UPPSALA
UNIVERSITET

UNIVERSITÉ
DE GENÈVE

Universal Dependencies

A Framework for Cross-Linguistically Consistent Grammatical Annotation

Joakim Nivre

Introduction

Introduction

Linguistic annotation is tremendously useful

- Computational linguistics use it for machine learning and evaluation
- Corpus linguistics use it for studying complex linguistic phenomena

Introduction

Linguistic annotation is tremendously useful

- Computational linguistics use it for machine learning and evaluation
- Corpus linguistics use it for studying complex linguistic phenomena

Linguistic annotation is available for many languages

- Multilingual evaluation to test generality of computational models
- Cross-lingual learning to support under-resourced languages
- Empirically grounded linguistic typology and comparative linguistics

Introduction

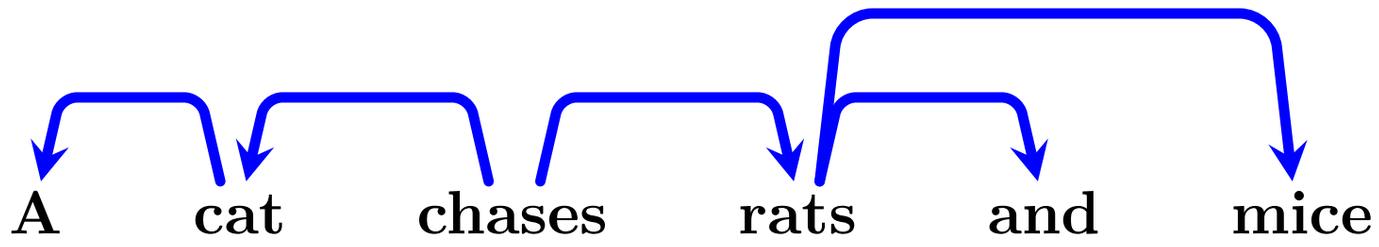
Linguistic annotation is tremendously useful

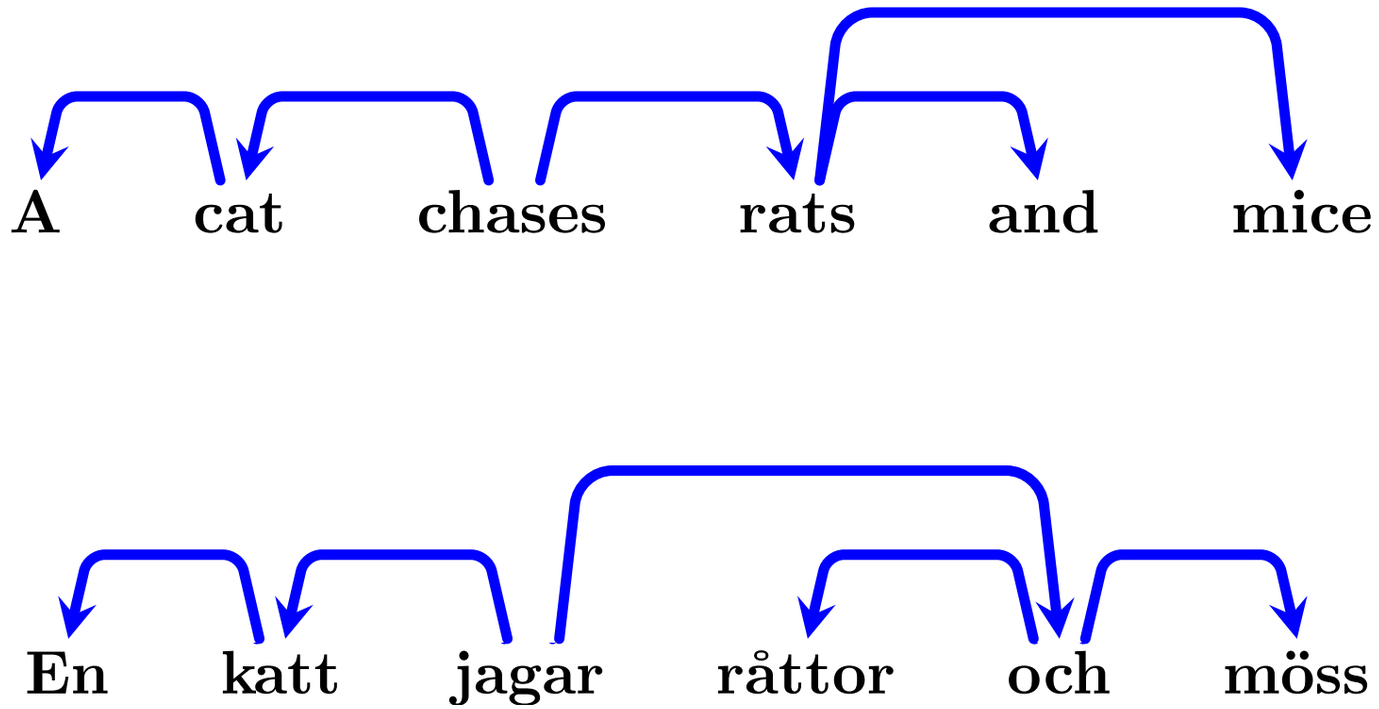
- Computational linguistics use it for machine learning and evaluation
- Corpus linguistics use it for studying complex linguistic phenomena

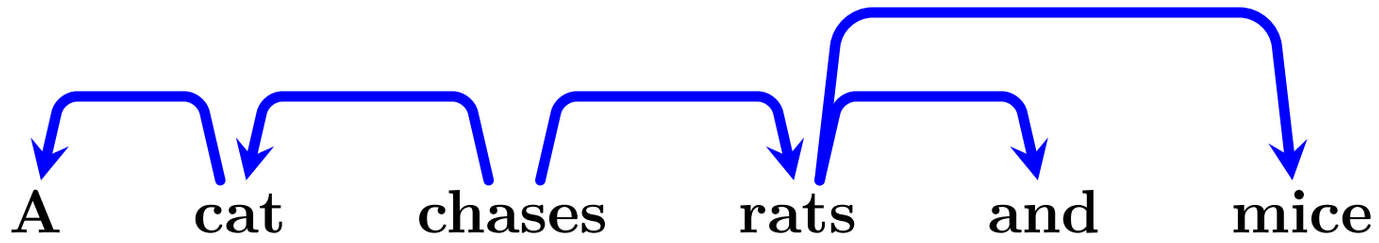
Linguistic annotation is available for many languages

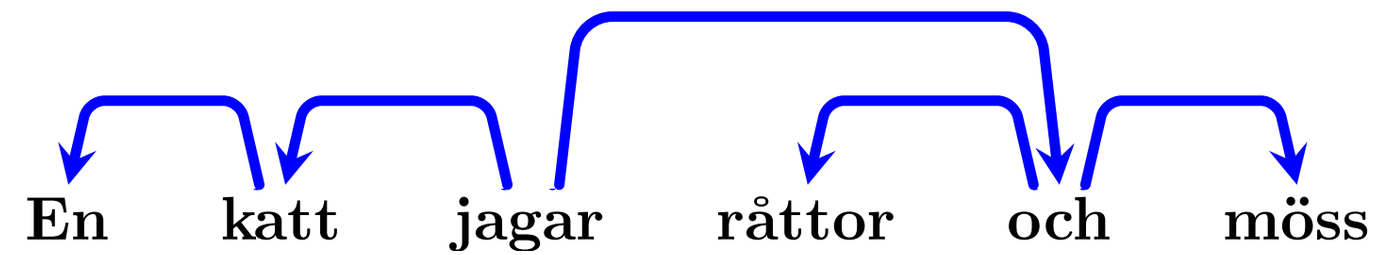
- Multilingual evaluation to test generality of computational models
- Cross-lingual learning to support under-resourced languages
- Empirically grounded linguistic typology and comparative linguistics

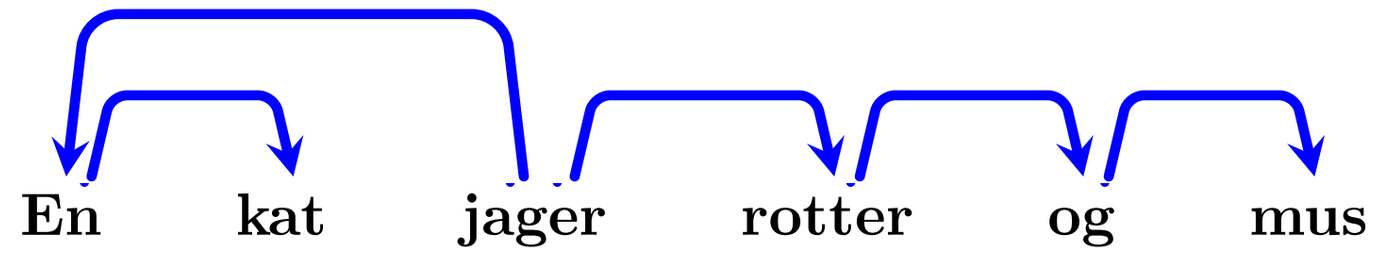
Linguistic annotation guidelines vary across languages 

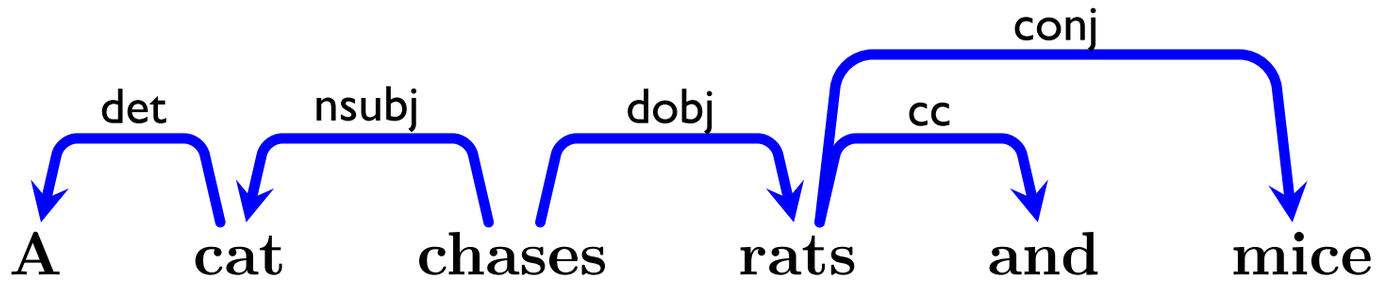

A
cat
chases
rats
and
mice

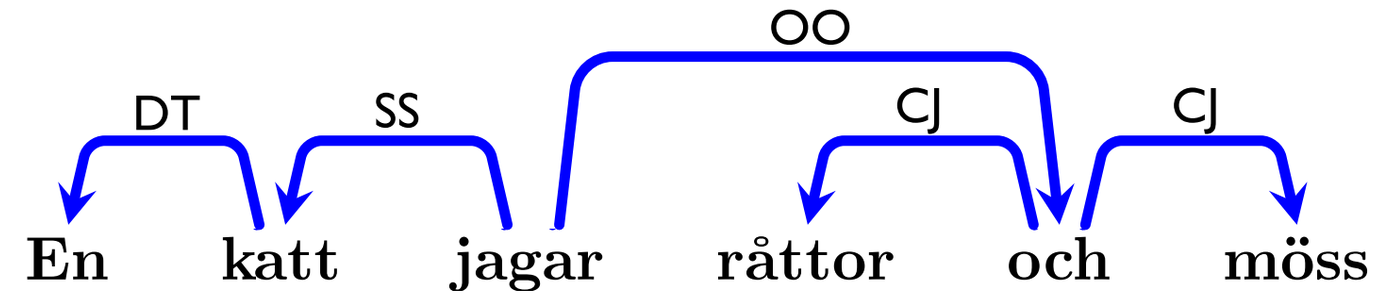

A
cat
chases
rats
and
mice
En
katt
jagar
råttor
och
möss

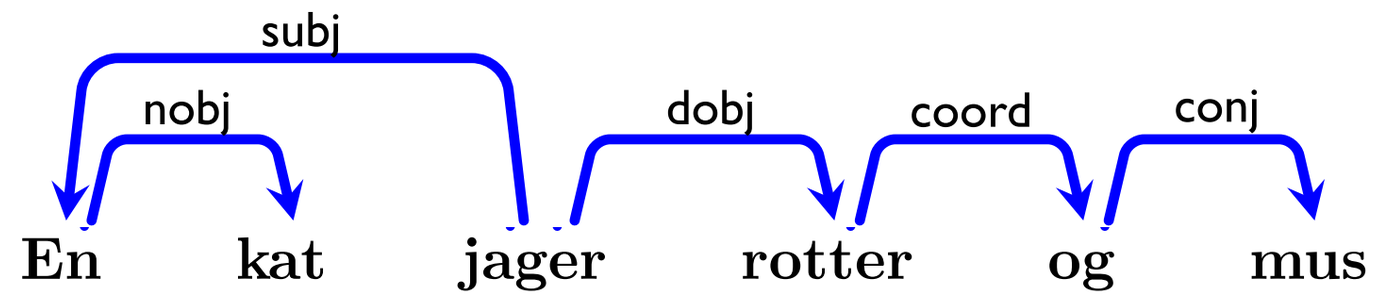

A
cat
chases
rats
and
mice

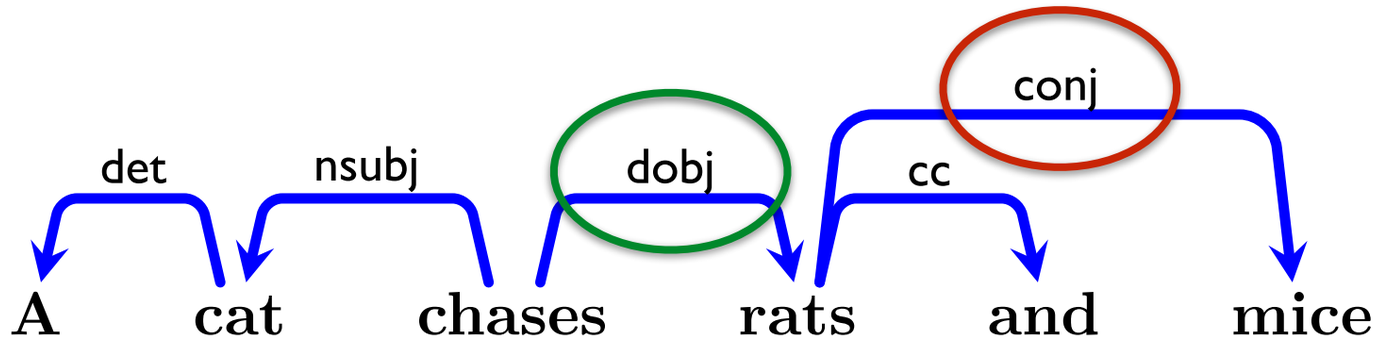

En
katt
jagar
råttor
och
möss

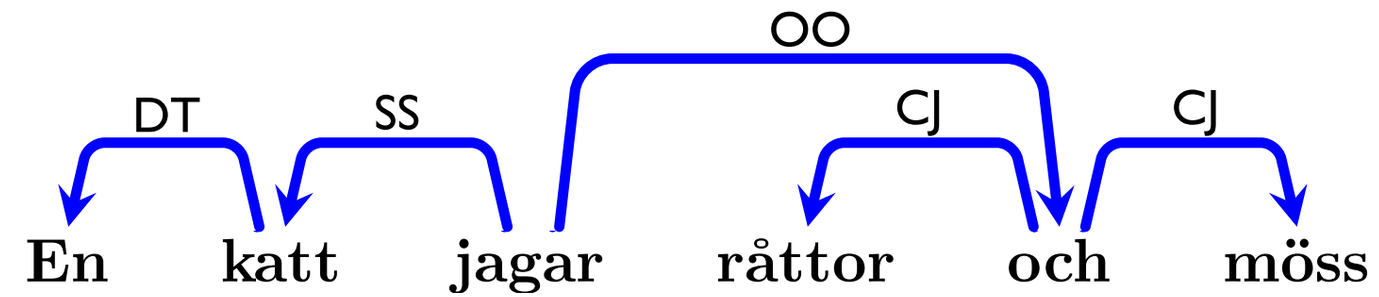

En
kat
jager
rotter
og
mus

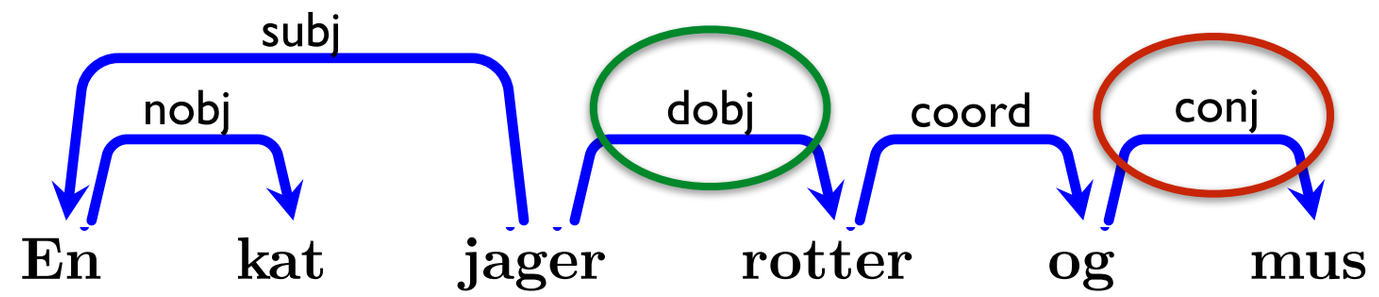

conj
det
nsubj
dobj
cc
A
cat
chases
rats
and
mice


OO
DT
SS
CJ
CJ
En
katt
jagar
råttor
och
möss


subj
nobj
dobj
coord
conj
En
kat
jager
rotter
og
mus


conj
det
nsubj
dobj
cc
A
cat
chases
rats
and
mice


OO
DT
SS
CJ
CJ
En
katt
jagar
råttor
och
möss


subj
nobj
dobj
coord
conj
En
kat
jager
rotter
og
mus

Why is this a problem?

Why is this a problem?

- Hard to compare empirical results across languages

Why is this a problem?

- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure learning

Why is this a problem?

- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure learning
- Hard to evaluate cross-lingual learning

Why is this a problem?

- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure learning
- Hard to evaluate cross-lingual learning
- Hard to build and maintain multilingual systems

Why is this a problem?

- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure learning
- Hard to evaluate cross-lingual learning
- Hard to build and maintain multilingual systems
- Hard to make comparative linguistic studies

Why is this a problem?

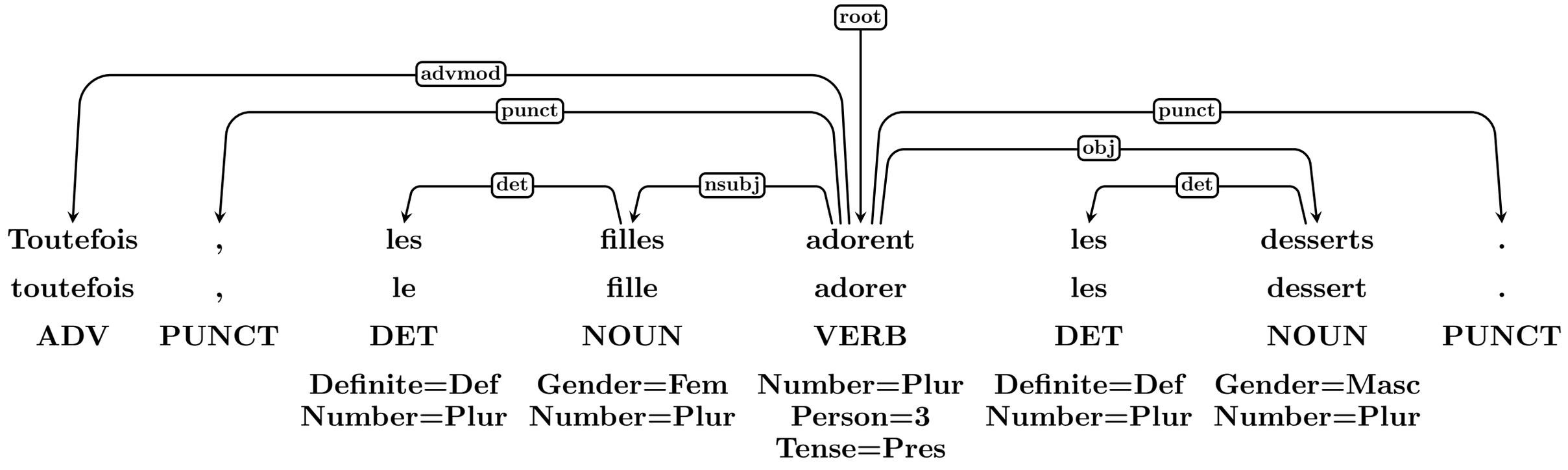
- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure learning
- Hard to evaluate cross-lingual learning
- Hard to build and maintain multilingual systems
- Hard to make comparative linguistic studies
- Hard to validate linguistic typology

Why is this a problem?

- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure learning
- Hard to evaluate cross-lingual learning
- Hard to build and maintain multilingual systems
- Hard to make comparative linguistic studies
- Hard to validate linguistic typology
- Hard to make progress towards a universal parser

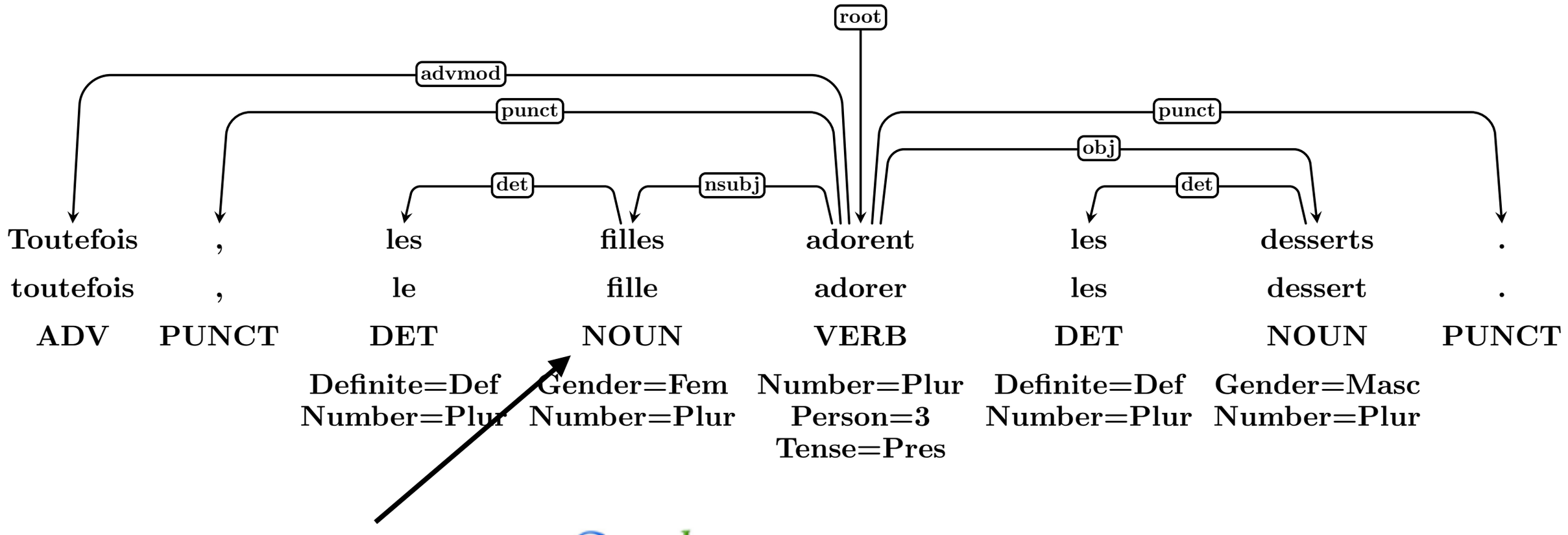
Universal Dependencies

http://universaldependencies.org



Universal Dependencies

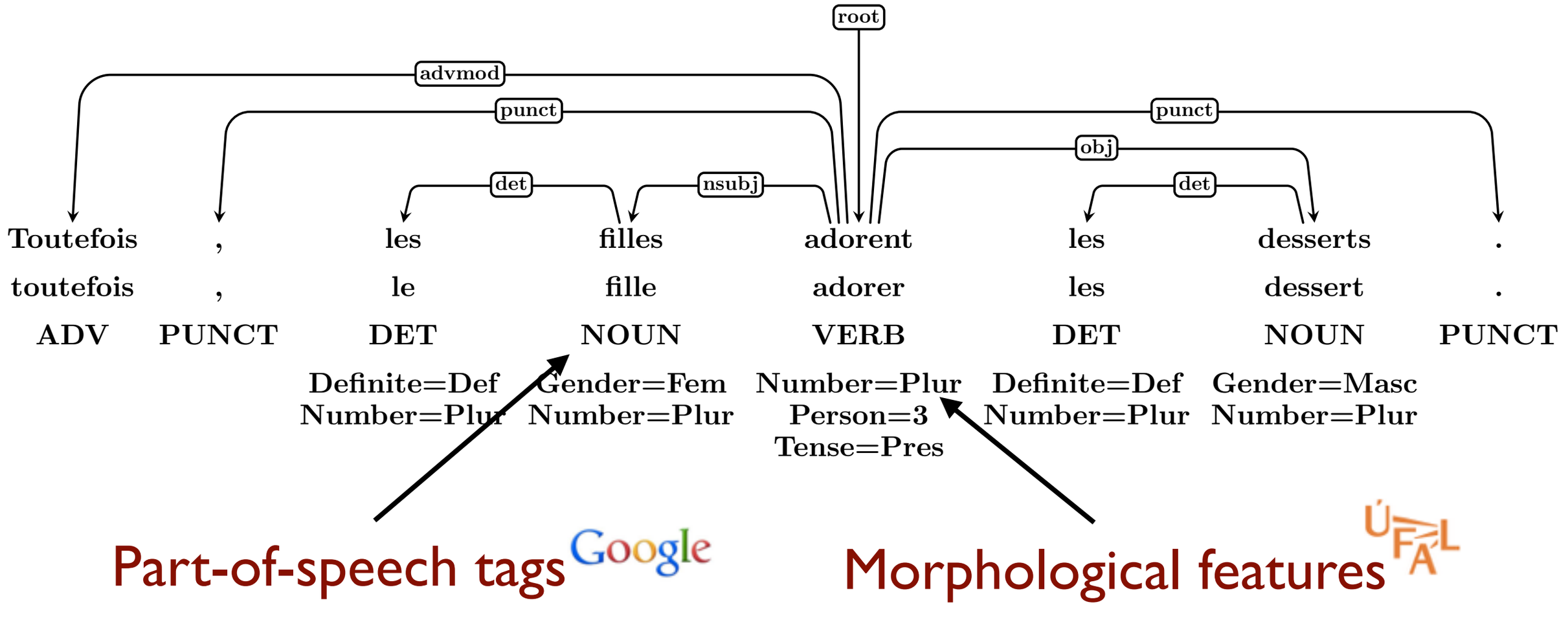
http://universaldependencies.org



Part-of-speech tags Google

Universal Dependencies

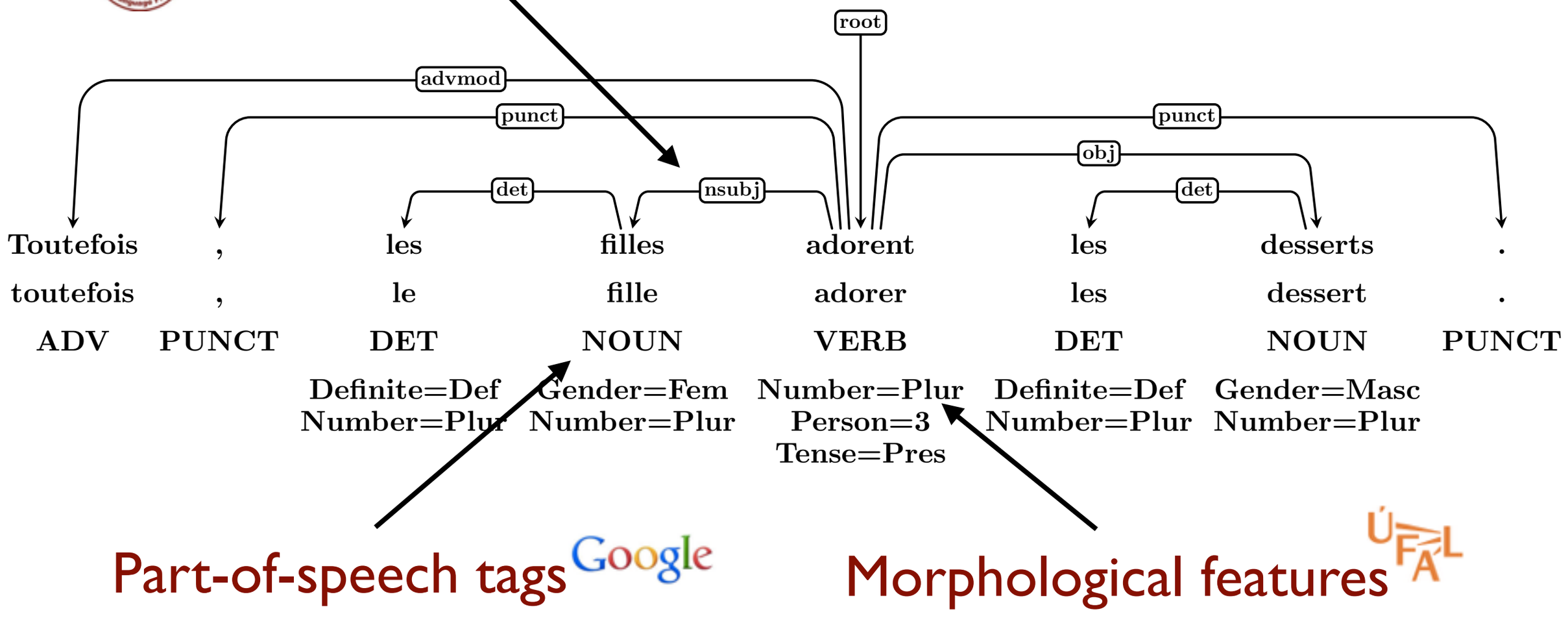
http://universaldependencies.org



Universal Dependencies

http://universaldependencies.org

Dependency relations



Part-of-speech tags Google

Morphological features ÚFAL

Universal Dependencies

http://universaldependencies.org

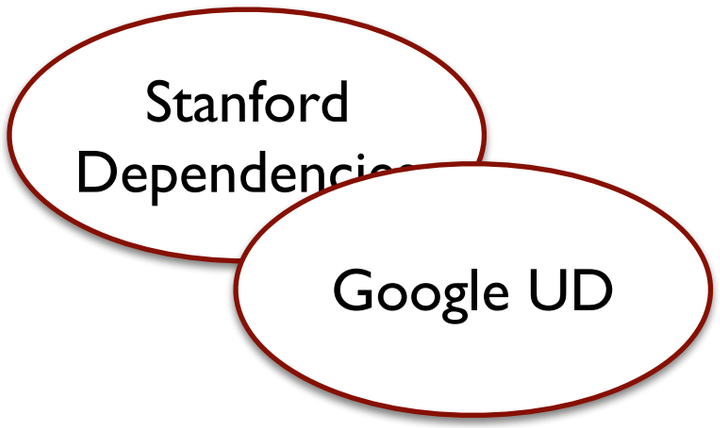
Universal Dependencies

http://universaldependencies.org

Stanford Dependencies

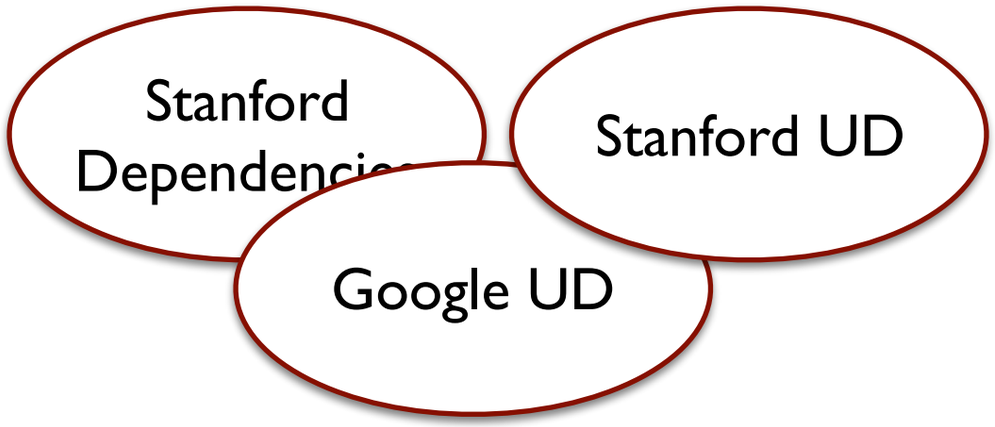
Universal Dependencies

http://universaldependencies.org



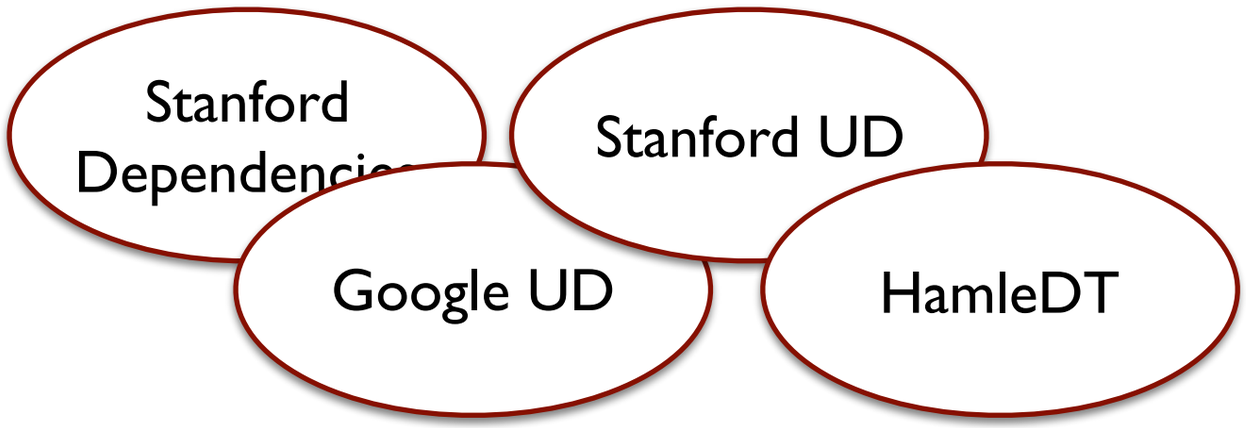
Universal Dependencies

http://universaldependencies.org



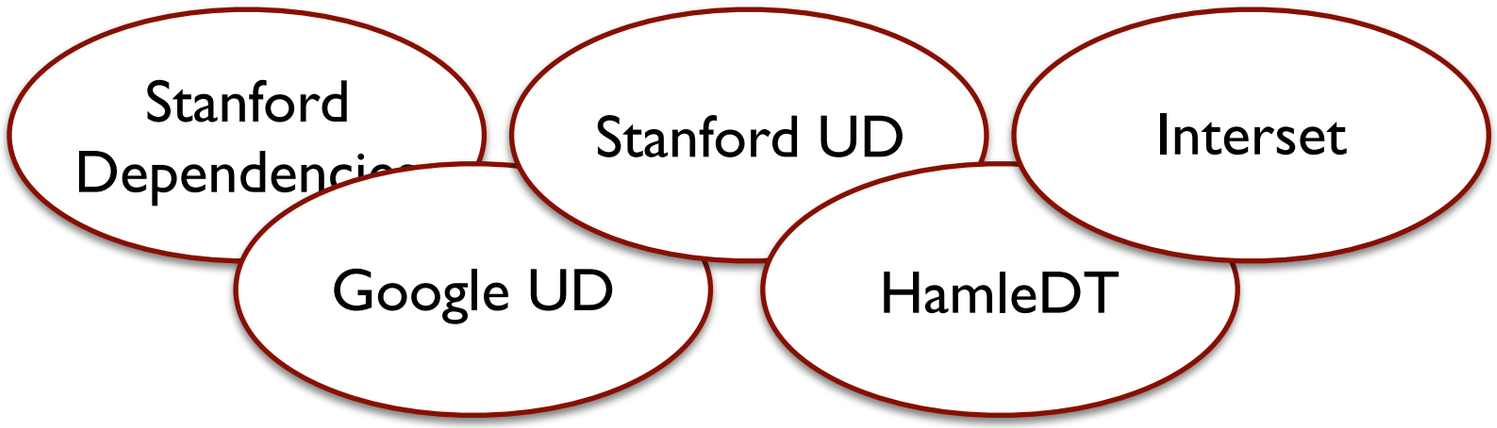
Universal Dependencies

http://universaldependencies.org



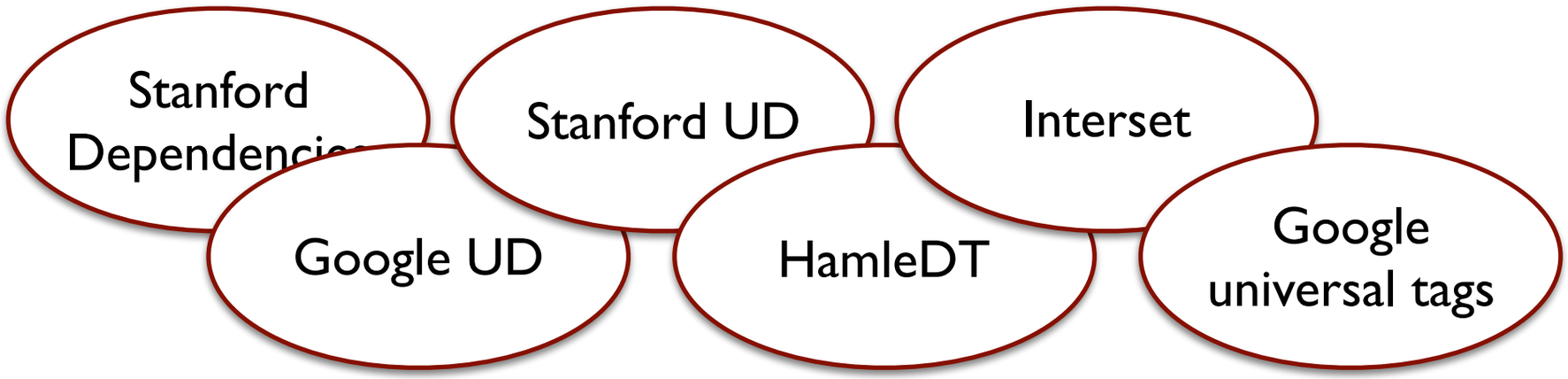
Universal Dependencies

http://universaldependencies.org



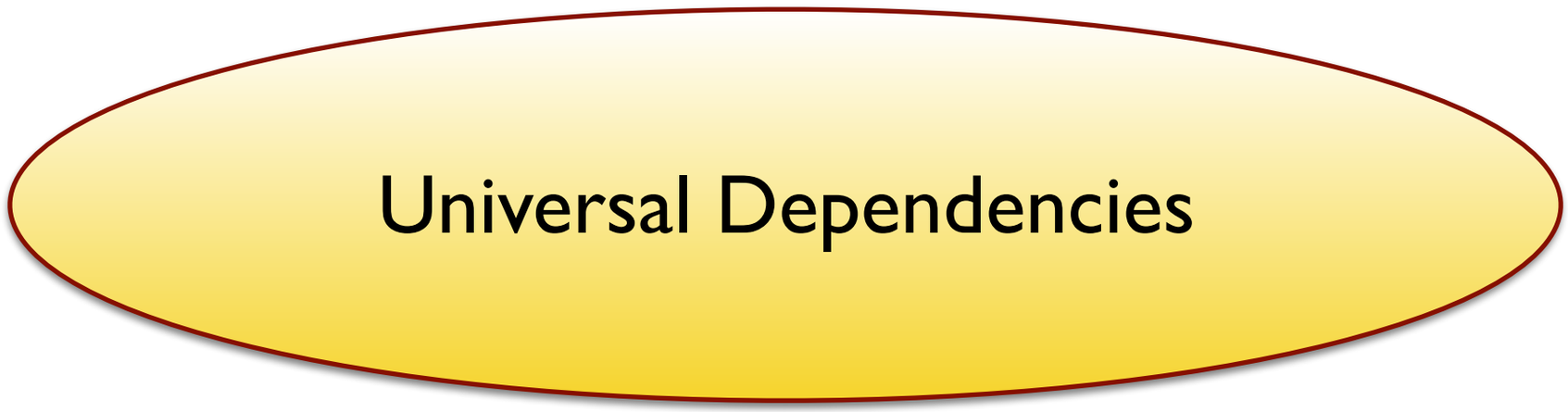
Universal Dependencies

http://universaldependencies.org



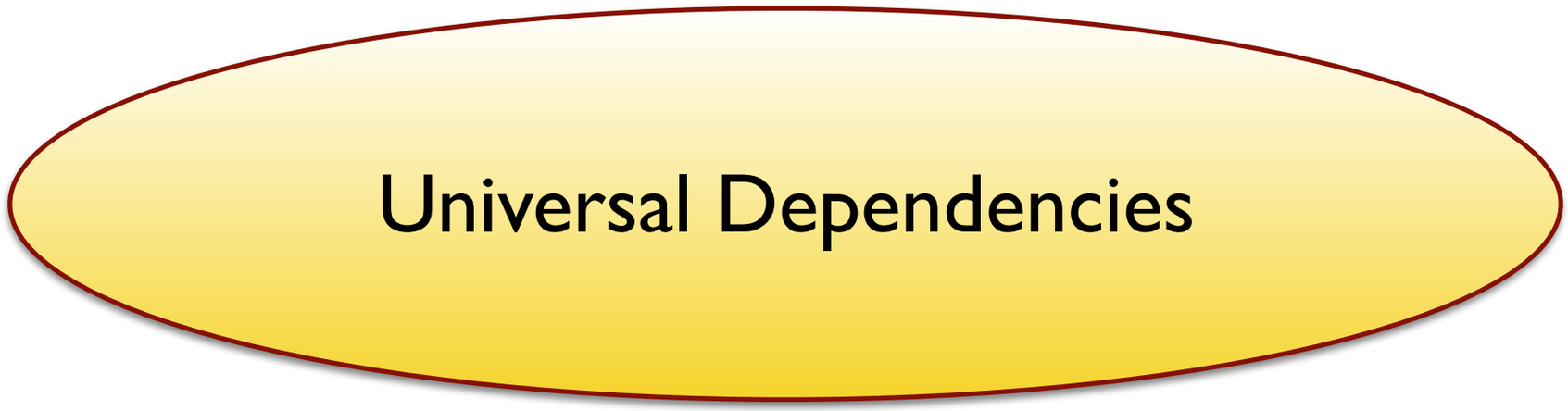
Universal Dependencies

http://universaldependencies.org



Universal Dependencies

http://universaldependencies.org



Milestones:

- Kick-off meeting at EACL in Gothenburg, April 2014
- Guidelines v1, October 2014
- Treebank releases every 6 months (v1.0–v1.4)
- Guidelines v2, December 2016
- Treebank release v2.0, March 2017 New!

Open community effort – anyone can contribute!

UD Treebanks

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

March 1, 2017:

- 50 languages
- 70 treebanks
- 162 contributors
- 8000+ downloads

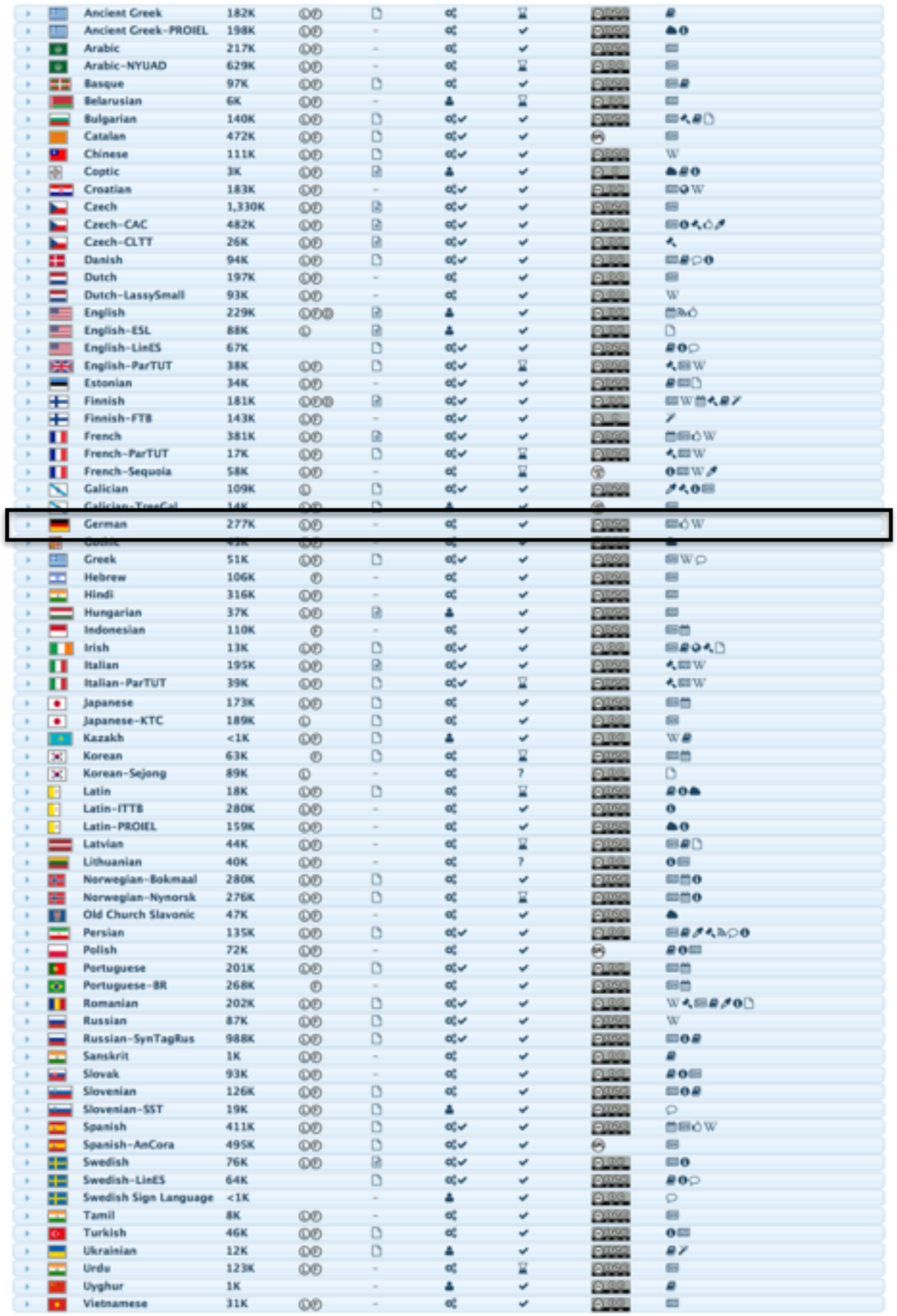
UD Treebanks

Ancient Greek	182K	UFP					
Ancient Greek-PROIEL	198K	UFP	-		✓		
Arabic	217K	UFP	-		✓		
Arabic-NYUAD	629K	UFP	-				
Basque	97K	UFP			✓		
Belarusian	6K	UFP	-				
Bulgarian	140K	UFP			✓		
Catalan	472K	UFP			✓		
Chinese	111K	UFP			✓		W
Coptic	3K	UFP			✓		
Croatian	183K	UFP	-		✓		W
Czech	1,330K	UFP			✓		
Czech-CAC	482K	UFP			✓		
Czech-CLTT	26K	UFP			✓		
Danish	94K	UFP			✓		
Dutch	197K	UFP	-		✓		
Dutch-LassySmall	93K	UFP	-		✓		W
English	229K	UFPB			✓		
English-ESL	88K	U			✓		
English-LinES	67K				✓		
English-ParTUT	38K	UFP			✓		W
Estonian	34K	UFP	-		✓		
Finnish	181K	UFPB			✓		W
Finnish-FTB	143K	UFP	-		✓		
French	381K	UFP			✓		W
French-ParTUT	17K	UFP			✓		W
French-Sequoia	58K	UFP	-				W
Galician	139K	U			✓		
Galician-TreeGal	14K	UFP			✓		
German	277K	UFP	-		✓		W
Gothic	45K	UFP	-		✓		
Greek	51K	UFP			✓		W
Hebrew	106K	F	-		✓		
Hindi	316K	UFP	-		✓		
Hungarian	37K	UFP			✓		
Indonesian	110K	F	-		✓		
Irish	13K	UFP			✓		
Italian	195K	UFP			✓		W
Italian-ParTUT	39K	UFP			✓		W
Japanese	173K	UFP			✓		
Japanese-KTC	189K	U			✓		
Kazakh	<1K	UFP			✓		W
Korean	63K	F			✓		
Korean-Sejong	89K	U	-		?		
Latin	18K	UFP			✓		
Latin-ITTB	280K	UFP	-		✓		
Latin-PROIEL	159K	UFP	-		✓		
Latvian	44K	UFP	-		✓		
Lithuanian	40K	UFP	-		?		
Norwegian-Bokmaal	280K	UFP			✓		
Norwegian-Nynorsk	276K	UFP			✓		
Old Church Slavonic	47K	UFP	-		✓		
Persian	135K	UFP			✓		
Polish	72K	UFP	-		✓		
Portuguese	201K	UFP			✓		
Portuguese-BR	268K	F	-		✓		
Romanian	202K	UFP			✓		W
Russian	87K	UFP			✓		W
Russian-SynTagRus	988K	UFP			✓		
Sanskrit	1K	UFP	-		✓		
Slovak	93K	UFP	-		✓		
Slovenian	126K	UFP			✓		
Slovenian-SST	19K	UFP			✓		
Spanish	411K	UFP			✓		W
Spanish-AnCora	495K	UFP			✓		
Swedish	76K	UFP			✓		
Swedish-LinES	64K				✓		
Swedish Sign Language	<1K		-		✓		
Tamil	8K	UFP	-		✓		
Turkish	46K	UFP			✓		
Ukrainian	12K	UFP			✓		
Urdu	123K	UFP	-				
Uyghur	1K		-		✓		
Vietnamese	31K	UFP	-		✓		

March 1, 2017:

- 50 languages
- 70 treebanks
- 162 contributors
- 8000+ downloads

UD Treebanks



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

March 1, 2017:

- 50 languages
- 70 treebanks
- 162 contributors
- 8000+ downloads

UD Treebanks

Ancient Greek	182K	UFP					
Ancient Greek-PROIEL	198K	UFP	-		✓		
Arabic	217K	UFP	-		✓		
Arabic-NYUAD	629K	UFP	-				
Basque	97K	UFP			✓		
Belarusian	6K	UFP	-				
Bulgarian	140K	UFP			✓		
Catalan	472K	UFP			✓		
Chinese	111K	UFP			✓		W
Coptic	3K	UFP			✓		
Croatian	183K	UFP	-		✓		W
Czech	1,330K	UFP			✓		
Czech-CAC	482K	UFP			✓		
Czech-CLTT	26K	UFP			✓		
Danish	94K	UFP			✓		
Dutch	197K	UFP	-		✓		
Dutch-LassySmall	93K	UFP	-		✓		W
English	229K	UFPB			✓		
English-ESL	88K	U			✓		
English-LinES	67K				✓		
English-ParTUT	38K	UFP			✓		W
Estonian	34K	UFP	-		✓		
Finnish	181K	UFPB			✓		W
Finnish-FTB	143K	UFP	-		✓		
French	381K	UFP			✓		W
French-ParTUT	17K	UFP			✓		W
French-Sequoia	58K	UFP	-				W
Galician	109K	U			✓		
Galician-TreeGal	14K	UFP			✓		
German	277K	UFP	-		✓		W
Gothic	45K	UFP	-		✓		
Greek	51K	UFP			✓		W
Hebrew	106K	F	-		✓		
Hindi	316K	UFP	-		✓		
Hungarian	37K	UFP			✓		
Indonesian	110K	F	-		✓		
Irish	13K	UFP			✓		
Italian	195K	UFP			✓		W
Italian-ParTUT	39K	UFP			✓		W
Japanese	175K	UFP			✓		
Japanese-KTC	189K	U			✓		
Kazakh	<1K	UFP			✓		W
Korean	63K	F					
Korean-Sejong	89K	U	-		?		
Latin	18K	UFP					
Latin-ITTB	280K	UFP	-		✓		
Latin-PROIEL	159K	UFP	-		✓		
Latvian	44K	UFP	-				
Lithuanian	40K	UFP	-		?		
Norwegian-Bokmaal	280K	UFP			✓		
Norwegian-Nynorsk	276K	UFP					
Old Church Slavonic	47K	UFP	-		✓		
Persian	135K	UFP			✓		
Polish	72K	UFP	-		✓		
Portuguese	201K	UFP			✓		
Portuguese-BR	268K	F	-		✓		
Romanian	202K	UFP			✓		W
Russian	87K	UFP			✓		W
Russian-SynTagRus	988K	UFP			✓		
Sanskrit	1K	UFP	-		✓		
Slovak	93K	UFP	-		✓		
Slovenian	126K	UFP			✓		
Slovenian-SST	19K	UFP			✓		
Spanish	411K	UFP			✓		W
Spanish-AnCora	495K	UFP			✓		
Swedish	76K	UFP			✓		
Swedish-LinES	64K				✓		
Swedish Sign Language	<1K		-		✓		
Tamil	8K	UFP	-		✓		
Turkish	46K	UFP			✓		
Ukrainian	12K	UFP			✓		
Urdu	123K	UFP	-				
Uyghur	1K		-		✓		
Vietnamese	31K	UFP	-		✓		

March 1, 2017:

- 50 languages
- 70 treebanks
- 162 contributors
- 8000+ downloads

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						

March 1, 2017:

- 50 languages
- 70 treebanks
- 162 contributors
- 8000+ downloads

Chief Cat Herder

Release and Documentation Task Force

Universal Guidelines Group

Goals and Requirements

Goals and Requirements

Cross-linguistically consistent grammatical annotation

Goals and Requirements

Cross-linguistically consistent grammatical annotation

Support multilingual research in NLP and linguistics

- Meaningful linguistic analysis within and across languages
- Syntactic parsing in monolingual and cross-lingual settings
- Useful information for downstream language understanding tasks

Goals and Requirements

Cross-linguistically consistent grammatical annotation

Support multilingual research in NLP and linguistics

- Meaningful linguistic analysis within and across languages
- Syntactic parsing in monolingual and cross-lingual settings
- Useful information for downstream language understanding tasks

Build on common usage and existing de facto standards

Goals and Requirements

Cross-linguistically consistent grammatical annotation

Support multilingual research in NLP and linguistics

- Meaningful linguistic analysis within and across languages
- Syntactic parsing in monolingual and cross-lingual settings
- Useful information for downstream language understanding tasks

Build on common usage and existing de facto standards

Complement – not replace – language-specific schemes

The UD Philosophy

The UD Philosophy

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

The UD Philosophy

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 language-specific elaboration

- Languages select from a universal pool of categories
- Allow language-specific extensions

Design Principles

Design Principles

Dependency

- Widely used in practical NLP systems
- Available in treebanks for many languages

Design Principles

Dependency

- Widely used in practical NLP systems
- Available in treebanks for many languages

Lexicalism

- Basic annotation units are words – syntactic words
- Words have morphological properties
- Words enter into syntactic relations

Design Principles

Dependency

- Widely used in practical NLP systems
- Available in treebanks for many languages

Lexicalism

- Basic annotation units are words – syntactic words
- Words have morphological properties
- Words enter into syntactic relations

Recoverability

- Transparent mapping from input text to word segmentation

Word Segmentation

What is a word?

- Single part-of-speech tag
- Real syntactic relation

Two-level segmentation

- Represent orthographic tokens in addition to syntactic words

Word Segmentation

What is a word?

- Single part-of-speech tag
- Real syntactic relation

Two-level segmentation

- Represent orthographic tokens in addition to syntactic words

Text	Words

Word Segmentation

What is a word?

- Single part-of-speech tag
- Real syntactic relation

Two-level segmentation

- Represent orthographic tokens in addition to syntactic words

Text	Words
del	di il

Word Segmentation

What is a word?

- Single part-of-speech tag
- Real syntactic relation

Two-level segmentation

- Represent orthographic tokens in addition to syntactic words

Text	Words
del	di il
dámelo	da me lo

Word Segmentation

What is a word?

- Single part-of-speech tag
- Real syntactic relation

Two-level segmentation

- Represent orthographic tokens in addition to syntactic words

Text	Words
del	di il
dámelo	da me lo
ושמהשמש	ו ש מ ה שמש

Word Segmentation

What is a word?

- Single part-of-speech tag
- Real syntactic relation

Two-level segmentation

- Represent orthographic tokens in addition to syntactic words

Text	Words
del	di il
dámelo	da me lo
ושמהשמש	ו ש מ ה שמש
大阪国際会議場	大阪 国際 会議場

Morphology

Le **chat** **chasse** **les** **chiens** **.**

Morphology

Le	**chat**	**chasse**	**les**	**chiens**	**.**
le	**chat**	**chasser**	**le**	**chien**	**.**

- Lemma representing the semantic content of the word

Morphology

Le	**chat**	**chasse**	**les**	**chiens**	**.**
le	**chat**	**chasser**	**le**	**chien**	**.**
DET	**NOUN**	**VERB**	**DET**	**NOUN**	**PUNCT**

- Lemma representing the semantic content of the word
- Part-of-speech tag representing its grammatical class

Morphology

Le … hiens .
le … hien .
DET … OUN PUNCT

Open	Closed	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

- Lemma rep … of the word
- Part-of-speech tag representing its grammatical class

Morphology

Le	chat	chasse	les	chiens	.
le	chat	chasser	le	chien	.
DET	NOUN	VERB	DET	NOUN	PUNCT
Definite=Def Gender=Masc Number=Sing	Gender=Masc Number=Sing	Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin	Definite=Def Gender=Masc Number=Plur	Gender=Masc Number=Plur	

- Lemma representing the semantic content of the word
- Part-of-speech tag representing its grammatical class
- Features representing lexical and grammatical properties of the lemma or the particular word form

Morphology

Le			hiens	.
le			hien	.
DET	N		OUN	PUNCT
Definite=Def Gender=Masc Number=Sing	Gend Num		er=Masc er=Plur	

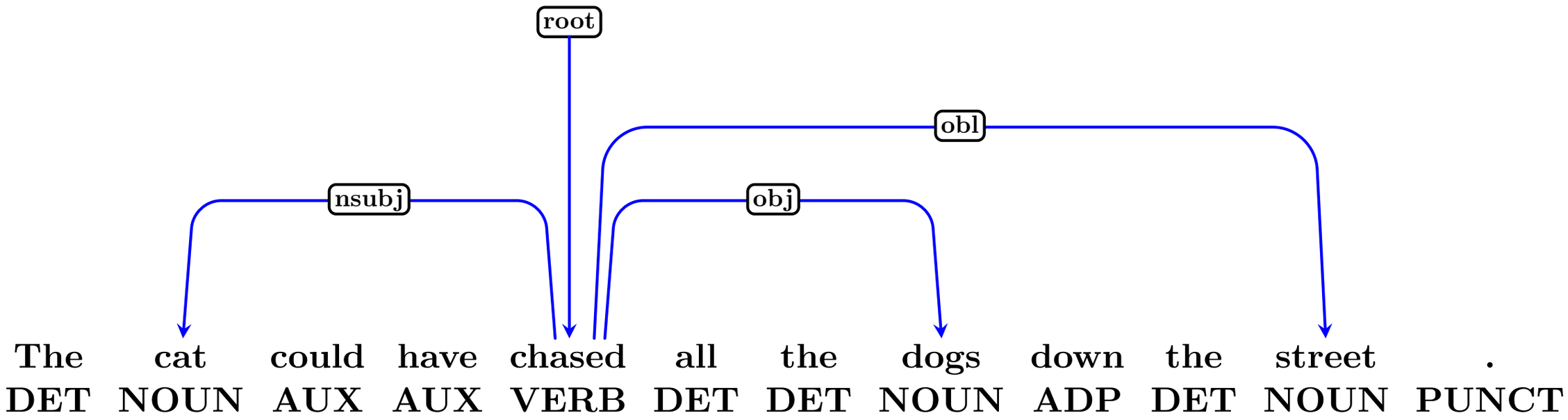
Lexical	Inflectional Nominal	Inflectional Verbal
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	Number	Tense
Reflex	Case	Aspect
Foreign	Definite	Voice
Abbr	Degree	Evident
		Polarity
		Person
		Polite

- Lemma rep of the word
- Part-of-spe tical class
- Features representing lexical and grammatical properties of the lemma or the particular word form

Syntax

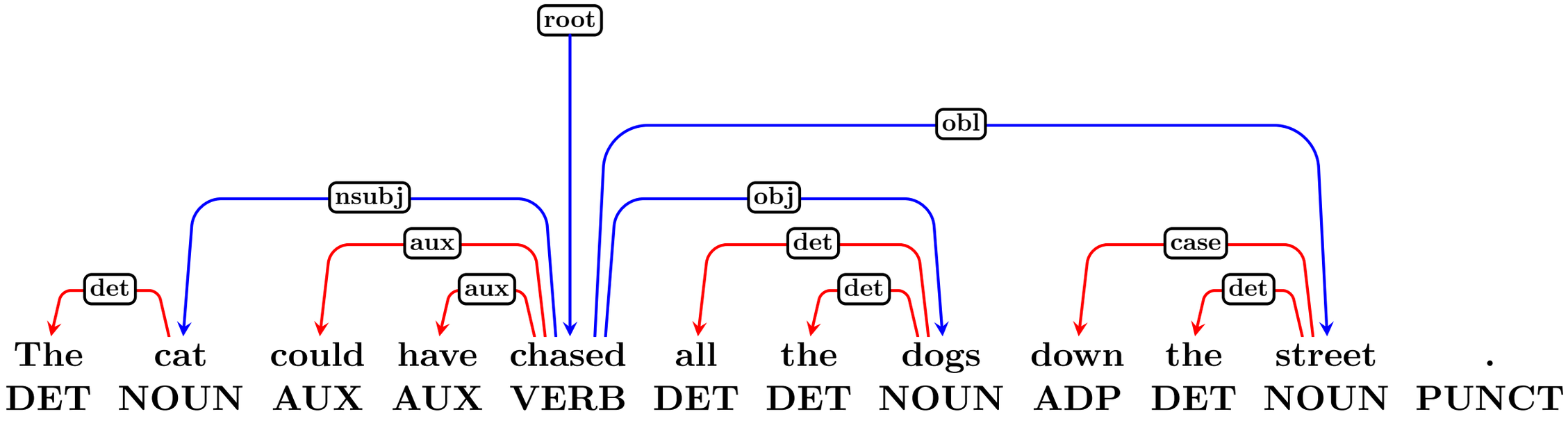
The	cat	could	have	chased	all	the	dogs	down	the	street	.
DET	NOUN	AUX	AUX	VERB	DET	DET	NOUN	ADP	DET	NOUN	PUNCT

Syntax



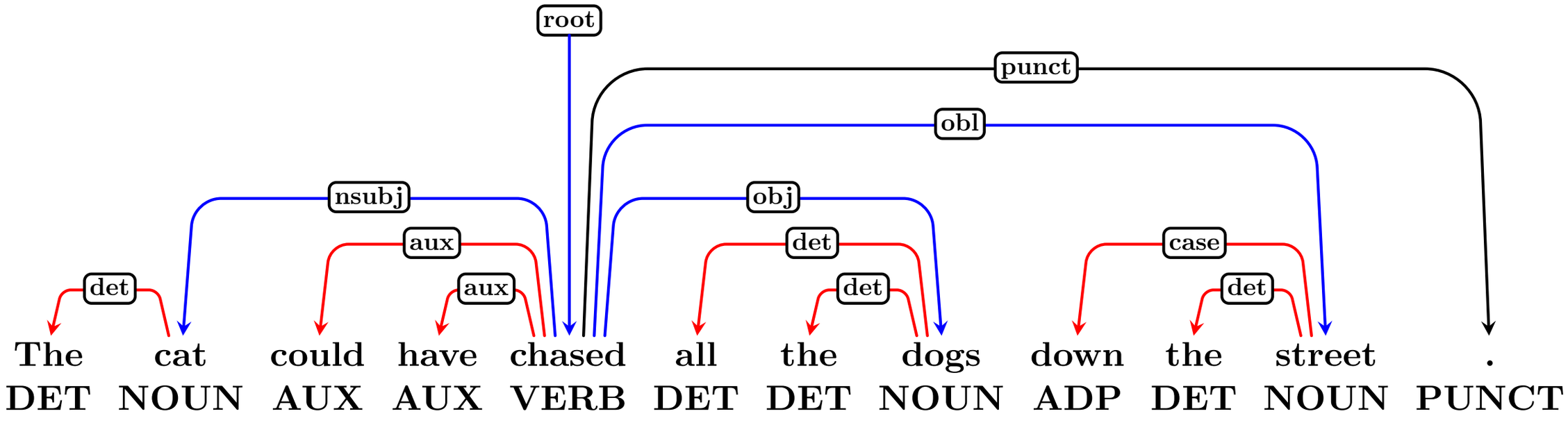
- Content words are related by dependency relations

Syntax

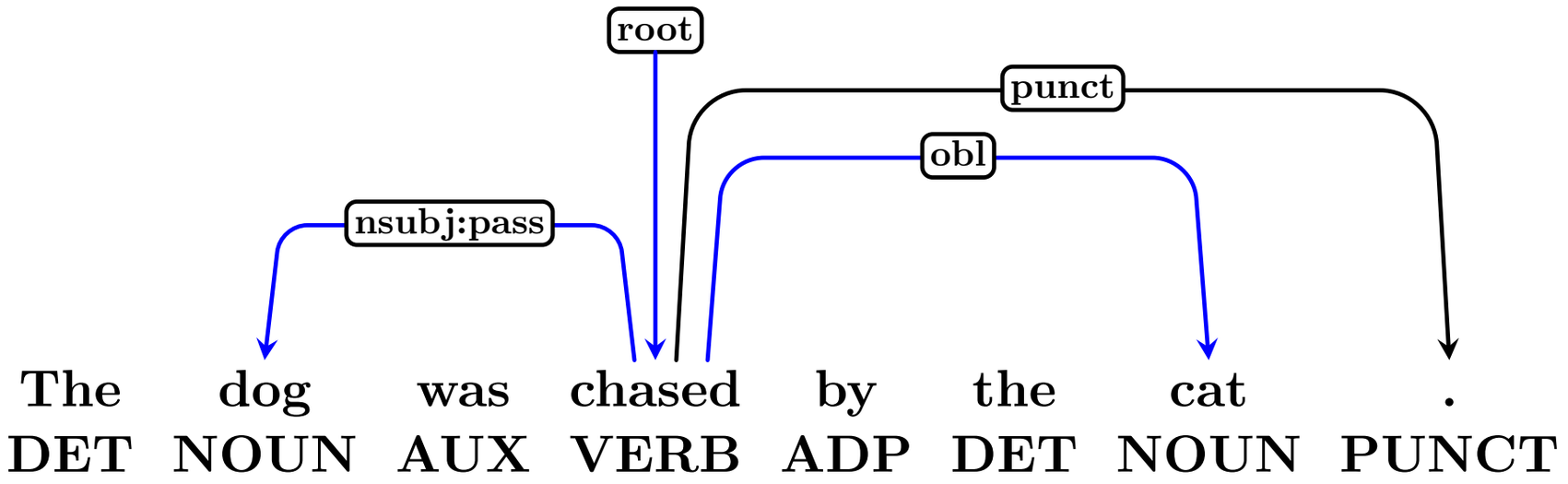


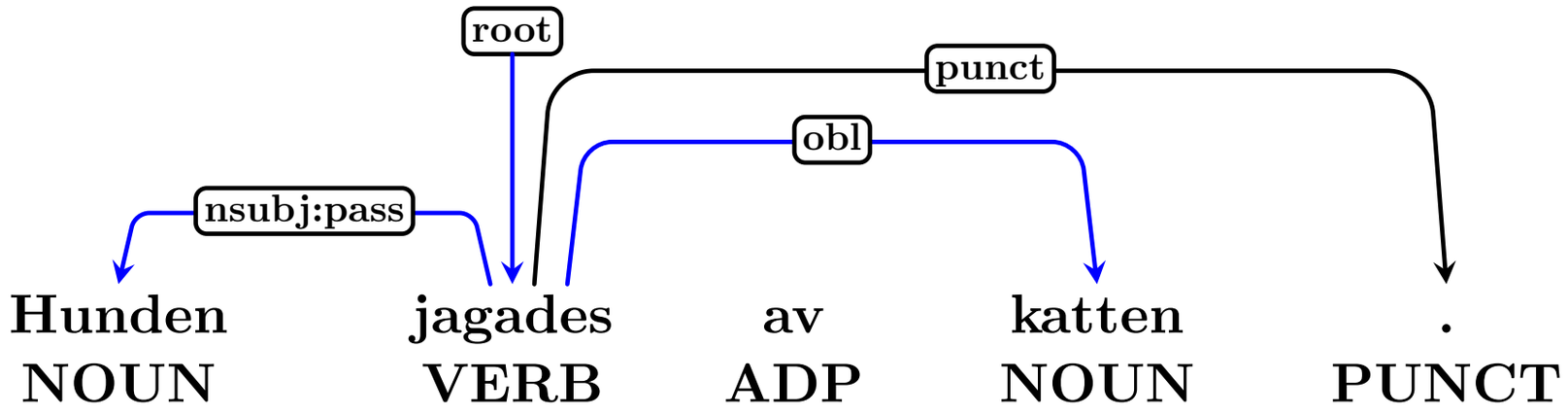
- Content words are related by dependency relations
- Function words attach to the content word they modify

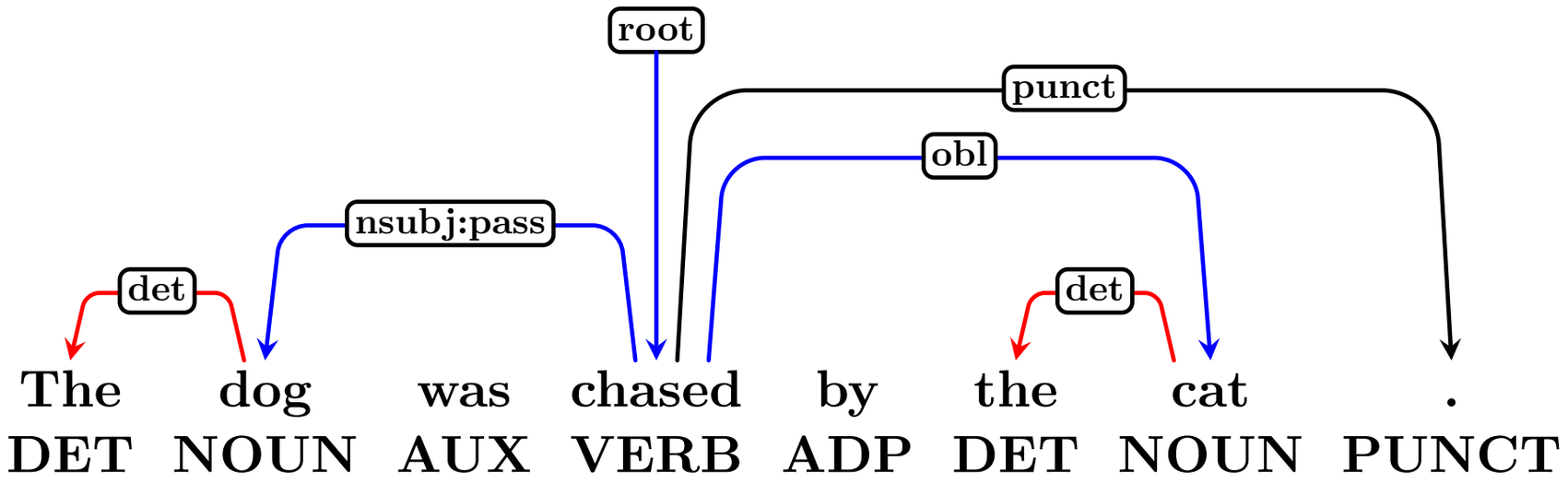
Syntax

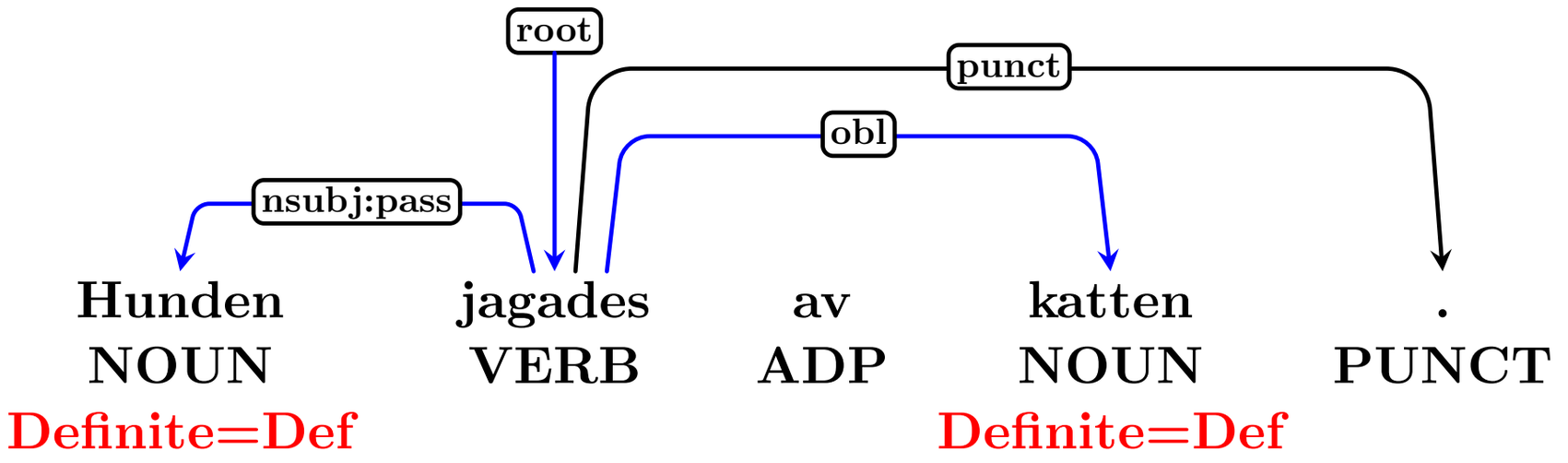


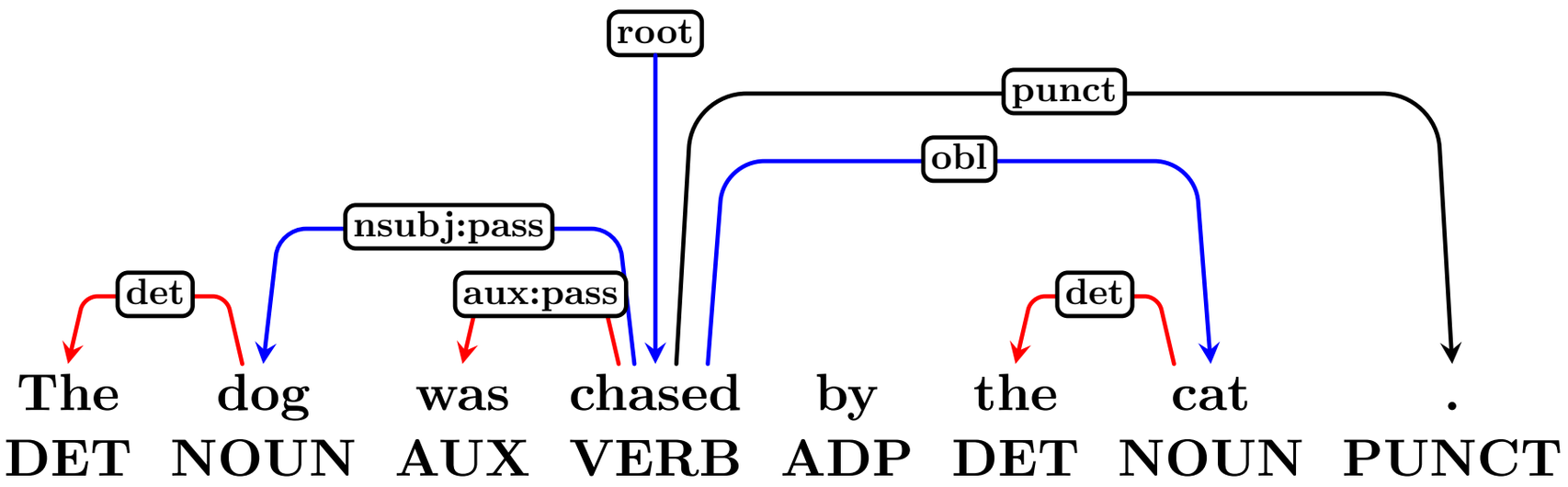
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

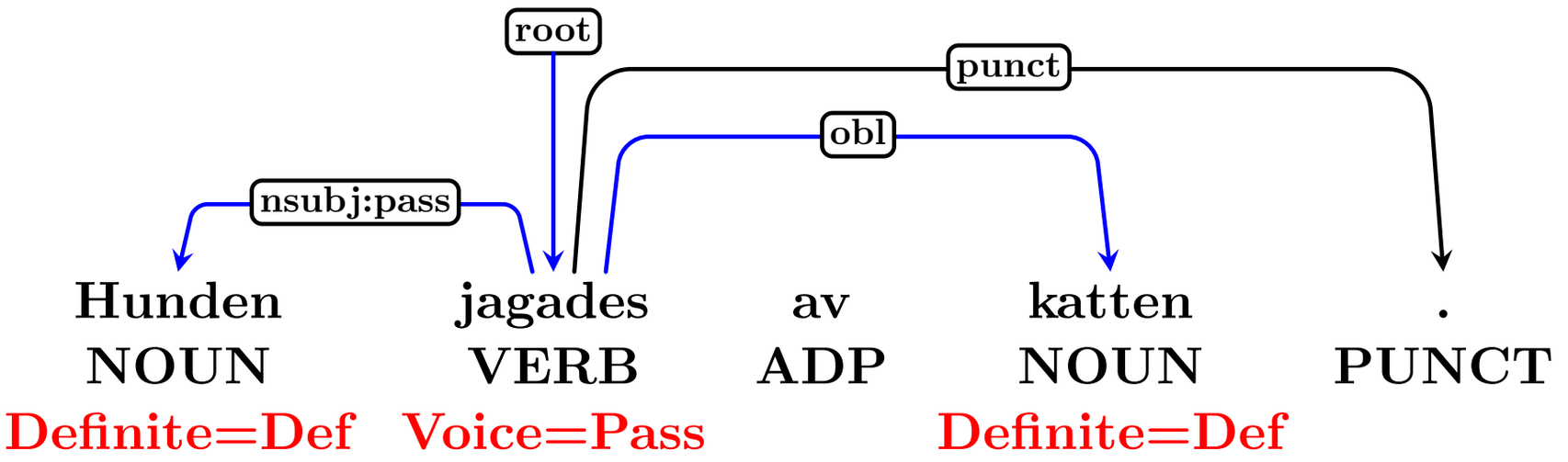

root
punct
obl
nsubj:pass
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT

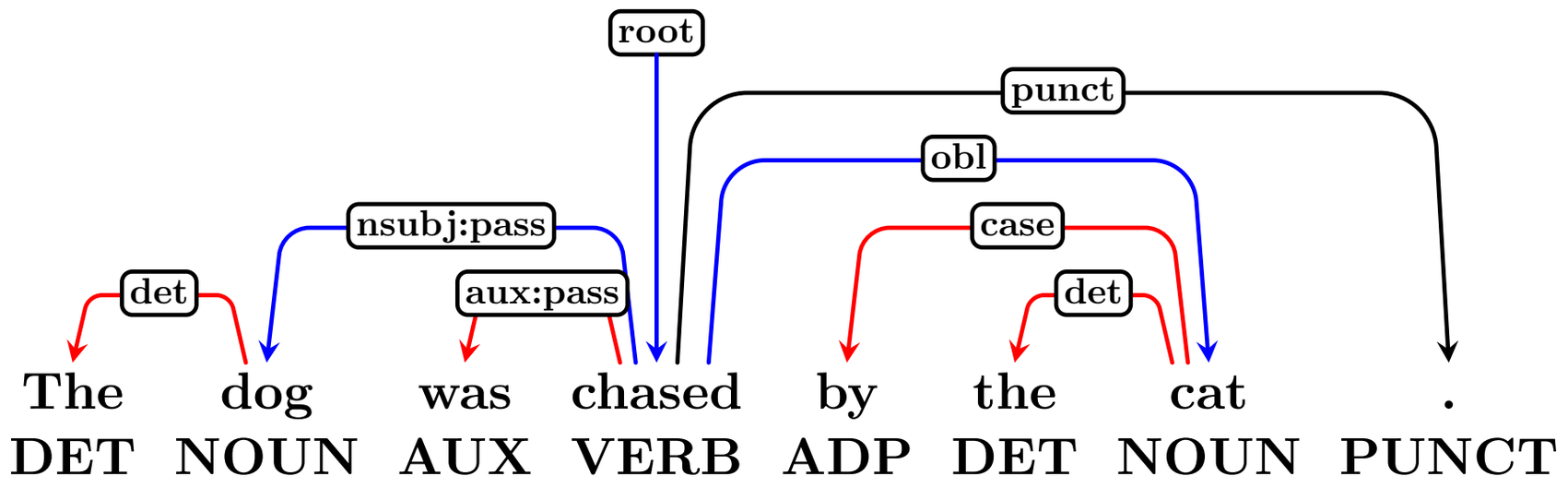

root
punct
obl
nsubj:pass
Hunden jagades av katten .
NOUN VERB ADP NOUN PUNCT

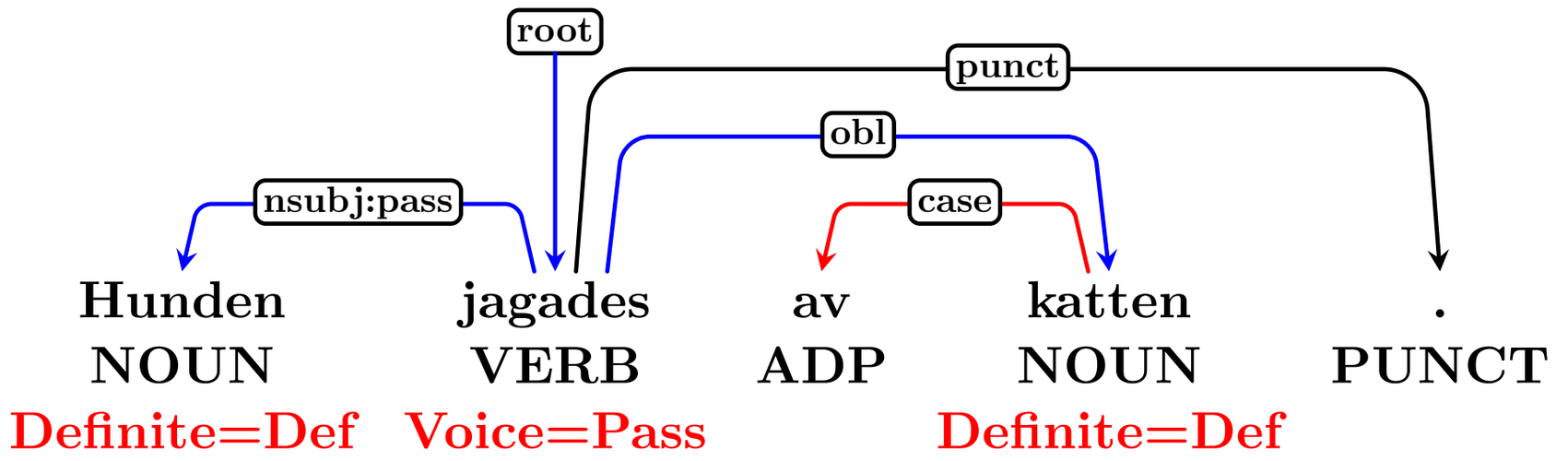

root
punct
obl
nsubj:pass
det
det
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT


root
punct
obl
nsubj:pass
Hunden jagades av katten .
NOUN VERB ADP NOUN PUNCT
Definite=Def
Definite=Def


root
punct
obl
nsubj:pass
det
aux:pass
det
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT


root
punct
obl
nsubj:pass
Hunden jagades av katten .
NOUN VERB ADP NOUN PUNCT
Definite=Def Voice=Pass Definite=Def


root
punct
obl
nsubj:pass
case
det
aux:pass
det
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT


root
punct
obl
nsubj:pass
case
Hunden jagades av katten .
NOUN VERB ADP NOUN PUNCT
Definite=Def Voice=Pass Definite=Def

Syntactic Relations

Syntactic Relations

Taxonomy of 37 universal syntactic relations

- Three types of structures: nominals, clauses, modifiers
- Core arguments vs. other dependents (not arguments vs. adjuncts)
- Language-specific subtypes

Syntactic Relations

Taxonomy of 37 universal syntactic relations

- Three types of structures: nominals, clauses, modifiers
- Core arguments vs. other dependents (not arguments vs. adjuncts)
- Language-specific subtypes

Basic and enhanced representations

- Basic dependencies form a (possibly non-projective) tree
- Additional dependencies in the enhanced representation

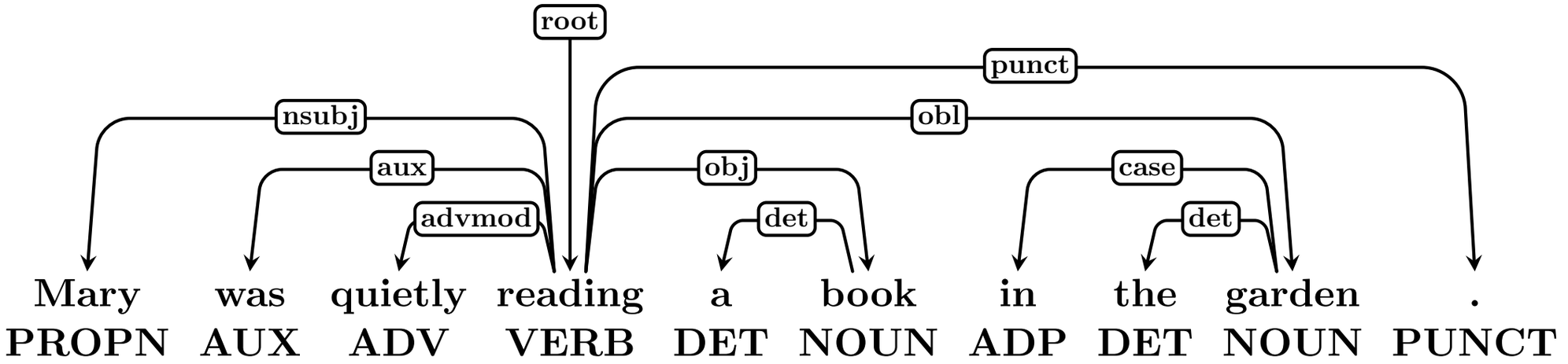
Syntactic Relations

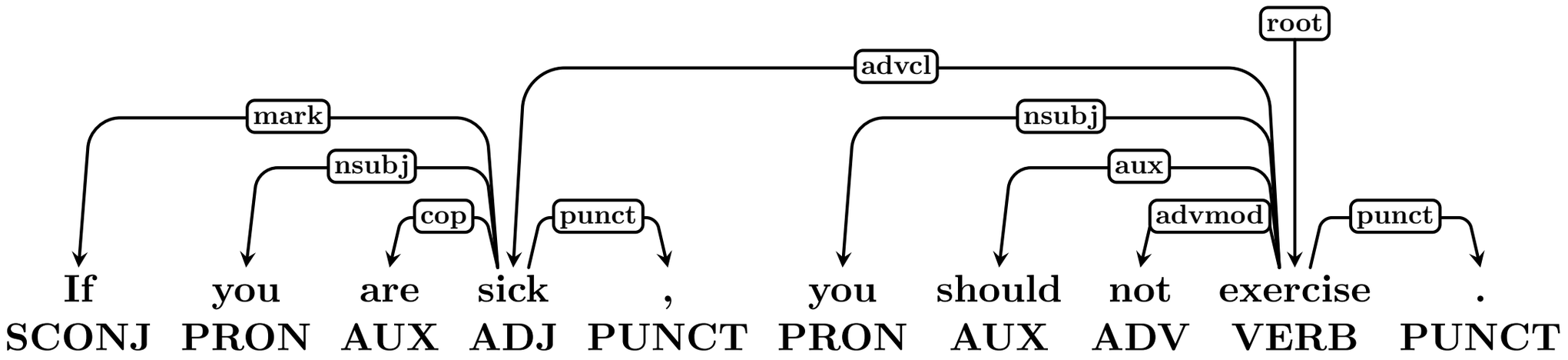
	Nominal	Clause	Modifier Word	Function Word
Core Predicate Dep	nsubj obj iobj	csubj ccomp xcomp		
Non-Core Predicate Dep	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal Dep	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed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

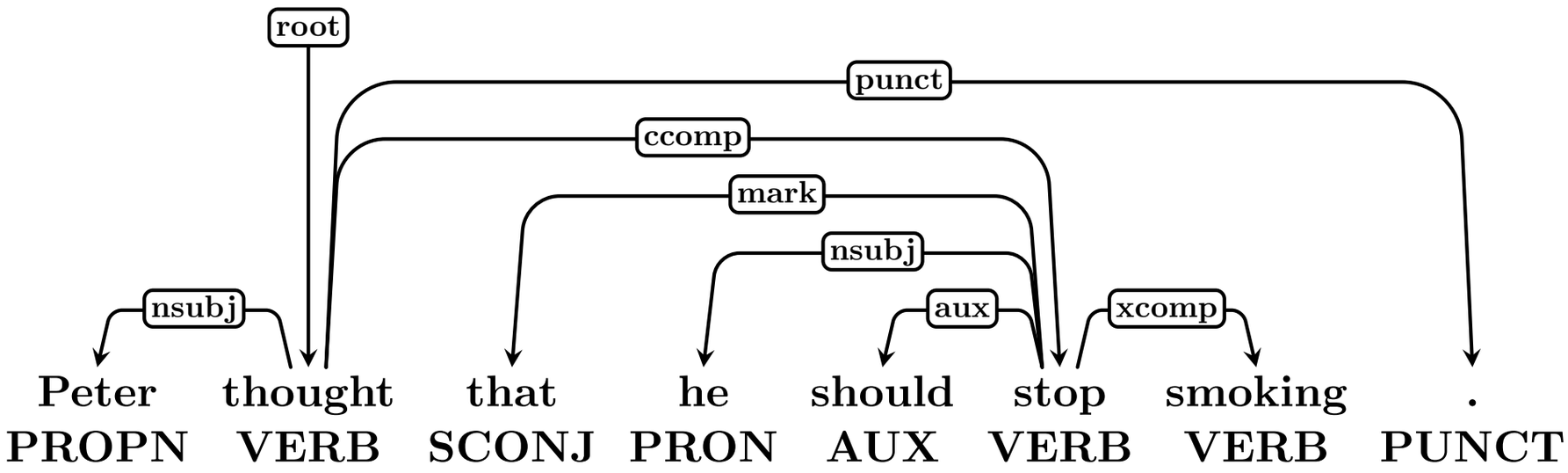
* Generalized modifier of predicates and (non-nominal) modifiers

Dependents of Clausal Predicates

	Nominal	Clause	Modifier Word	Function Word
Core	nsubj obj iobj	csubj ccomp xcomp		
Non-Core	obl vocative expl dislocated	advcl	advmod discourse	aux cop mark


root
punct
nsubj
obl
aux
obj
case
advmod
det
det
Mary was quietly reading a book in the garden .
PROPN AUX ADV VERB DET NOUN ADP DET NOUN PUNCT


root
advcl
mark
nsubj
nsubj
aux
cop
punct
advmod
punct
If you are sick , you should not exercise .
SCONJ PRON AUX ADJ PUNCT PRON AUX ADV VERB PUNCT


root
punct
ccomp
mark
nsubj
nsubj
aux
xcomp
Peter thought that he should stop smoking .
PROPN VERB SCONJ PRON AUX VERB VERB PUNCT

Core Arguments

Core Arguments

Arguments of basic intransitive and transitive verbs

- Verbs usually only agree with core arguments
- Core arguments normally appear as bare nominals without adpositions
- Certain cases, traditionally called nominative, accusative, and absolutive are typically reserved core arguments
- Core arguments often occupy special positions in the clause
- Syntactic phenomena like control, relativization and passivization can be restricted to core arguments

Core Arguments

Arguments of basic intransitive and transitive verbs

- Verbs usually only agree with core arguments
- Core arguments normally appear as bare nominals without adpositions
- Certain cases, traditionally called nominative, accusative, and absolutive are typically reserved core arguments
- Core arguments often occupy special positions in the clause
- Syntactic phenomena like control, relativization and passivization can be restricted to core arguments

Do not confuse

- Core arguments vs. oblique dependents – encoding of grammatical function
- Arguments vs. adjuncts – valency or subcategorization

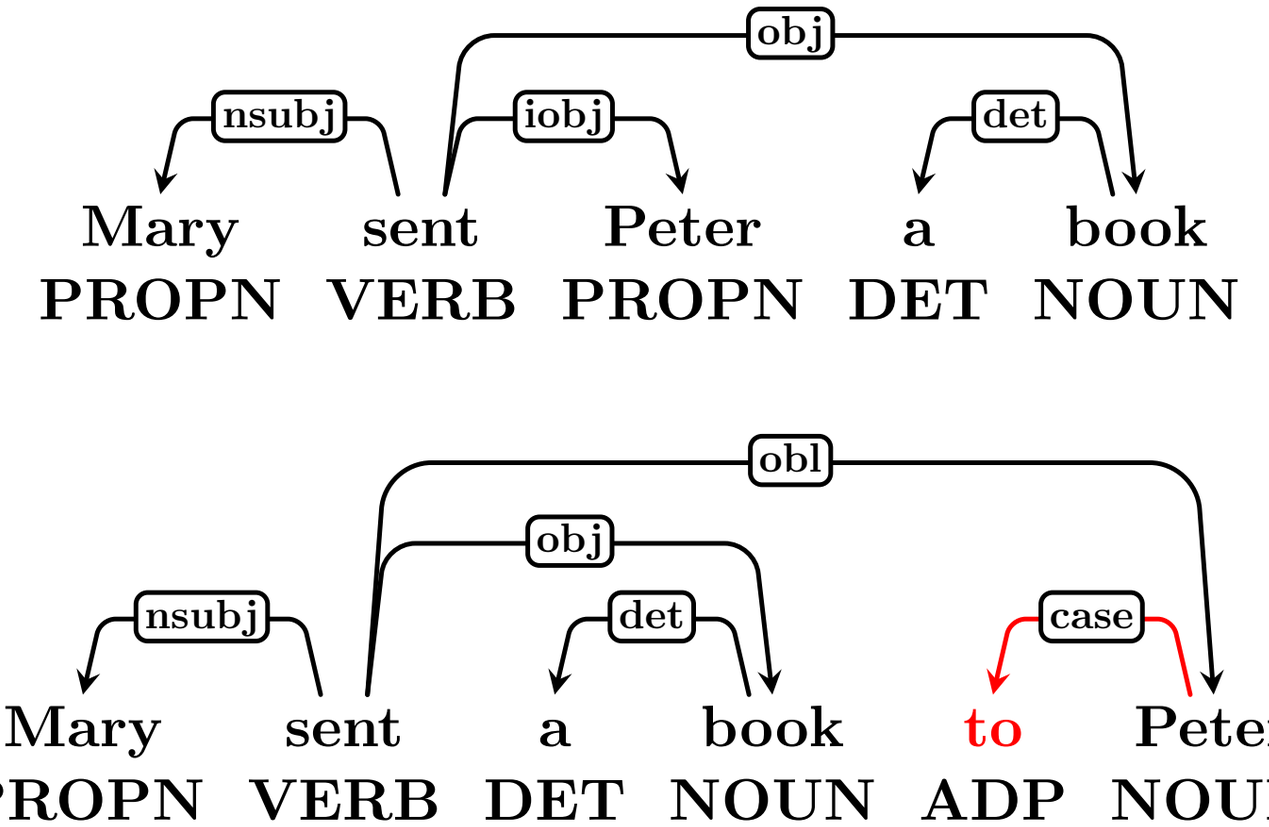
Core Arguments

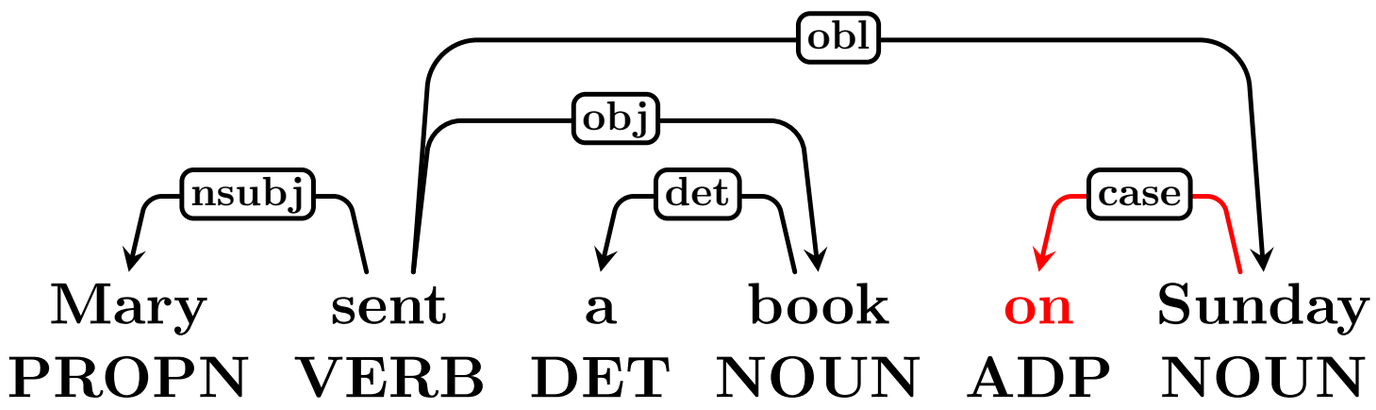
Arguments of basic intransitive and transitive verbs

- Verbs usually only agree with core arguments
- Core arguments normally appear as bare nominals without adpositions
- Certain cases, traditionally called nominative, accusative, and absolutive are typically reserved core arguments
- Core arguments often occupy special positions in the clause
- Syntactic phenomena like control, relativization and passivization can be restricted to core arguments

Do not confuse

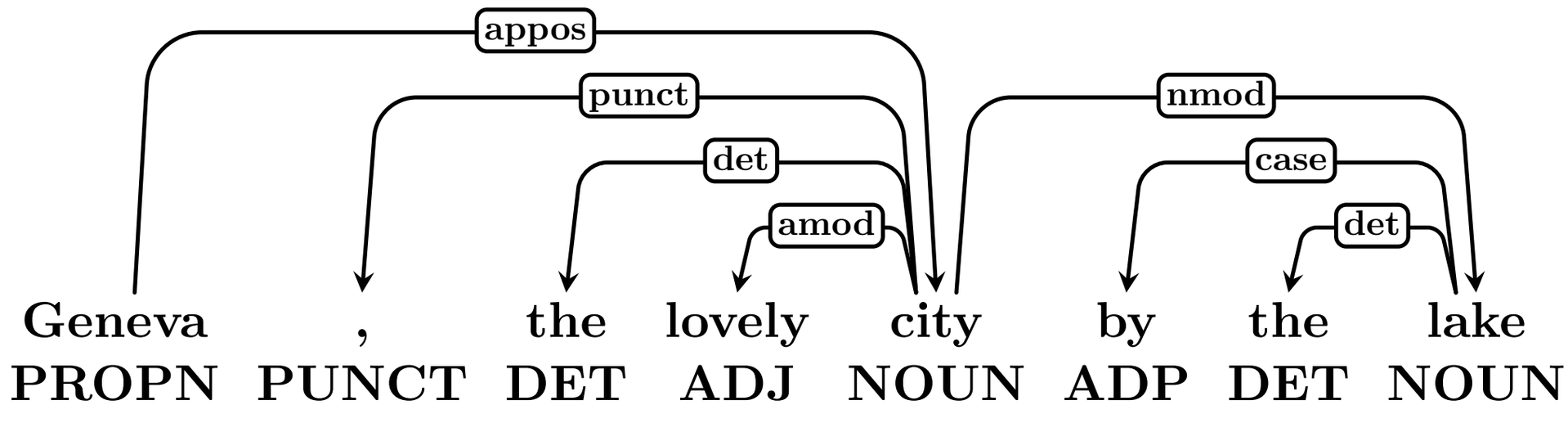
- Core arguments vs. oblique dependents – encoding of grammatical function
- Arguments vs. adjuncts – valency or subcategorization


obj
nsubj
iobj
det
Mary sent Peter a book
PROPN VERB PROPN DET NOUN
obl
obj
nsubj
det
case
Mary sent a book to Peter
PROPN VERB DET NOUN ADP NOUN

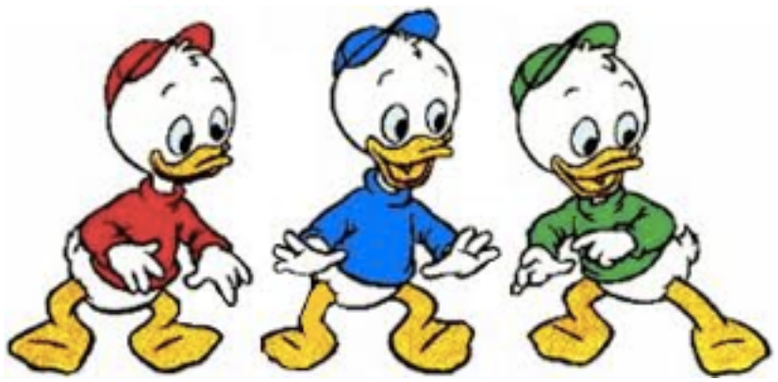

obl
obj
nsubj
det
case
Mary sent a book on Sunday
PROPN VERB DET NOUN ADP NOUN

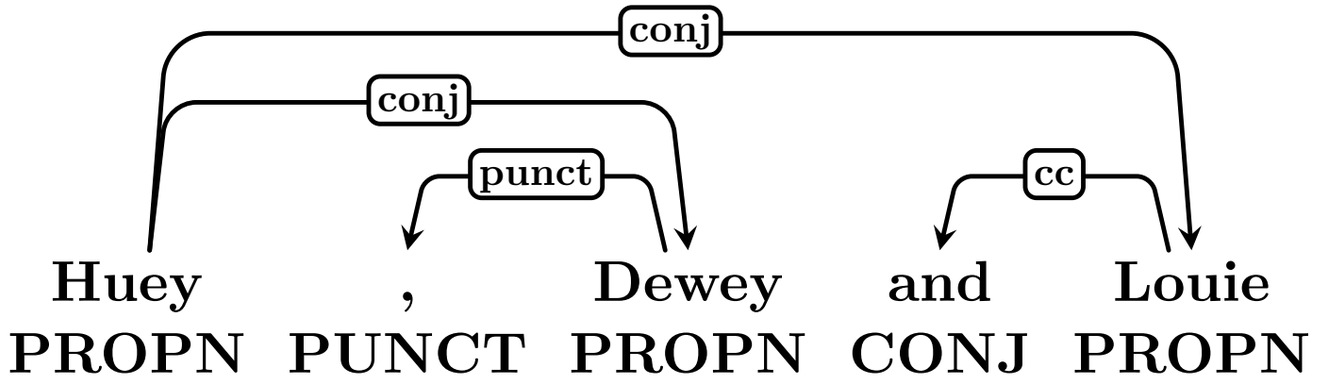
Dependents of Nominals

Nominal	Clause	Modifier Word	Function Word
nmod appos nummod	acl	amod	det clf case



Coordination

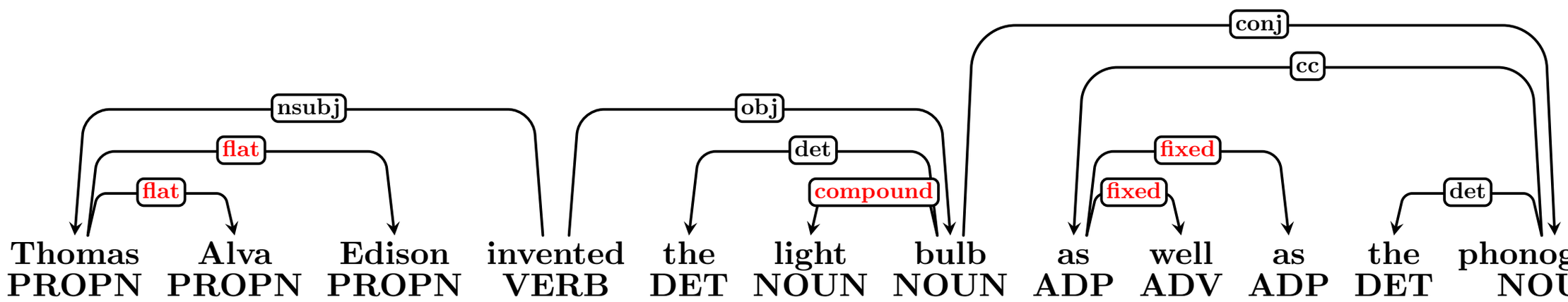




Coordinate structures are headed by the first conjunct

- Subsequent conjuncts depend on it via the conj relation
- Conjunction depends on following conjunct via the cc relation
- Punctuation depends on following conjunct via the punct relation

Multiword Expressions



Only restricted classes of MWEs get special treatment:

- Fixed grammaticized expressions (fixed)
- Semi-fixed expressions with no clear head (flat)
- Lexical compounds – normally headed (compound)

Loose Joining Relations

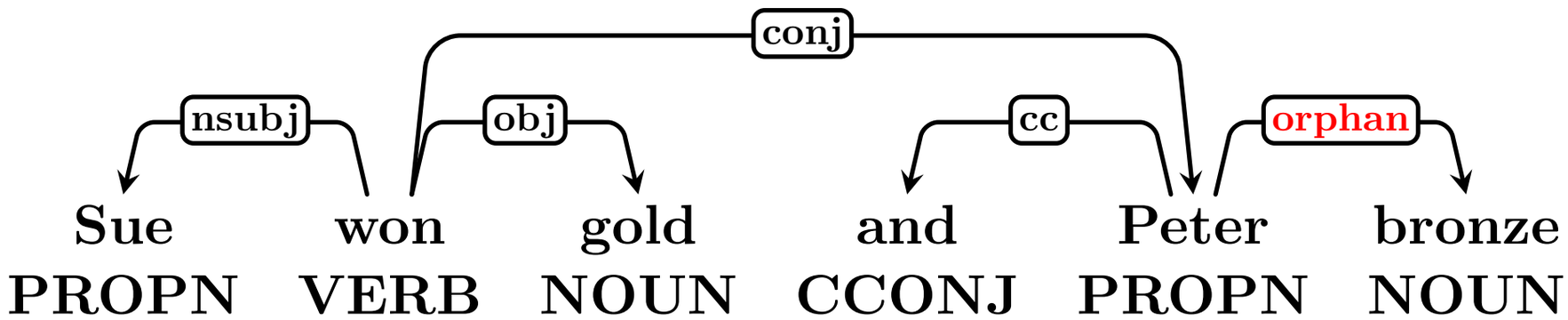
The parataxis relation:

- Side-by-side sentences ("run-on sentences")
  Bearded dragons are sight hunters, they need to see the food to move.
- Injective clauses (parentheticals)
  Calafia has great fries (they are to die for!) and decent burgers.
- Certain types of reported speech
  That guy, he said, left early this morning.
- Tag questions
  It's not me, is it?

The list relation:

- Chains of comparable items
  Steve Jones Phone: 555-9814 Email: jones@abc.edf

Ellipsis

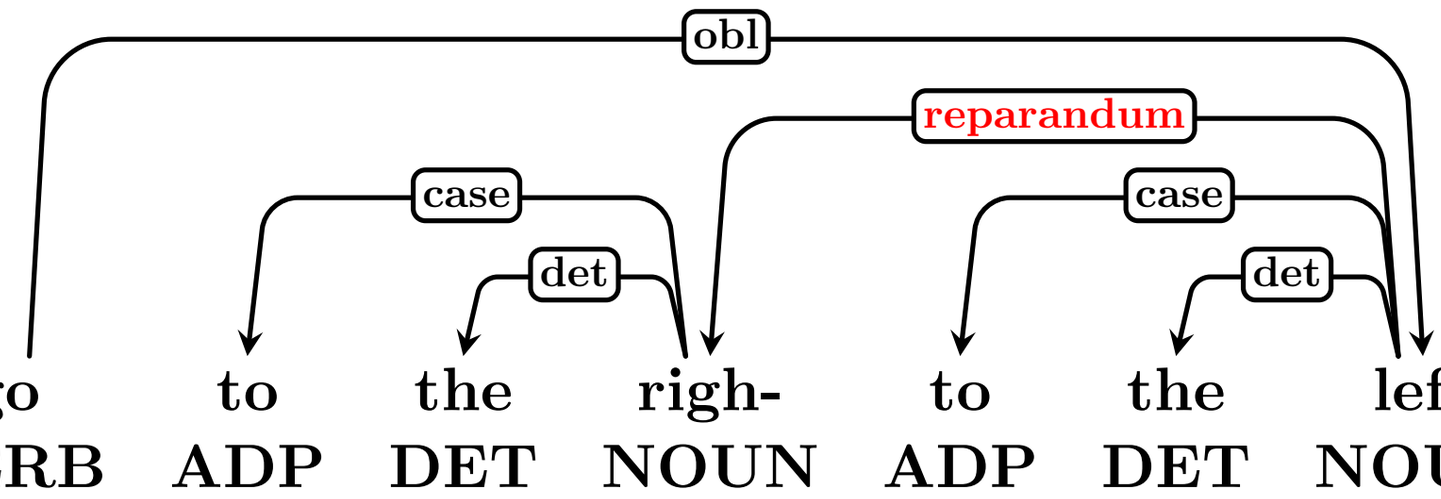


The UD approach to ellipsis (from v2):

1. If the elided word has no children, do nothing.
2. If the elided word has children, promote one of them to be the head.
3. If the elided word is a predicate and the new head a core argument, attach other non-functional elements with the orphan relation.

Implicit relations are recovered in enhanced dependencies

Disfluencies



The reparandum relation:

- Disfluencies that are overridden in a speech repair

The goeswith relation:

- Parts of words resulting from orthographic or editing mistakes

Punctuation

- A punctuation mark separating coordinated units is attached to the following conjunct.
- A punctuation mark preceding or following a subordinated unit is attached to this unit.
- Within the relevant unit, a punctuation mark is attach at the highest possible node that preserves projectivity.
- Paired punctuation marks should be attached to the same word unless that would create non-projectivity.

Special Relations

The root relation:

- The word at the root of the dependency tree
- Normally the predicate of the main clause
- Exactly one word in each tree

The dep relation:

- Unspecified syntactic relation (when all else fails)

A Two-Level Architecture

- Universal relations to allow cross-linguistic comparison
- Subtypes to capture language-specific phenomena

A Two-Level Architecture

- Universal relations to allow cross-linguistic comparison
- Subtypes to capture language-specific phenomena

Universal	Subtype

A Two-Level Architecture

- Universal relations to allow cross-linguistic comparison
- Subtypes to capture language-specific phenomena

Universal	Subtype
acl	acl:relcl

A Two-Level Architecture

- Universal relations to allow cross-linguistic comparison
- Subtypes to capture language-specific phenomena

Universal	Subtype
acl	acl:relcl
compound	compound:prt

A Two-Level Architecture

- Universal relations to allow cross-linguistic comparison
- Subtypes to capture language-specific phenomena

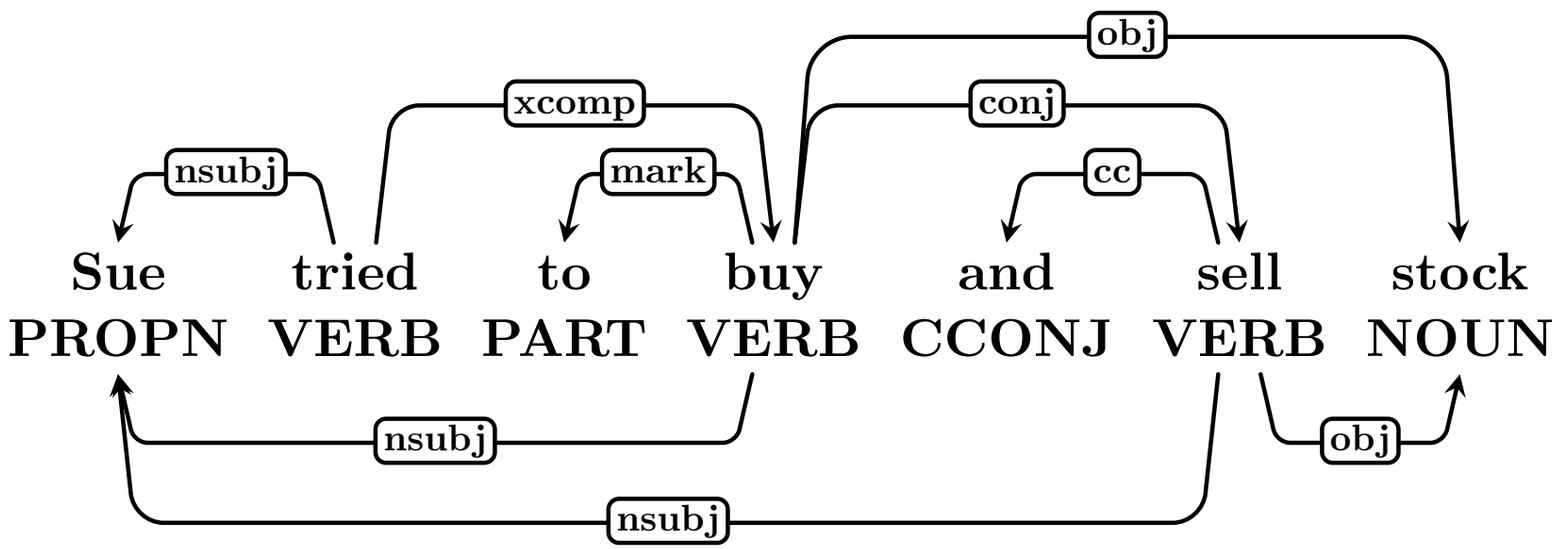
Universal	Subtype
acl	acl:relcl
compound	compound:prt
nmod	nmod:poss

A Two-Level Architecture

- Universal relations to allow cross-linguistic comparison
- Subtypes to capture language-specific phenomena

Universal	Subtype
acl	acl:relcl
compound	compound:prt
nmod	nmod:poss
flat	flat:name

Enhanced Dependencies



An extended dependency graph containing

- Null nodes for elided predicates
- Additional subject relations for control and raising constructions
- Propagation of dependents over coordination
- Coreference in relative clause constructions
- Labels augmented with function word information

CoNLL-U Format

- Revised version of the CoNLL-X format
- Two-level segmentation and enhanced dependencies

CoNLL-U Format

ID
1–2
1
2
3–4
3
4
5
6

- Revised version of the CoNLL-X format
- Two-level segmentation and enhanced dependencies

CoNLL-U Format

ID	FORM
1-2	Vámonos
1	Vamos
2	nos
3-4	al
3	a
4	el
5	mar
6	.

- Revised version of the CoNLL-X format
- Two-level segmentation and enhanced dependencies

CoNLL-U Format

ID	FORM	LEMMA
1-2	Vámonos	_
1	Vamos	ir
2	nos	nosotros
3-4	al	_
3	a	a
4	el	el
5	mar	mar
6	.	.

- Revised version of the CoNLL-X format
- Two-level segmentation and enhanced dependencies

CoNLL-U Format

ID	FORM	LEMMA	UPOSTAG
1-2	Vámonos	_	_
1	Vamos	ir	VERB
2	nos	nosotros	PRON
3-4	al	_	_
3	a	a	ADP
4	el	el	DET
5	mar	mar	NOUN
6	.	.	.

- Revised version of the CoNLL-X format
- Two-level segmentation and enhanced dependencies

CoNLL-U Format

ID	FORM	LEMMA	UPOSTAG	XPOSTAG
1-2	Vámonos	_	_	_
1	Vamos	ir	VERB	_
2	nos	nosotros	PRON	_
3-4	al	_	_	_
3	a	a	ADP	_
4	el	el	DET	_
5	mar	mar	NOUN	_
6	.	.	.	_

- Revised version of the CoNLL-X format
- Two-level segmentation and enhanced dependencies

CoNLL-U Format

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS
1-2	Vámonos	_	_	_	_
1	Vamos	ir	VERB	_	Mood=Imp|Number=Plur|Person=1
2	nos	nosotros	PRON	_	PronType=Per|Number=Plur|Person=1
3-4	al	_	_	_	_
3	a	a	ADP	_	_
4	el	el	DET	_	Definite=Def|Number=Sing|
5	mar	mar	NOUN	_	Number=Sing|Gender=Masc
6	.	.	.	_	_

- Revised version of the CoNLL-X format
- Two-level segmentation and enhanced dependencies

CoNLL-U Format

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD
1-2	Vámonos	_	_	_	_	_
1	Vamos	ir	VERB	_	Mood=Imp|Number=Plur|Person=1	0
2	nos	nosotros	PRON	_	PronType=Per|Number=Plur|Person=1	1
3-4	al	_	_	_	_	_
3	a	a	ADP	_	_	5
4	el	el	DET	_	Definite=Def|Number=Sing|	5
5	mar	mar	NOUN	_	Number=Sing|Gender=Masc	1
6	.	.	.	_	_	1

- Revised version of the CoNLL-X format
- Two-level segmentation and enhanced dependencies

CoNLL-U Format

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL
1-2	Vámonos	_	_	_	_	_	_
1	Vamos	ir	VERB	_	Mood=Imp|Number=Plur|Person=1	0	root
2	nos	nosotros	PRON	_	PronType=Per|Number=Plur|Person=1	1	expl
3-4	al	_	_	_	_	_	_
3	a	a	ADP	_	_	5	case
4	el	el	DET	_	Definite=Def|Number=Sing|	5	det
5	mar	mar	NOUN	_	Number=Sing|Gender=Masc	1	obl
6	.	.	.	_	_	1	punct

- Revised version of the CoNLL-X format
- Two-level segmentation and enhanced dependencies

CoNLL-U Format

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS
1-2	Vámonos	_	_	_	_	_	_	_
1	Vamos	ir	VERB	_	Mood=Imp|Number=Plur|Person=1	0	root	_
2	nos	nosotros	PRON	_	PronType=Per|Number=Plur|Person=1	1	expl	_
3-4	al	_	_	_	_	_	_	_
3	a	a	ADP	_	_	5	case	_
4	el	el	DET	_	Definite=Def|Number=Sing|	5	det	_
5	mar	mar	NOUN	_	Number=Sing|Gender=Masc	1	obl	_
6	.	.	.	_	_	1	punct	_

- Revised version of the CoNLL-X format
- Two-level segmentation and enhanced dependencies

CoNLL-U Format

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1-2	Vámonos	_	_	_	_	_	_	_	_
1	Vamos	ir	VERB	_	Mood=Imp|Number=Plur|Person=1	0	root	_	_
2	nos	nosotros	PRON	_	PronType=Per|Number=Plur|Person=1	1	expl	_	_
3-4	al	_	_	_	_	_	_	_	_
3	a	a	ADP	_	_	5	case	_	_
4	el	el	DET	_	Definite=Def|Number=Sing|	5	det	_	_
5	mar	mar	NOUN	_	Number=Sing|Gender=Masc	1	obl	_	_
6	.	.	.	_	_	1	punct	_	_

- Revised version of the CoNLL-X format
- Two-level segmentation and enhanced dependencies

http://universaldependencies.org

So what exactly is UD?

So what exactly is UD?

A new linguistic theory?

- Not at all, but we like to think it is informed by linguistic theory and definitely useful also for linguistic studies

So what exactly is UD?

A new linguistic theory?

- Not at all, but we like to think it is informed by linguistic theory and definitely useful also for linguistic studies

A better parsing framework?

- Maybe not, since parsers seem to prefer function words as heads so we may have to tweak the representations for parsing – but stay tuned

So what exactly is UD?

A new linguistic theory?

- Not at all, but we like to think it is informed by linguistic theory and definitely useful also for linguistic studies

A better parsing framework?

- Maybe not, since parsers seem to prefer function words as heads so we may have to tweak the representations for parsing – but stay tuned

The ultimate annotation scheme?

- Not quite, more like a lingua franca for treebank developers and definitely useful for some annotation projects

So what exactly is UD?

A new linguistic theory?

- Not at all, but we like to think it is informed by linguistic theory and definitely useful also for linguistic studies

A better parsing framework?

- Maybe not, since parsers seem to prefer function words as heads so we may have to tweak the representations for parsing – but stay tuned

The ultimate annotation scheme?

- Not quite, more like a lingua franca for treebank developers and definitely useful for some annotation projects

A universal grammar?

- Not in any strong sense, but hopefully in the sense of providing comparable concepts for meaningful cross-linguistic comparison

Manning's Law

The secret to understanding the design of UD is to realize that it 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.**
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.

UD Events in 2017

Tutorial on Universal Dependencies

- Tutorial at EACL, April 4, 2017, Valencia, Spain

CoNLL-2017 Shared Task

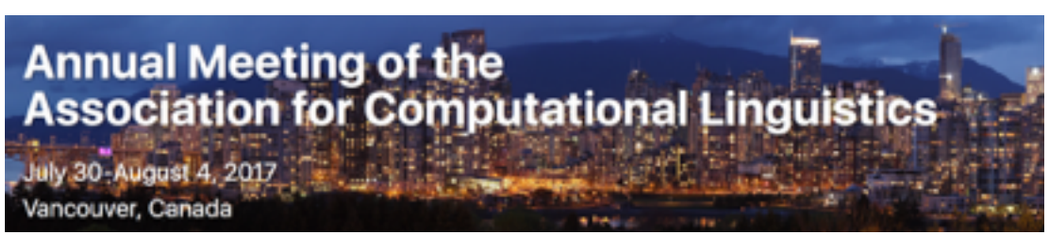
http://universaldependencies.org/conll17/

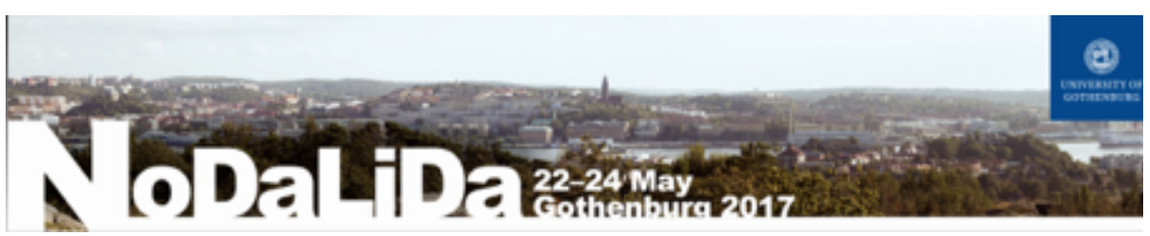
- Multilingual parsing from raw text to universal dependencies
- Collocated with ACL, August 3–4, 2017, Vancouver, Canada

First Workshop on Universal Dependencies

http://universaldependencies.org/udw17/

- Collocated with NoDaLiDa, May 20, 2017, Gothenburg, Sweden





Thanks to all UD contributors!

Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Yevgeni Berzak, Riyaz Ahmad Bhat, Eckhard Bick, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Minho Choi, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Claudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Jessica Kenney, Natalia Kotsyba, Simon Krek, Veronika Laippala, Lucia Lam, Phương Lê Hồng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Keiko Sophie Mori, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Joakim Nivre, Hanna Nurmi, Petya Osenova, Robert Östling, Lilja Øvrelid, Valeria de Paiva, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Manuela Sanguinetti, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Dmitry Sichinava, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jing Xian Wang, Jonathan North Washington, Mats Wirén, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Hanzhi Zhu