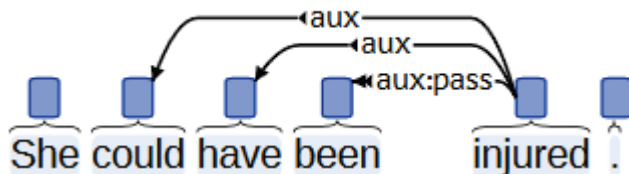


Status of Function Words

- *copula* as a function word

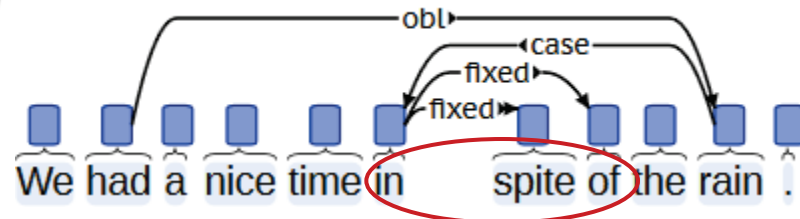


- *multiple function* words related to the same content word → as siblings



- 4 exceptions:

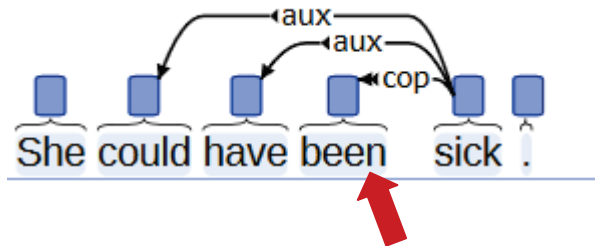
- *multiword function words* ... relation *fixed*
- coordinated function words
- function word modifiers
- promotion by head elision



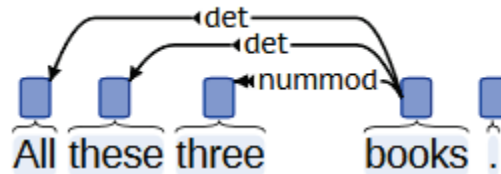
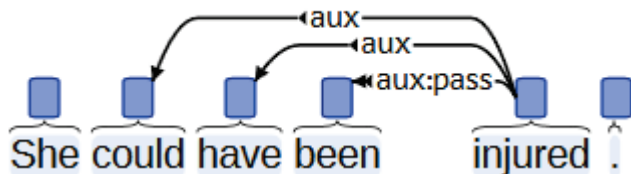


Status of Function Words

- *copula* as a function word

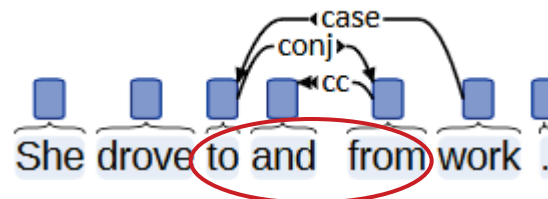


- *multiple function* words related to the same content word → as siblings



- 4 exceptions:

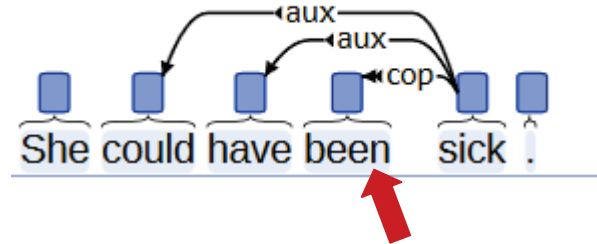
- multiword function words
- *coordinated function words*
- function word modifiers
- promotion by head elision



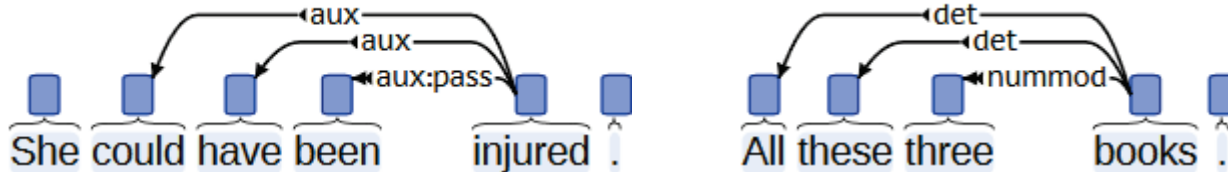


Status of Function Words

- *copula* as a function word



- *multiple function* words related to the same content word as siblings



- 4 exceptions:

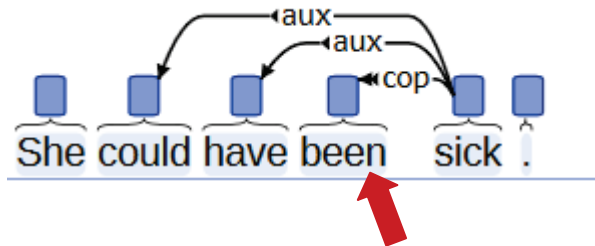
- multiword function words
- coordinated function words
- *function word modifiers* ... e.g., *modified determiners*
- promotion by head elision



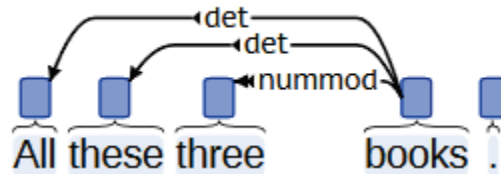
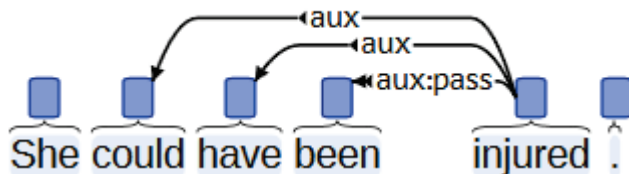


Status of Function Words

- *copula* as a function word

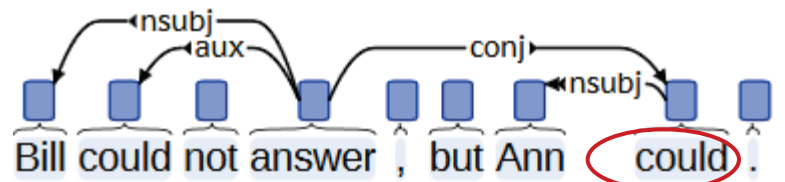


- *multiple function* words related to the same content word as siblings

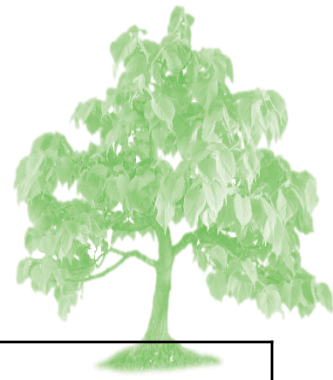


- 4 exceptions:

- multiword function words
- coordinated function words
- function word modifiers
- promotion by *head elision*



Others



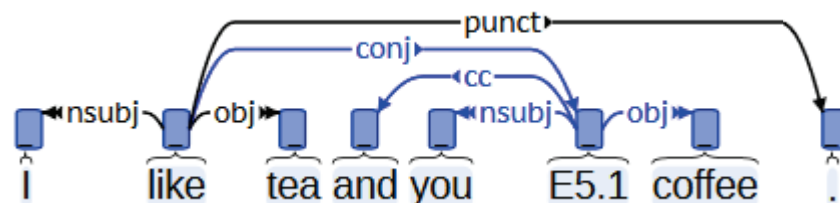
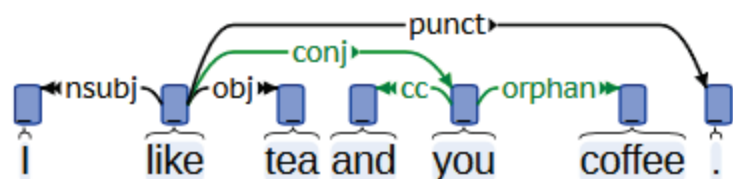
loose	<u>list</u>	tel. numbers, addresses, ... (without syntactic structure)
	<u>parataxis</u>	loosely linked clauses of same rank
special	<u>orphan</u>	orphans in ellipsis linked together
	<u>goeswith</u>	for badly segmented words
	<u>reparandum</u>	disfluency linked to (speech) repair
other	<u>punct</u>	
	<u>root</u>	syntactically independent element of clause/phrase
	<u>dep</u>	unspecified dependency



Enhanced Dependencies

To capture more fine information:

- null nodes for elided predicates



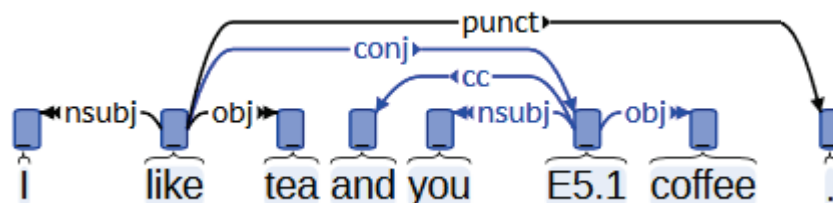
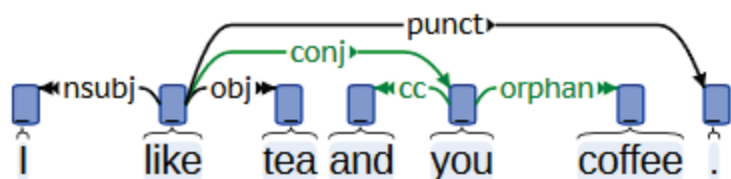
The enhanced graph is not necessarily a supergraph of the basic tree.



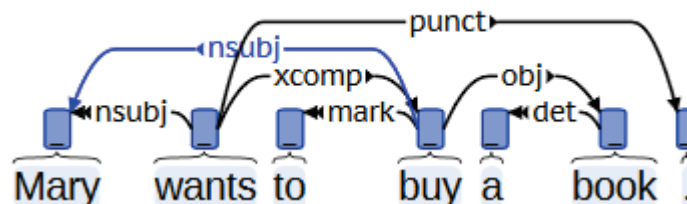
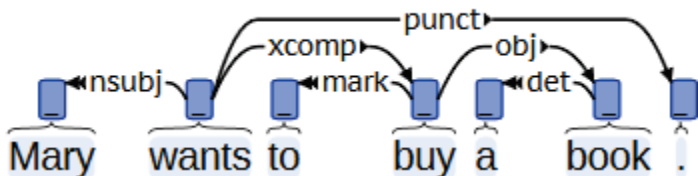
Enhanced Dependencies

To capture more fine information:

- null nodes for elided predicates



- additional subject relations for control and raising constructions



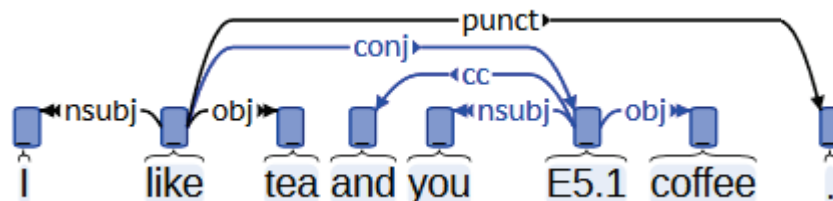
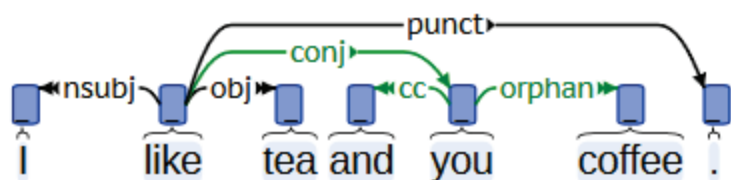
The enhanced graph is not necessarily a supergraph of the basic tree.



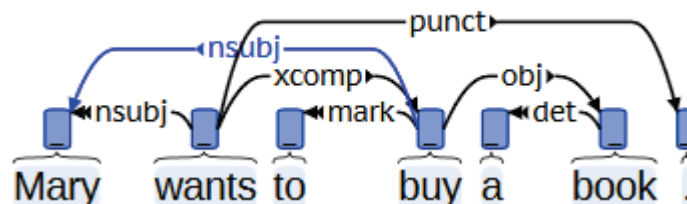
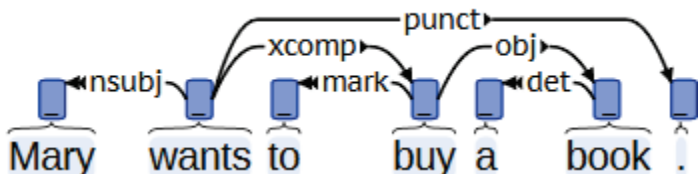
Enhanced Dependencies

To capture more fine information:

- null nodes for elided predicates



- additional subject relations for control and raising constructions



- propagation of conjuncts (~ effective children/parents in PDT)
- coreference in relative clause constructions
- modifier labels that contain the preposition or other case-marking information

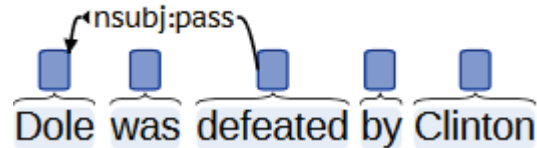
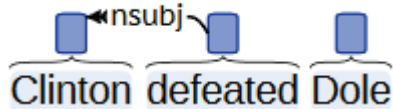
The enhanced graph is not necessarily a supergraph of the basic tree.



"Language-specific" relations

- relations annotated for specific languages (not necessary specific to the language!!)
- subtypes of universal relations

e.g., nsubj:pass



CONLL-U Format



- Revised and extended version of CoNLL-X format

The cat drinks milk.

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	Le	le	DET	-	-	2	det	-	-
2	chat	chat	NOUN	-	-	3	nsubj	-	-
3	boit	boire	VERB	-	-	0	root	-	-
4-5	du	-	-	-	-	-	-	-	-
4	de	de	ADP	-	-	6	case	-	-
5	le	le	DET	-	-	6	det	-	-
6	lait	lait	NOUN	-	-	3	obj	-	SpaceAfter=No
7	.	.	PUNCT	-	-	3	punct	-	-

CONLL-U Format



- Revised and extended version of CoNLL-X format

Lemma or stem of word form

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	Le	le	DET	-	-	2	det	-	-
2	chat	chat	NOUN	-	-	3	nsubj	-	-
3	boit	boire	VERB	-	-	0	root	-	-
4-5	du	-	-	-	-	-	-	-	-
4	de	de	ADP	-	-	6	case	-	-
5	le	le	DET	-	-	6	det	-	-
6	lait	lait	NOUN	-	-	3	obj	-	SpaceAfter=No
7	.	.	PUNCT	-	-	3	punct	-	-



CONLL-U Format

- Revised and extended version of CoNLL-X format

Universal part-of-speech tag

Language-specific part-of-speech tag

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	Le	le	DET	-	-	2	det	-	-
2	chat	chat	NOUN	-	-	3	nsubj	-	-
3	boit	boire	VERB	-	-	0	root	-	-
4-5	du	-	-	-	-	-	-	-	-
4	de	de	ADP	-	-	6	case	-	-
5	le	le	DET	-	-	6	det	-	-
6	lait	lait	NOUN	-	-	3	obj	-	SpaceAfter=No
7	.	.	PUNCT	-	-	3	punct	-	-



CONLL-U Format

- Revised and extended version of CoNLL-X format

List of morphological features
(universal as well as lang. specific)

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	Le	le	DET	-	-	2	det	-	-
2	chat	chat	NOUN	-	-	3	nsubj	-	-
3	boit	boire	VERB	-	-	0	root	-	-
4-5	du	-	-	-	-	-	-	-	-
4	de	de	ADP	-	-	6	case	-	-
5	le	le	DET	-	-	6	det	-	-
6	lait	lait	NOUN	-	-	3	obj	-	SpaceAfter=No
7	.	.	PUNCT	-	-	3	punct	-	-

CONLL-U Format



- Revised and extended version of CoNLL-X format

Head of the current word - ID or zero

Universal dependency relation to the HEAD/root

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	Le	le	DET	-	-	2	det	-	-
2	chat	chat	NOUN	-	-	3	nsubj	-	-
3	boit	boire	VERB	-	-	0	root	-	-
4-5	du	-	-	-	-	-	-	-	-
4	de	de	ADP	-	-	6	case	-	-
5	le	le	DET	-	-	6	det	-	-
6	lait	lait	NOUN	-	-	3	obj	-	SpaceAfter=No
7	.	.	PUNCT	-	-	3	punct	-	-

CONLL-U Format



- Revised and extended version of CoNLL-X format

Enhanced dependency graph
(head-deprel pairs)

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	Le	le	DET	-	-	2	det	-	-
2	chat	chat	NOUN	-	-	3	nsubj	-	-
3	boit	boire	VERB	-	-	0	root	-	-
4-5	du	-	-	-	-	-	-	-	-
4	de	de	ADP	-	-	6	case	-	-
5	le	le	DET	-	-	6	det	-	-
6	lait	lait	NOUN	-	-	3	obj	-	SpaceAfter=No
7	.	.	PUNCT	-	-	3	punct	-	-

Universal Dependencies (UDs)



idea:

- cross-linguistically consistent treebank annotation
- annotation for different languages as similar as possible
- support multilingual NLP (parser development, cross-lingual learning, ...) and linguistic research

people:

- J. Nivre, D. Zeman, F. Ginter, (more than 160 contributors)
- open community effort – anyone can contribute!

main facts:

- from 2014
- version 2 in December 2016 (guidelines) and March 2017 (treebanks)
- next to 80 treebanks for 49 languages

<http://universaldependencies.org/>