UPPSALA
UNIVERSITET

Universal Dependencies

Dubious Linguistics and Crappy Parsing?

Joakim Nivre

Uppsala University
Department of Linguistics and Philology

Based on collaborative work with Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher Manning, Ryan McDonald, Natalia Silveira, Slav Petrov, Sampo Pyysalo, Sebastian Schuster, Reut Tsarfaty, Francis Tyers, Daniel Zeman and many others

Introduction

Introduction

Growing interest in multilingual and cross-lingual NLP

- Multilingual evaluation campaigns to test generality of approaches
- Cross-lingual learning to support low-resource languages

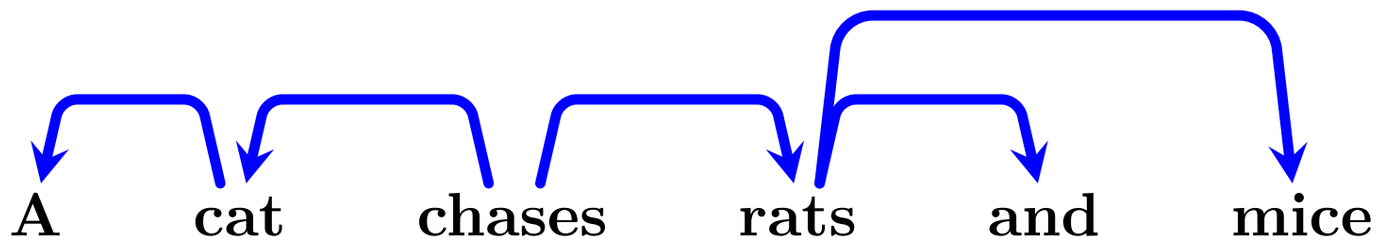
Introduction

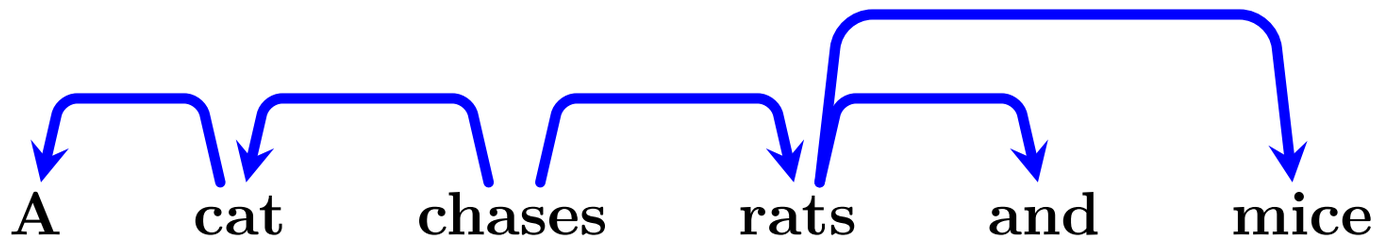
Growing interest in multilingual and cross-lingual NLP

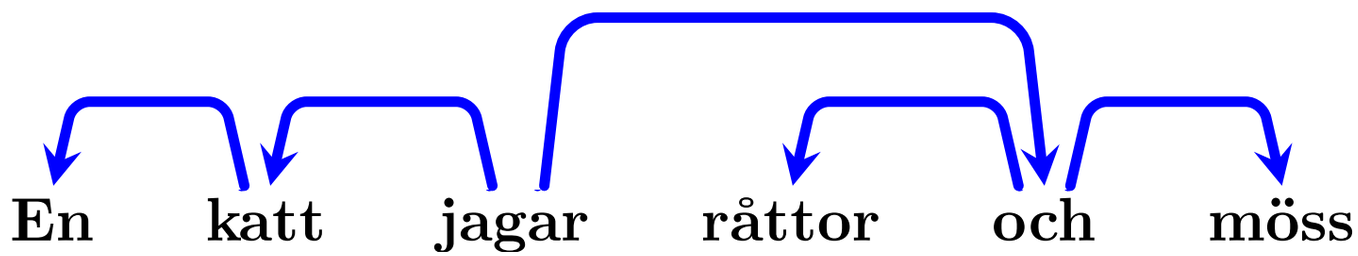
- Multilingual evaluation campaigns to test generality of approaches
- Cross-lingual learning to support low-resource languages

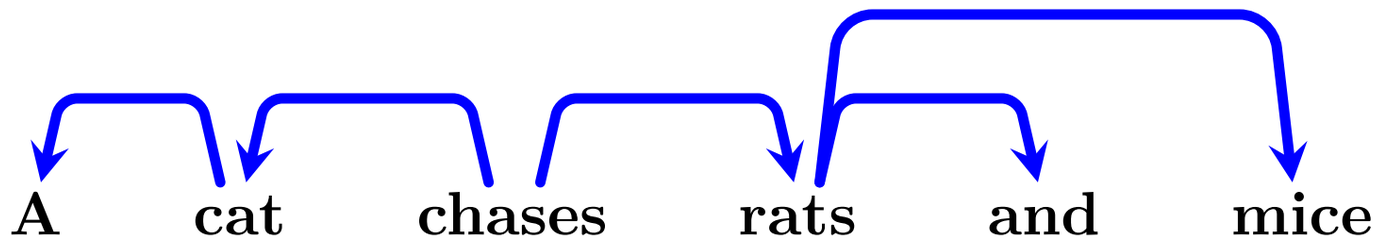
Growing awareness of methodological problems

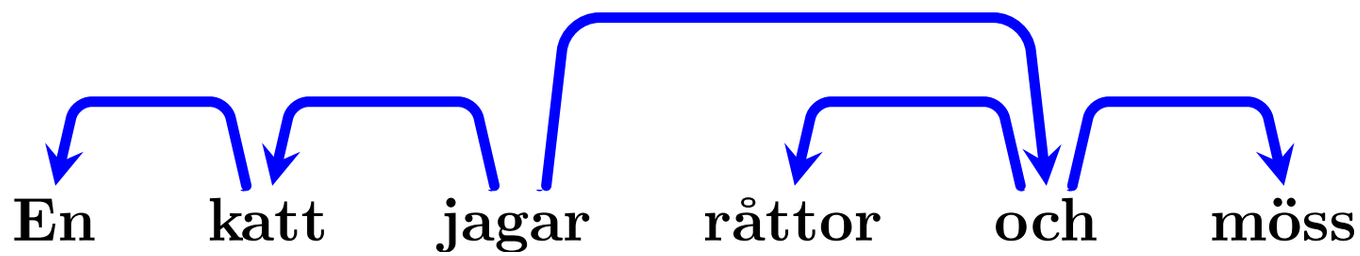
- Current NLP relies heavily on linguistic annotation
- Annotation guidelines vary across languages

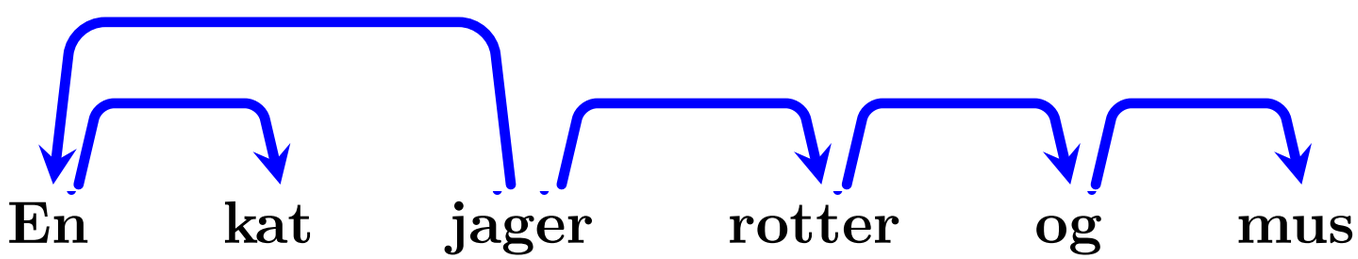

A
cat
chases
rats
and
mice

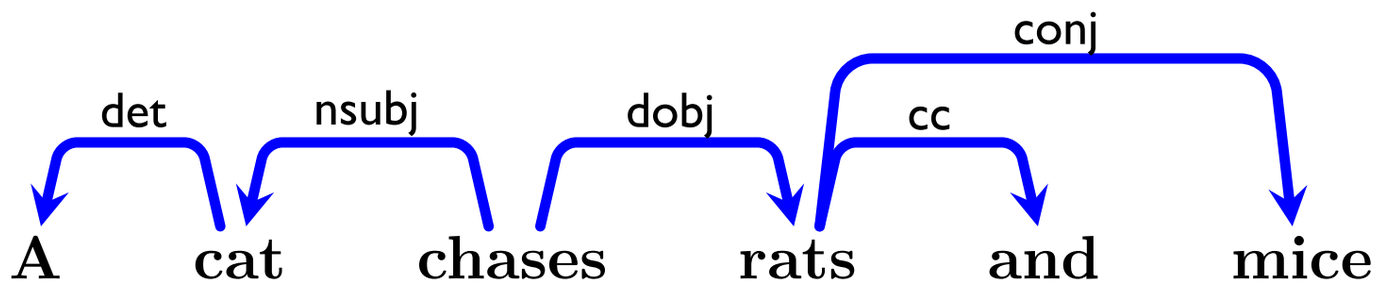

A
cat
chases
rats
and
mice

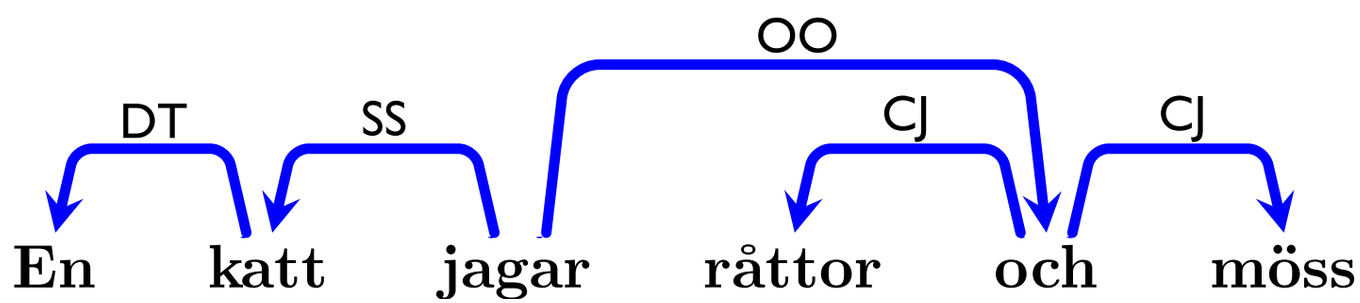

En
katt
jagar
råttor
och
möss

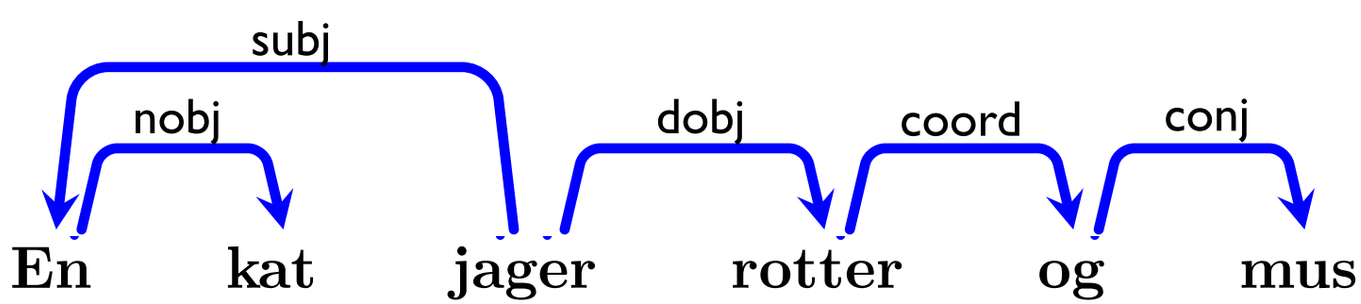

A
cat
chases
rats
and
mice

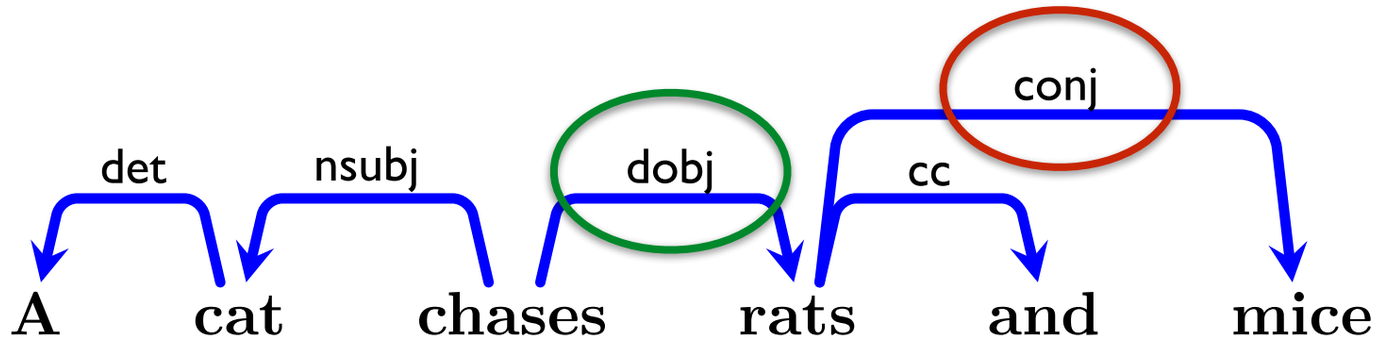

En
katt
jagar
råttor
och
möss

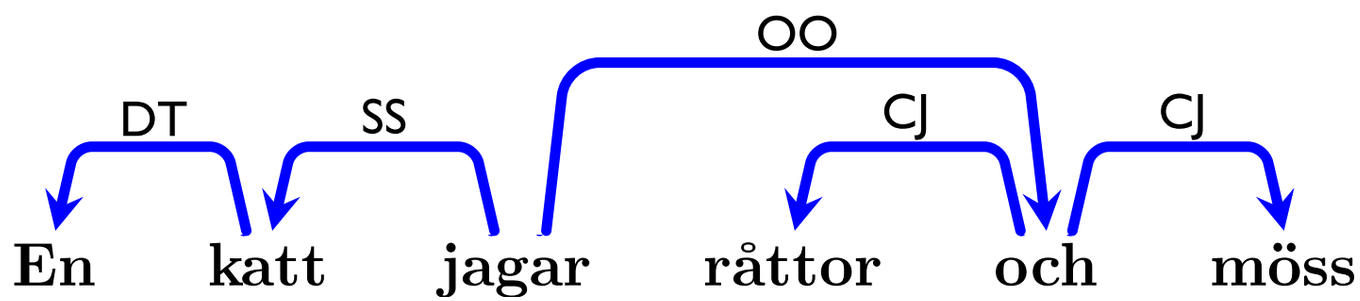

En
kat
jager
rotter
og
mus

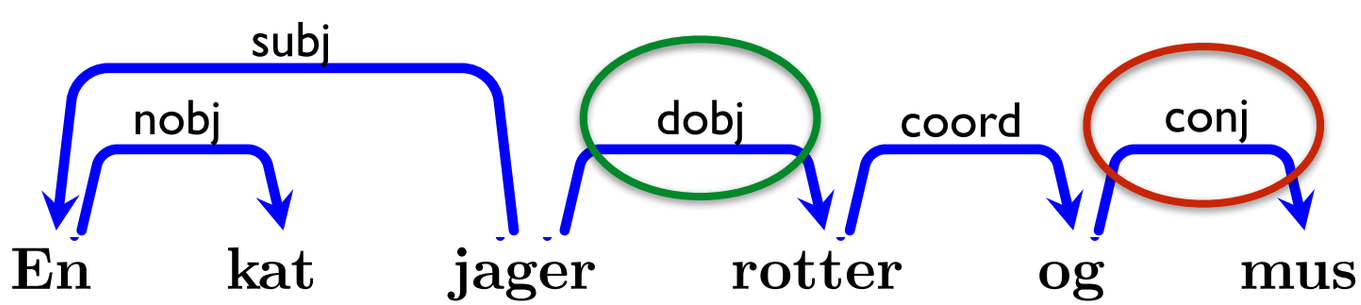

conj
det
nsubj
dobj
cc
A
cat
chases
rats
and
mice


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 transfer

Why is this a problem?

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

Why is this a problem?

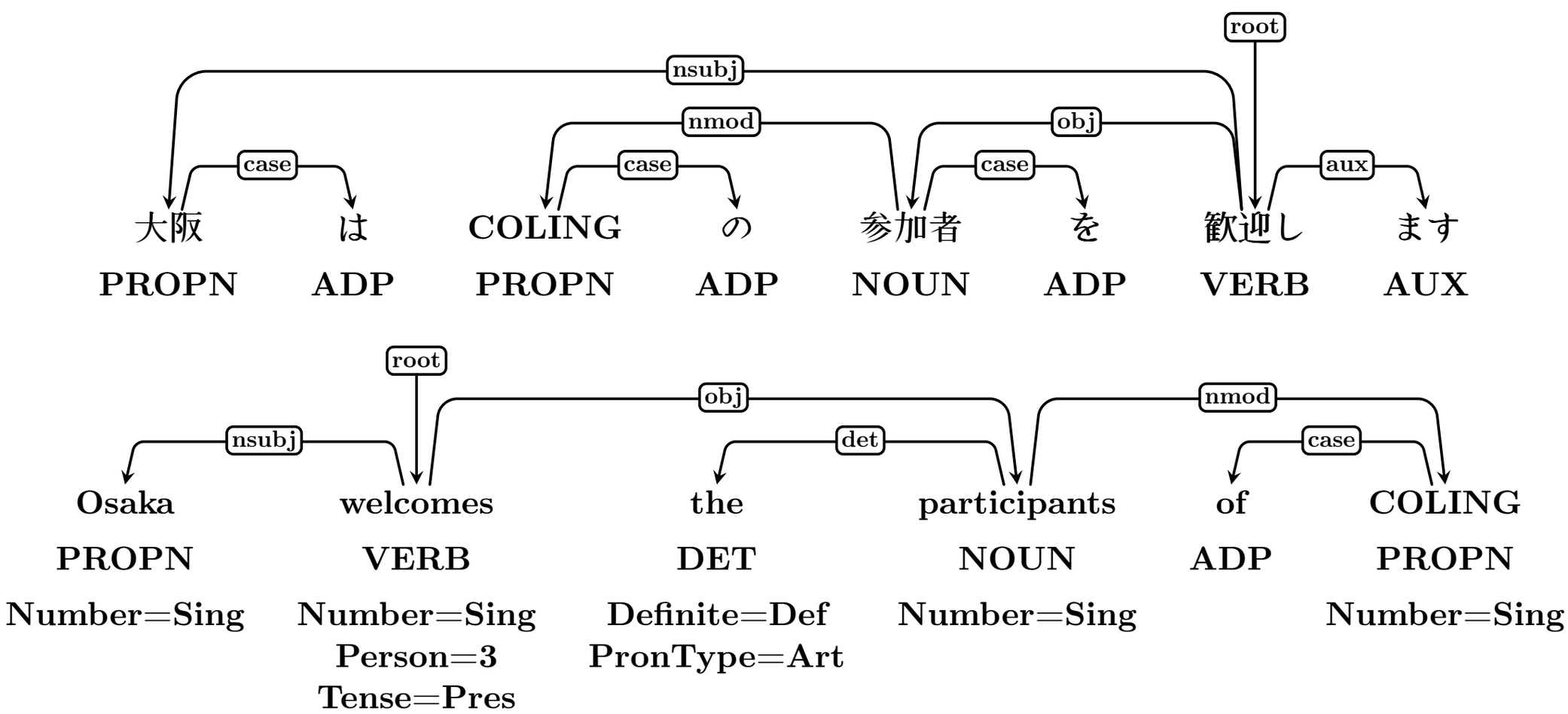
- Hard to compare empirical results across languages
- Hard to usefully do cross-lingual structure transfer
- Hard to evaluate cross-lingual learning
- Hard to build and maintain multilingual systems
- Hard to make comparative linguistic studies
- Hard to validate linguistic typology

Why is this a problem?

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

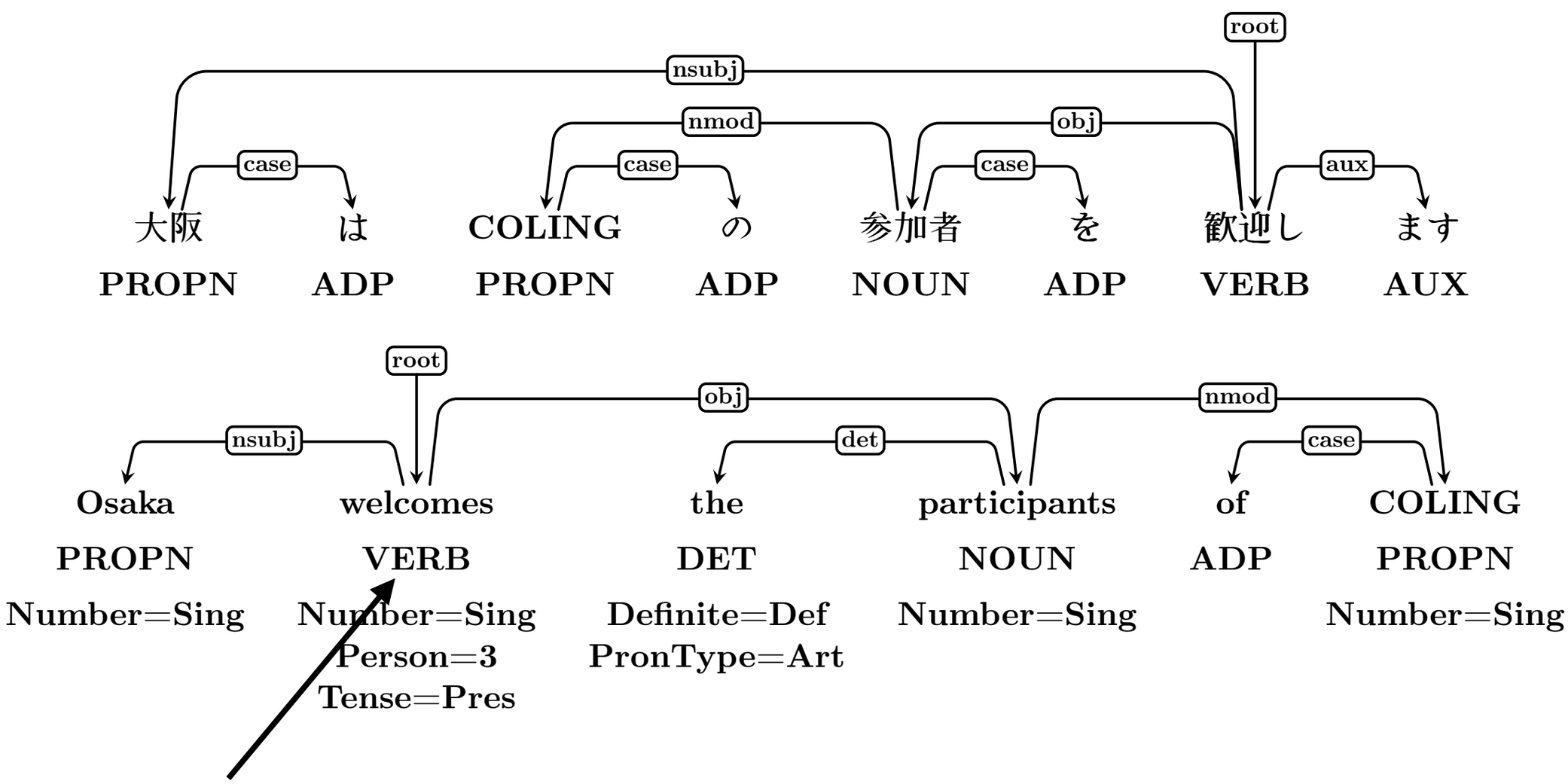
Universal Dependencies

http://universaldependencies.org



Universal Dependencies

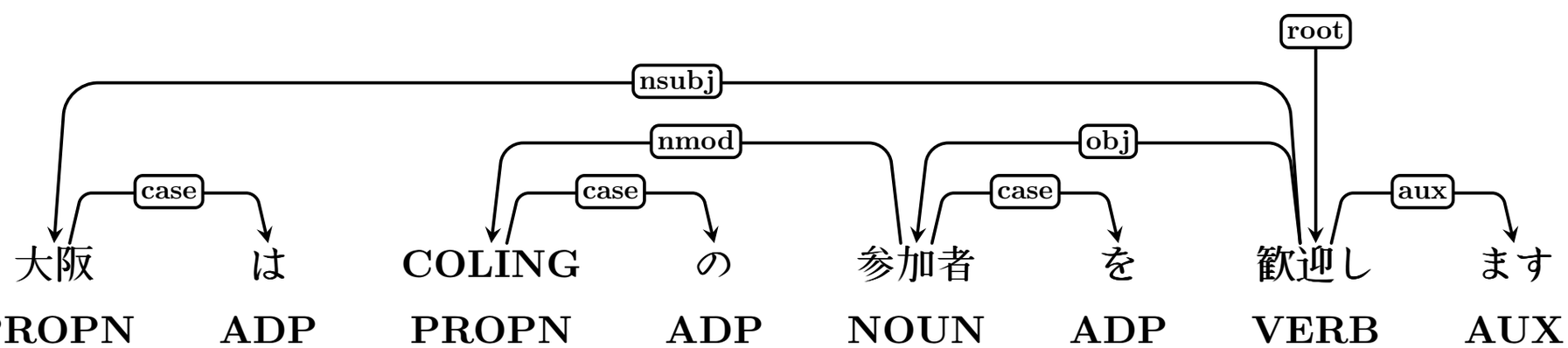
http://universaldependencies.org

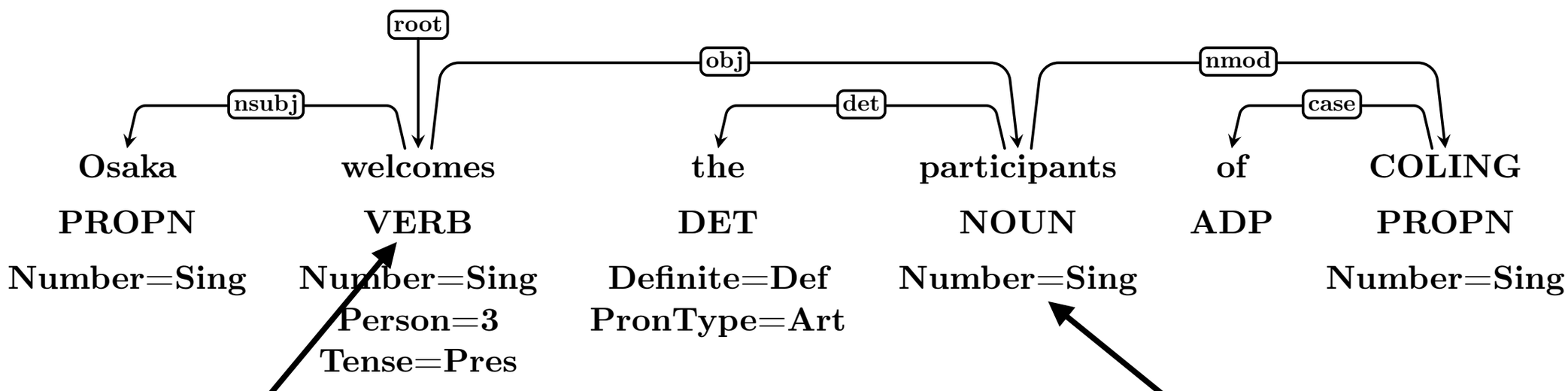


Part-of-speech tags Google

Universal Dependencies

http://universaldependencies.org





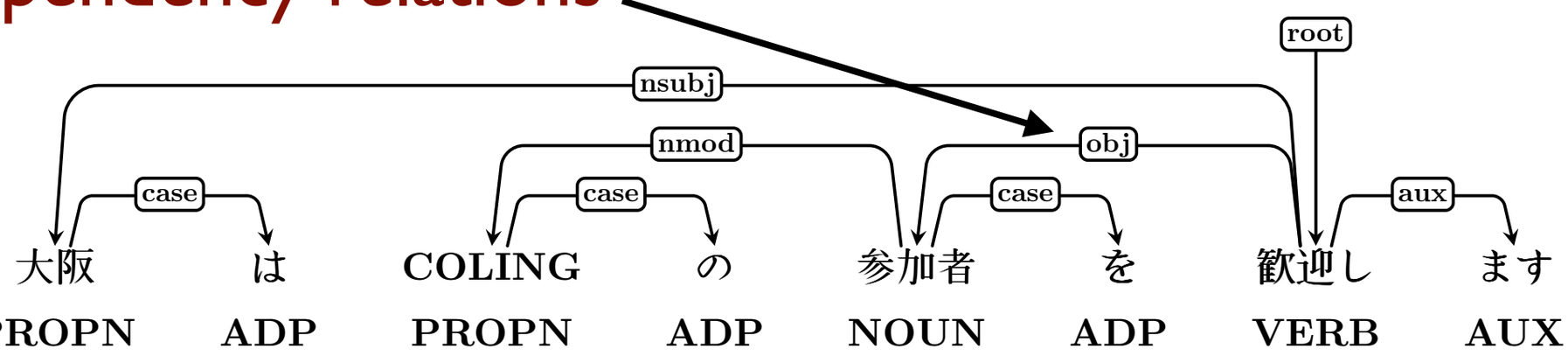
Part-of-speech tags Google

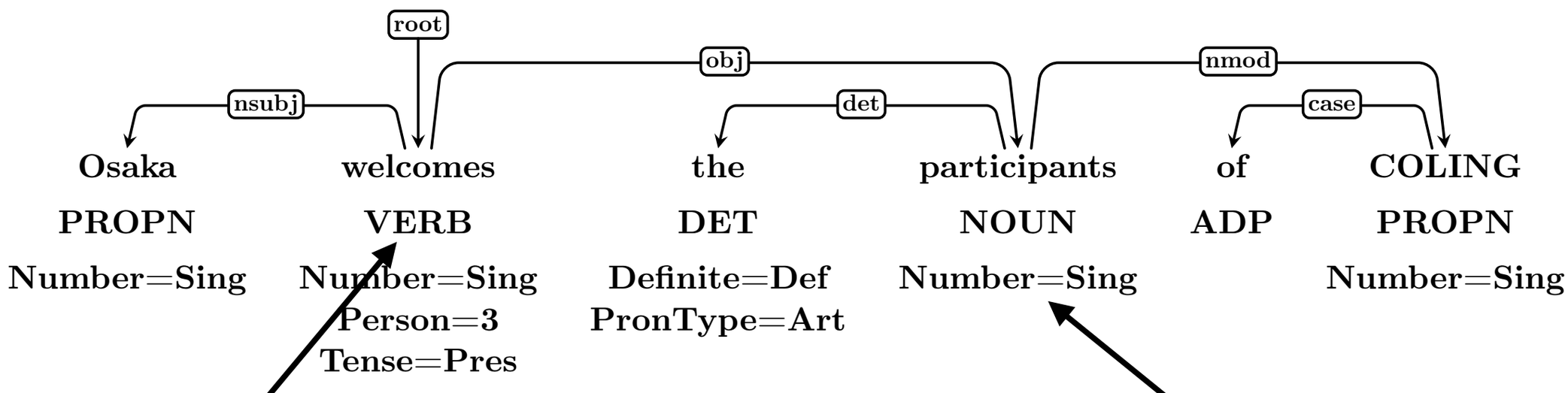
Morphological features ÚFAL

Universal Dependencies

http://universaldependencies.org

Dependency relations





Part-of-speech tags

Morphological features

Universal Dependencies

http://universaldependencies.org

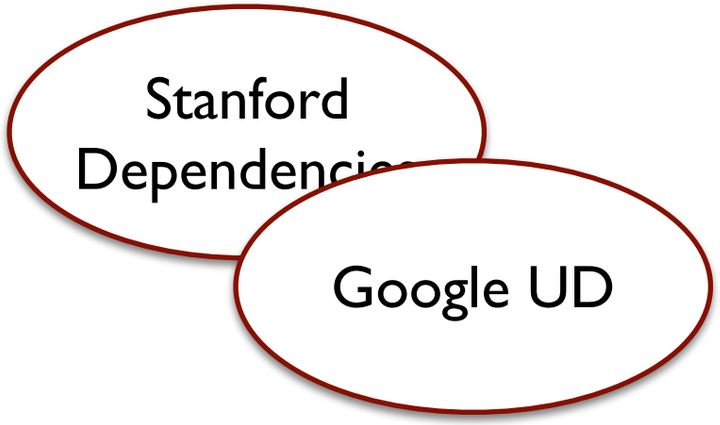
Universal Dependencies

http://universaldependencies.org

Stanford Dependencies

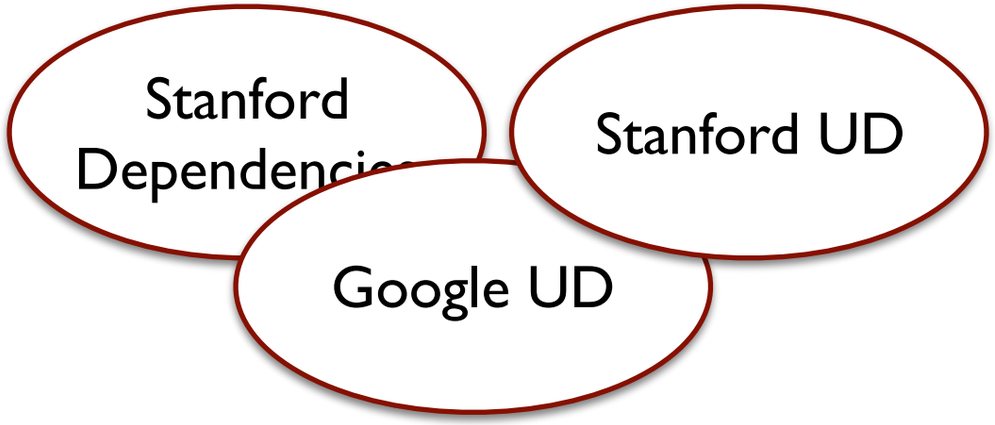
Universal Dependencies

http://universaldependencies.org



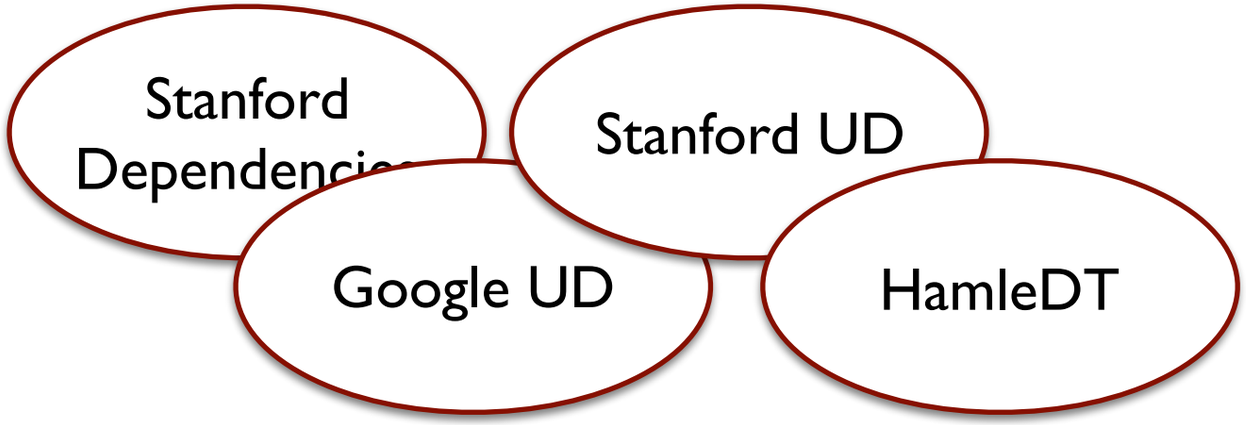
Universal Dependencies

http://universaldependencies.org



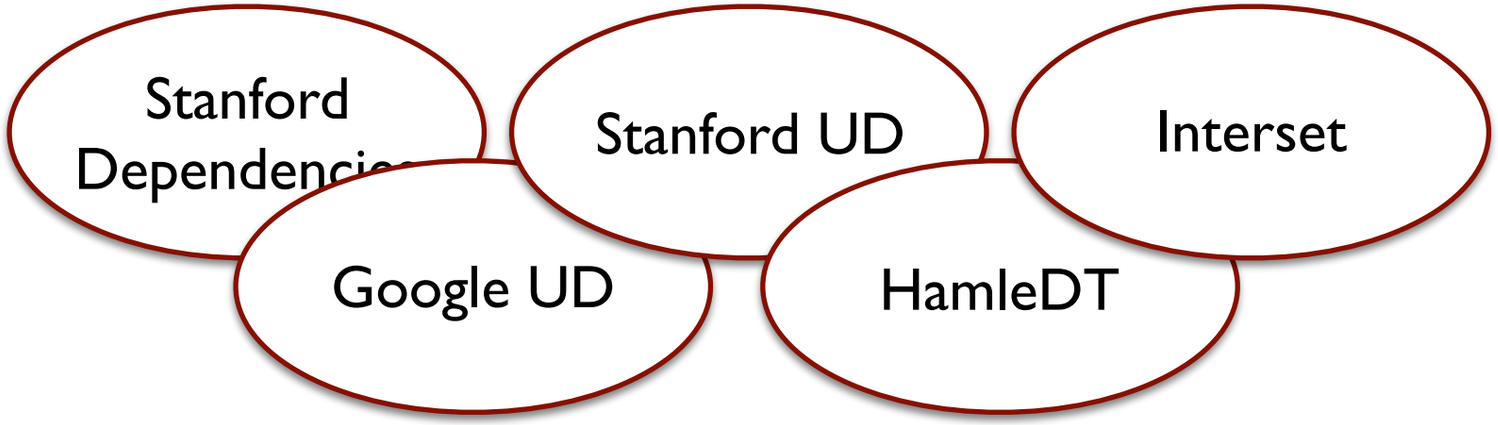
Universal Dependencies

http://universaldependencies.org



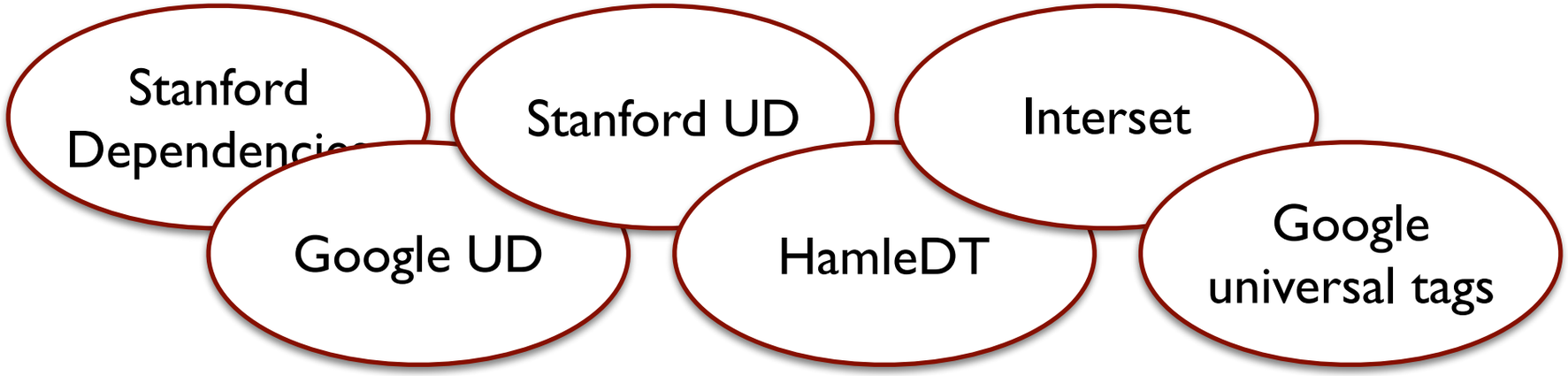
Universal Dependencies

http://universaldependencies.org



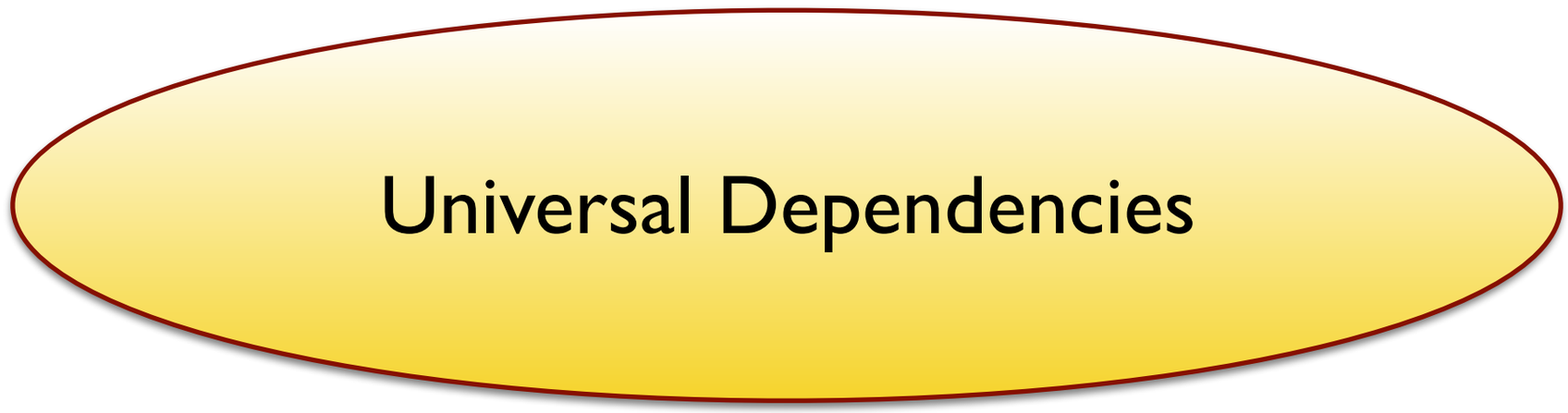
Universal Dependencies

http://universaldependencies.org



Universal Dependencies

http://universaldependencies.org



Universal Dependencies

http://universaldependencies.org

Universal Dependencies

Milestones:

- Kick-off meeting at EACL in Gothenburg, April 2014
- Release of annotation guidelines, v1, October 2014
- Releases of treebanks every 6 months, v1.0–v1.4
- Release of annotation guidelines, v2, December 2016 New!

Open community effort – anyone can contribute!

UD Treebanks

Ancient Greek	244K						
Ancient Greek-PROIEL	206K		-				
Arabic	242K		-				
Basque	121K						
Bulgarian	156K						
Buryat	9K		-				
Catalan	530K						
Chinese	123K						W
Coptic	5K						
Croatian	139K		-				W
Czech	1,503K						
Czech-CAC	493K						
Czech-CLTT	35K						
Danish	100K						
Dutch	209K		-				
Dutch-LassySmall	98K		-				W
English	254K						
English-ESL	97K						
English-LinES	82K						
Estonian	234K		-				
Faroese	132K		-				
Finnish	181K						W
Finnish-FTB	159K		-				
French	391K						W
Galician	138K						
Galician-TreeGal	24K						
German	293K		-				W
Gothic	56K		-				
Greek	59K						W
Hebrew	115K		-				
Hindi	351K		-				
Hungarian	42K						
Indonesian	121K		-				
Irish	23K						
Italian	272K						W
Japanese	92K		-				
Japanese-KTC	267K						
Kazakh	6K						W
Latin	47K		-				
Latin-ITTB	291K		-				
Latin-PROIEL	165K		-				
Latvian	20K		-				
Norwegian-Bokmaal	310K						
Old Church Slavonic	57K		-				
Persian	151K						
Polish	83K		-				
Portuguese	209K		-				
Portuguese-BR	298K		-				
Portuguese-Bosque	227K						
Romanian	218K						W
Russian	99K						W
Russian-SynTagRus	1,068K						
Sanskrit	1K		-				
Slovak	106K		-				
Slovenian	140K						
Slovenian-SST	29K						
Spanish	423K						W
Spanish-AnCora	547K						
Swedish	96K						
Swedish-LinES	79K						
Swedish Sign Language	<1K		-				
Tamil	8K		-				
Turkish	56K						
Ukrainian	1K						
Uyghur	6K		-				
Vietnamese	43K		-				

December 13, 2016:

- 47 languages
- 64 treebanks
- 145 contributors
- 7000+ downloads

UD Treebanks

Ancient Greek	244K						
Ancient Greek-PROIEL	206K		-				
Arabic	242K		-				
Basque	121K						
Bulgarian	156K						
Buryat	9K		-				
Catalan	530K						
Chinese	123K						
Coptic	5K						
Croatian	139K		-				
Czech	1,503K						
Czech-CAC	493K						
Czech-CLTT	35K						
Danish	100K						
Dutch	209K		-				
Dutch-LassySmall	98K		-				
English	254K						
English-ESL	97K						
English-LinES	82K						
Estonian	234K		-				
Faroese	132K		-				
Finnish	181K						
Finnish-FTB	159K		-				
French	391K						
Galician	138K						
Galician-TreeGal	24K						
German	293K		-				
Gothic	56K		-				
Greek	59K						
Hebrew	115K		-				
Hindi	351K		-				
Hungarian	42K						
Indonesian	121K		-				
Irish	23K						
Italian	272K						
Japanese	92K		-				
Japanese-KTC	267K						
Kazakh	6K						
Latin	47K		-				
Latin-ITTB	291K		-				
Latin-PROIEL	165K		-				
Latvian	20K		-				
Norwegian-Bokmaal	310K						
Old Church Slavonic	57K		-				
Persian	151K						
Polish	83K		-				
Portuguese	209K		-				
Portuguese-BR	298K		-				
Portuguese-Bosque	227K						
Romanian	218K						
Russian	99K						
Russian-SynTagRus	1,068K						
Sanskrit	1K		-				
Slovak	106K		-				
Slovenian	140K						
Slovenian-SST	29K						
Spanish	423K						
Spanish-AnCora	547K						
Swedish	96K						
Swedish-LinES	79K						
Swedish Sign Language	<1K		-				
Tamil	8K		-				
Turkish	56K						
Ukrainian	1K						
Uyghur	6K		-				
Vietnamese	43K		-				

December 13, 2016:

- 47 languages
- 64 treebanks
- 145 contributors
- 7000+ downloads

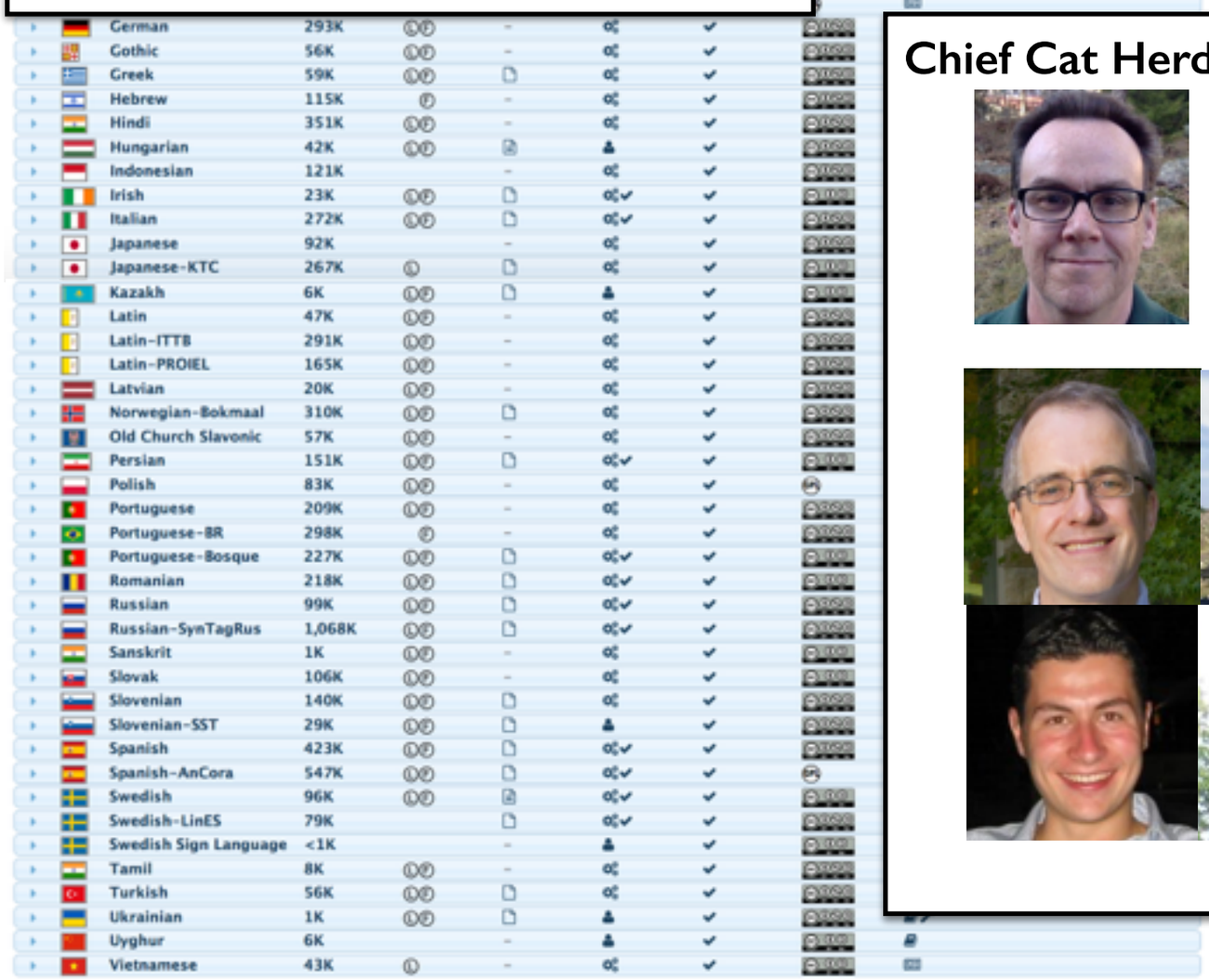
Chief Cat Herder

Release and Documentation Task Force

Universal Guidelines Group

UD Japanese

Masayuki Asahara
Hiroshi Kanayama
Yuji Matsumoto
Yusuke Miyao
Shinsuke Mori
Takaaki Tanaka
Sumire Uematsu



German	293K	
Gothic	56K	
Greek	59K	
Hebrew	115K	
Hindi	351K	
Hungarian	42K	
Indonesian	121K	
Irish	23K	
Italian	272K	
Japanese	92K	
Japanese-KTC	267K	
Kazakh	6K	
Latin	47K	
Latin-ITTB	291K	
Latin-PROIEL	165K	
Latvian	20K	
Norwegian-Bokmaal	310K	
Old Church Slavonic	57K	
Persian	151K	
Polish	83K	
Portuguese	209K	
Portuguese-BR	298K	
Portuguese-Bosque	227K	
Romanian	218K	
Russian	99K	
Russian-SynTagRus	1,068K	
Sanskrit	1K	
Slovak	106K	
Slovenian	140K	
Slovenian-SST	29K	
Spanish	423K	
Spanish-AnCora	547K	
Swedish	96K	
Swedish-LinES	79K	
Swedish Sign Language	<1K	
Tamil	8K	
Turkish	56K	
Ukrainian	1K	
Uyghur	6K	
Vietnamese	43K	

December 13, 2016:

- 47 languages
- 64 treebanks
- 145 contributors
- 7000+ downloads

Chief Cat Herder

Release and Documentation Task Force

Universal Guidelines Group

A guided tour of the UD framework

A guided tour of the UD framework

Why such weird dependency trees?

A guided tour of the UD framework

Why such weird dependency trees?

Dubious linguistics?

Crappy parsing?

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 ... niens .

le ... hien .

DET N ... 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 le DET Definite=Def Gender=Masc Number=Sing

N Gend Num

niens hien OUN er=Masc er=Plur

. . PUNCT

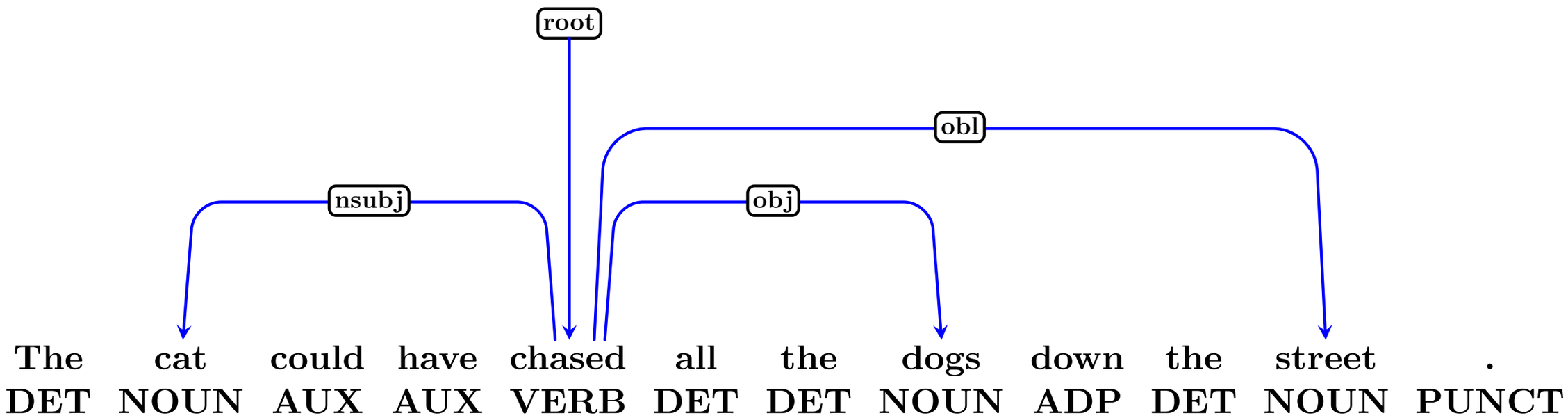
Lexical	Inflectional Nominal	Inflectional Verbal
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	Number	Tense
Reflex	Case	Aspect
Foreign	Definite	Voice
Abbr	Degree	Evident
		Polarity
		Person
		Polite

- 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

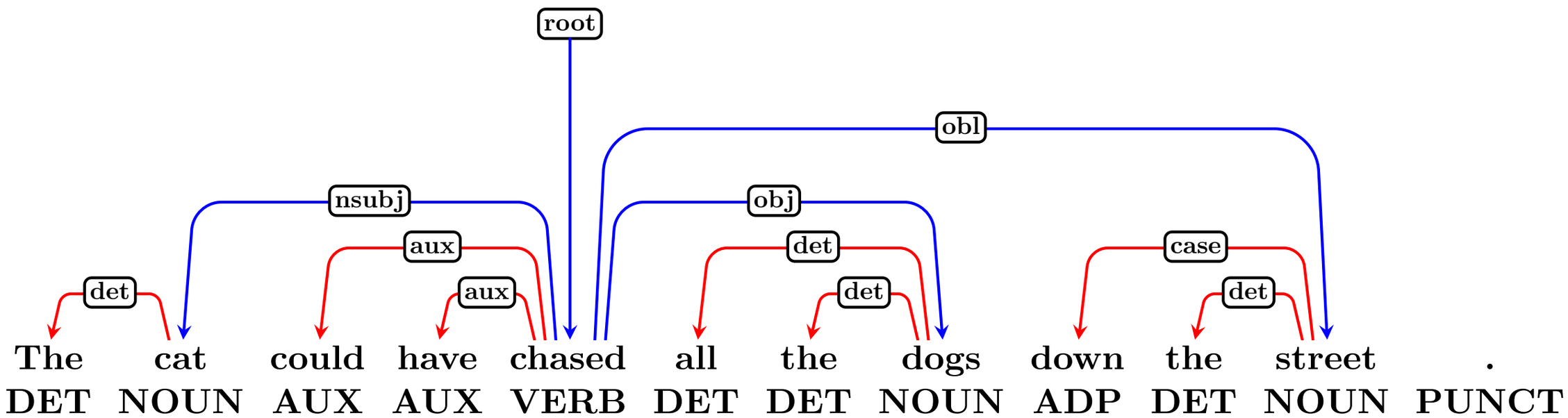
The	**cat**	**could**	**have**	**chased**	**all**	**the**	**dogs**	**down**	**the**	**street**	**.**
DET	**NOUN**	**AUX**	**AUX**	**VERB**	**DET**	**DET**	**NOUN**	**ADP**	**DET**	**NOUN**	**PUNCT**

Syntax



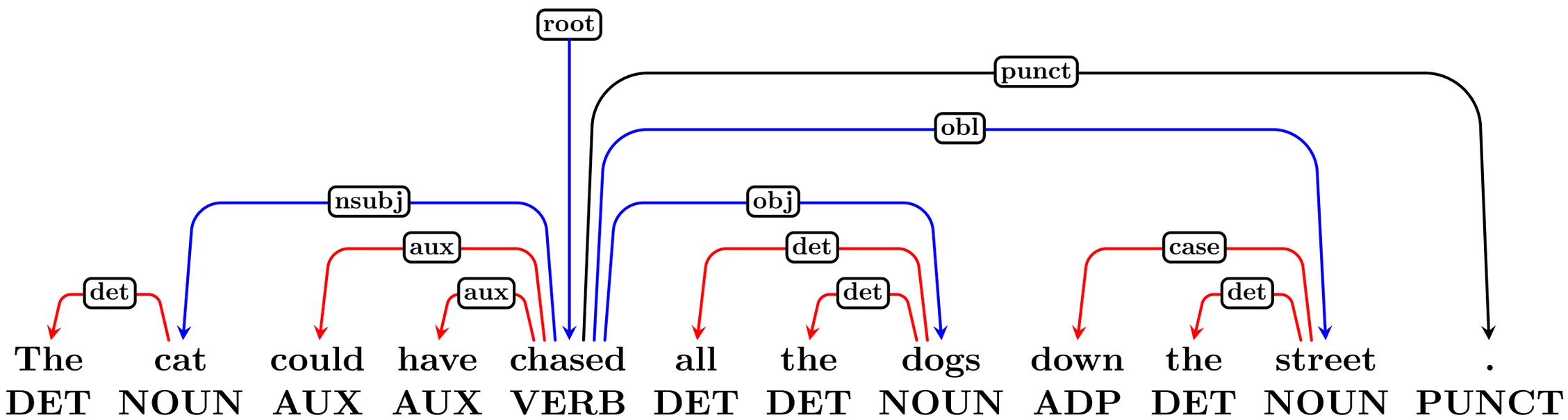
- Content words are related by dependency relations

Syntax

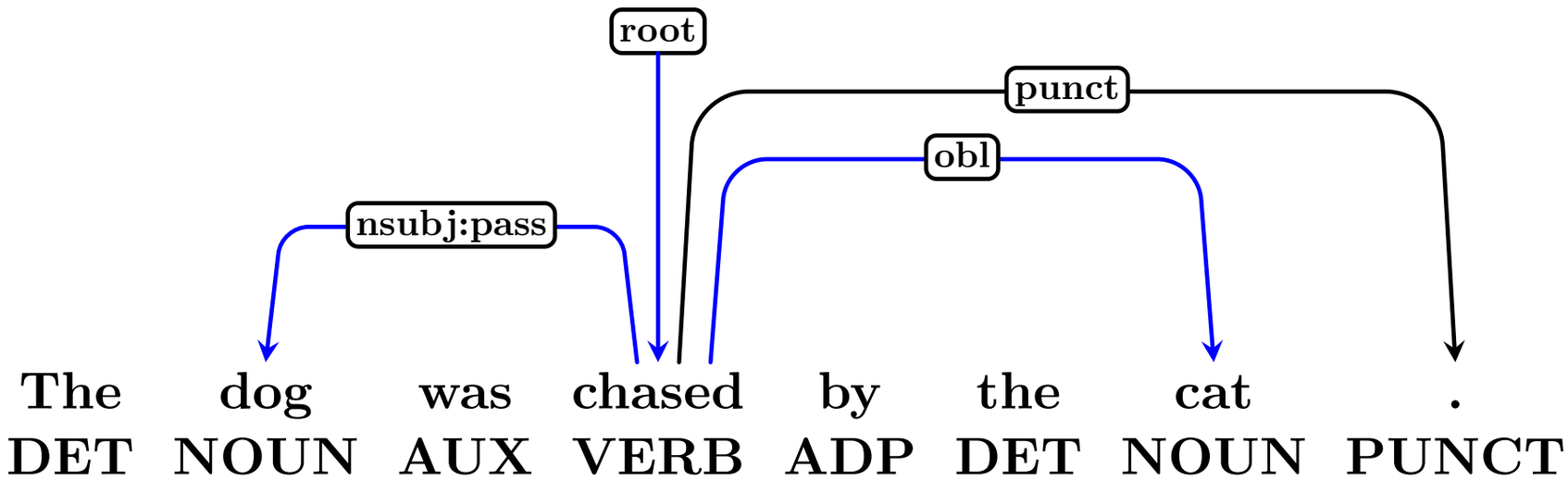


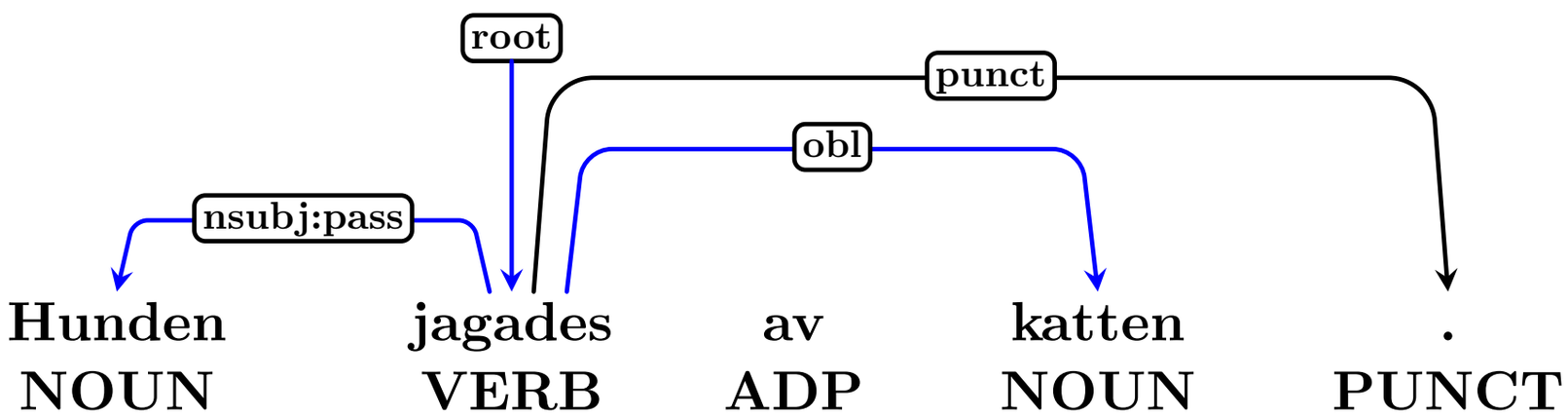
- Content words are related by dependency relations
- Function words attach to the content word they modify

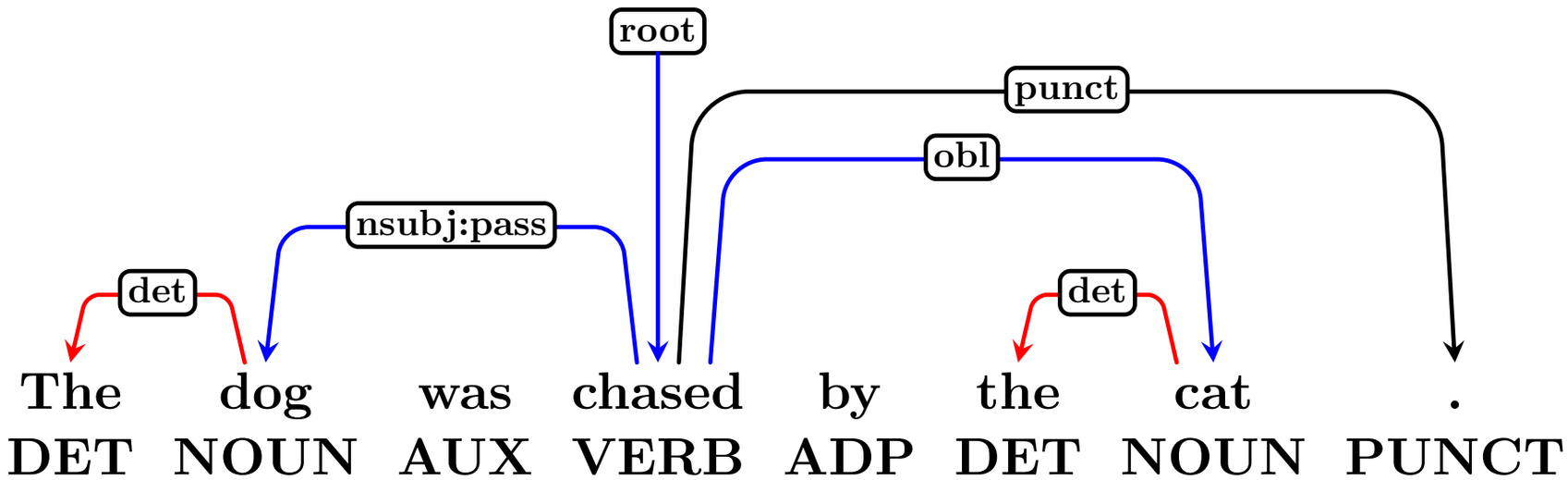
Syntax

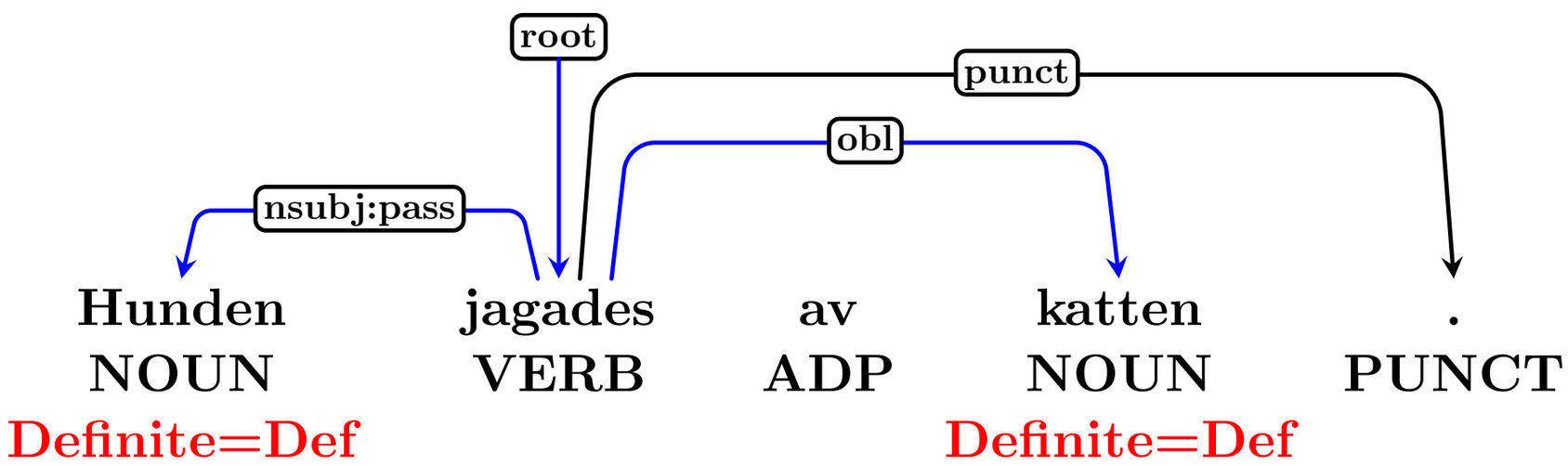


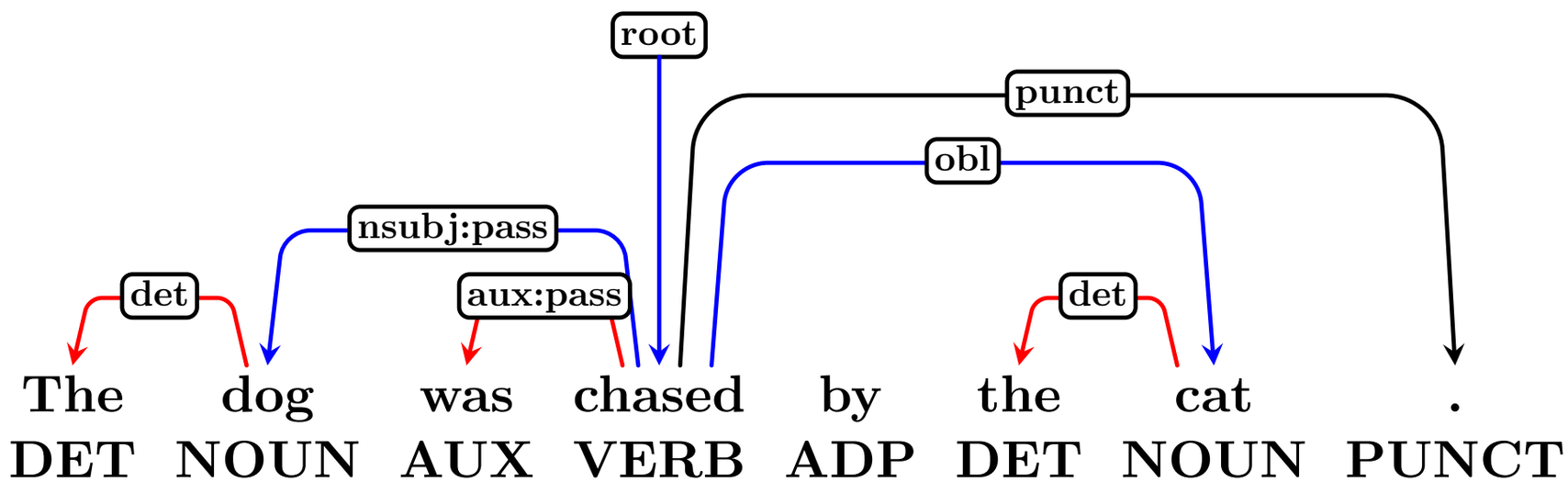
- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

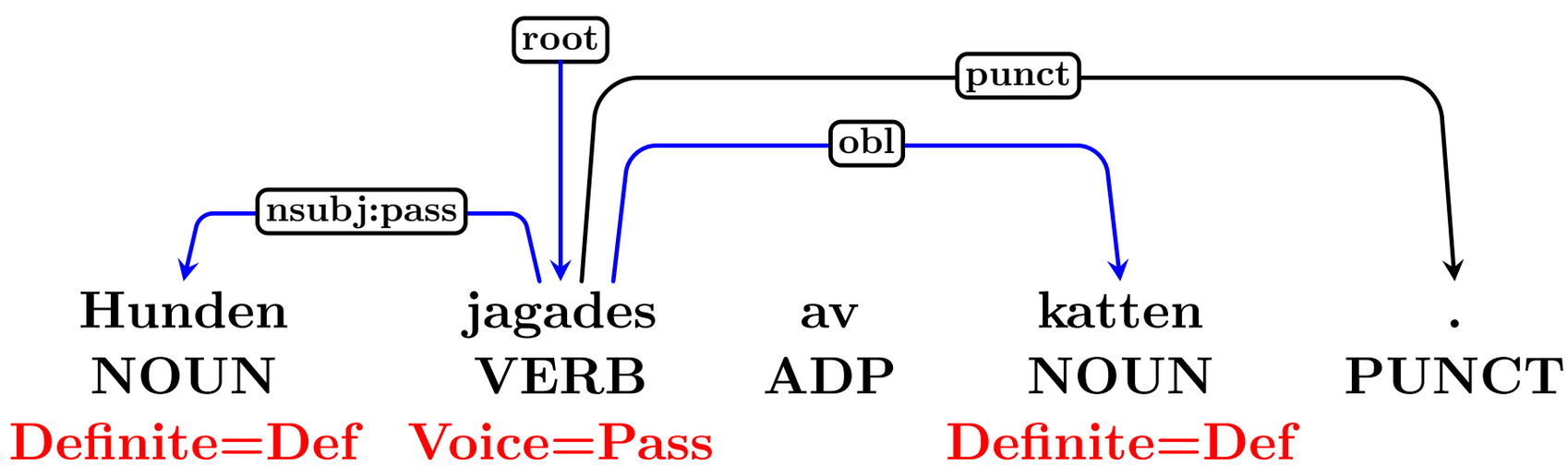

root
punct
obl
nsubj:pass
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT

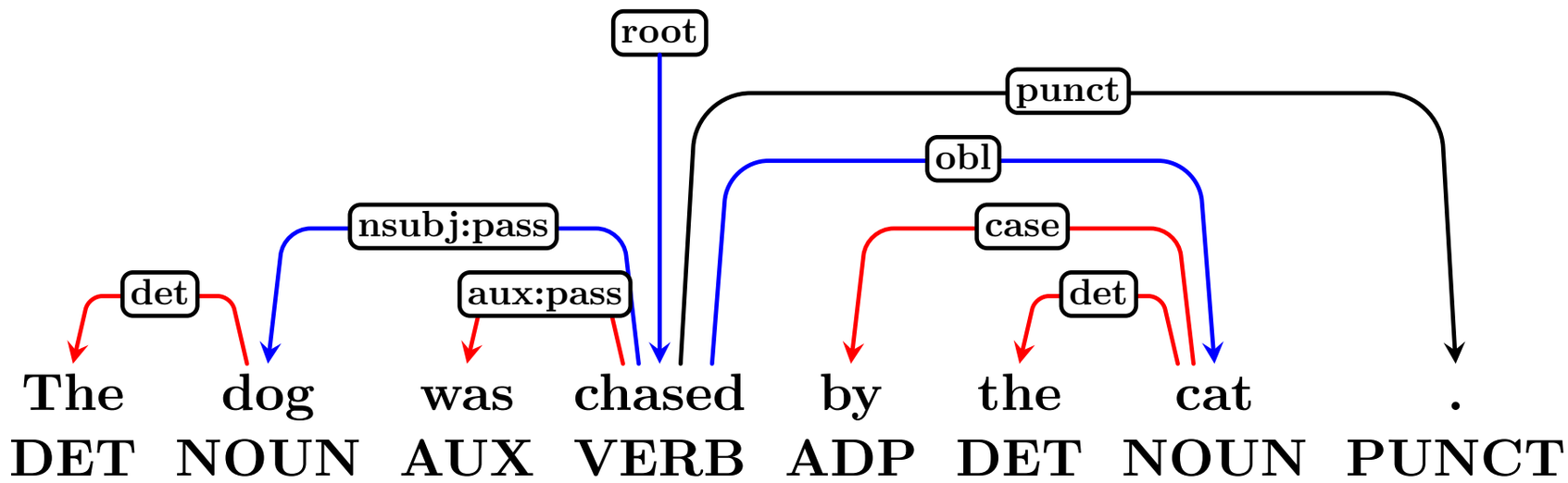

root
punct
obl
nsubj:pass
Hunden jagades av katten .
NOUN VERB ADP NOUN PUNCT

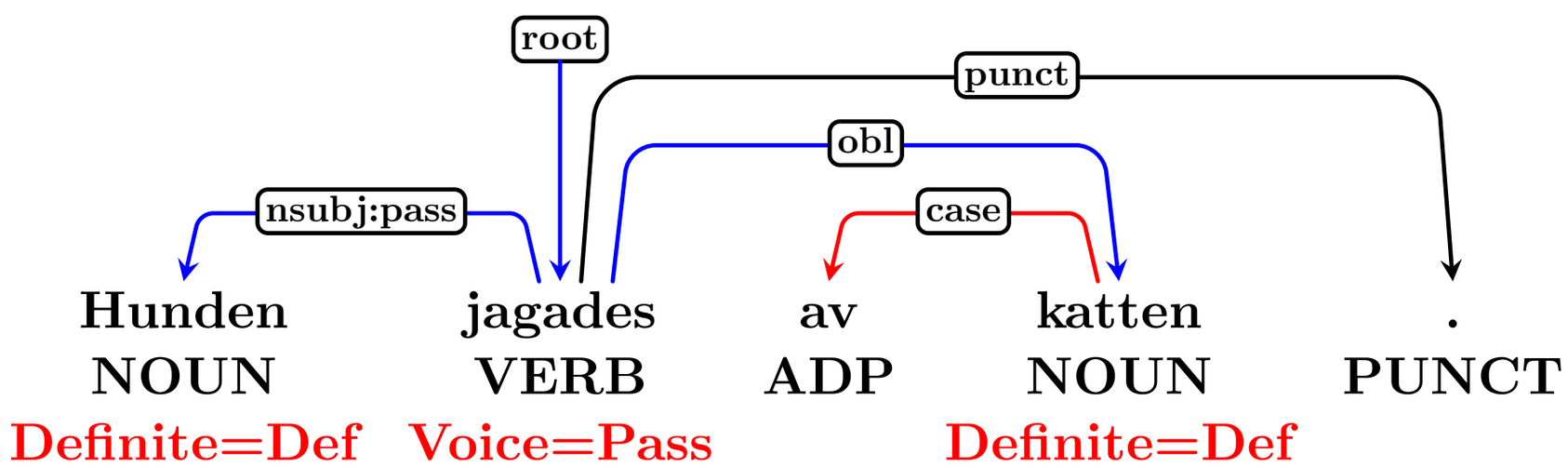

root
punct
obl
nsubj:pass
det
det
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT


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 complements 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 complements 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

A Two-Level Architecture

Universal relations

- Broad categories to allow cross-linguistic comparison

Language-specific relations

- Subtypes to capture language-specific phenomena

A Two-Level Architecture

Universal relations

- Broad categories to allow cross-linguistic comparison

Language-specific relations

- Subtypes to capture language-specific phenomena

Universal	Subtype

A Two-Level Architecture

Universal relations

- Broad categories to allow cross-linguistic comparison

Language-specific relations

- Subtypes to capture language-specific phenomena

Universal	Subtype
acl	acl:relcl

A Two-Level Architecture

Universal relations

- Broad categories to allow cross-linguistic comparison

Language-specific relations

- Subtypes to capture language-specific phenomena

Universal	Subtype
acl	acl:relcl
compound	compound:prt

A Two-Level Architecture

Universal relations

- Broad categories to allow cross-linguistic comparison

Language-specific relations

- Subtypes to capture language-specific phenomena

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

http://universaldependencies.org

home edit page issue tracker

This page pertains to UD version 2.

Universal Dependencies v2

This is the online documentation for Universal Dependencies, version 2 (2016-12-01). **Note:** The treebanks listed below still follow the v1 guidelines available here.

http://universaldependencies.org

home edit page issue tracker

This page pertains to UD version 2.

Universal D

Executive summary of c

- Tokenization and
- Morphology
  - General pr
  - Universal
  - Universal
  - Language
  - Conversio
- Syntax
  - General pr
  - Basic dep
    - Sim
    - No
    - Co
    - Oth
  - Enhanced
  - Universal
  - Language
- CoNLL-U format

home edit page issue tracker

This page pertains to UD version 2.

Simple Clauses

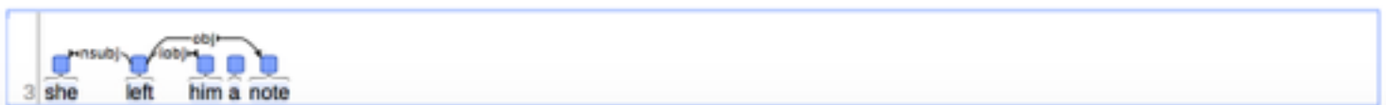
The UD annotation assumes the clause as one of the basic structures that we expect to find in all languages. A simple clause minimally consists of a predicate together with its core argument dependents, but may be extended with oblique modifiers. Core arguments are typically nominals, while oblique modifiers are either (oblique) nominals or adverbial modifiers. (In complex clauses, both core arguments and oblique modifiers can also be realized as subordinate clauses.) Finally, the predicate may be associated with function words that express different types of grammatical information such as tense, mood, aspect, voice, evidentiality, or type of subordination.

Intransitive and Transitive Clauses

In most clauses, the predicate takes the form of a verb, which may be intransitive or transitive.

An intransitive verb takes a single argument (usually referred to as S in the literature on linguistic typology) with the `nsubj` relation. A transitive verb in addition takes an argument with the `obj` relation. When deciding which relation to use with which argument in a transitive clause, the `nsubj` relation should be used with the argument that most resembles the proto-agent (often called A in linguistic typology) and that satisfies additional language-internal criteria for subjecthood based on case-marking, agreement and/or linear position with respect to the predicate. The `obj` relation should be used for the argument that most resembles the proto-patient (often called O or P in linguistic typology) and that satisfies relevant language-internal criteria. Note that, while case-marking (whether morphological or analytic) can provide important evidence in specific languages, case alignment should not be used to decide the assignment of core argument roles. Thus, in ergative languages, the patient-like argument of a transitive verb (O/P) will take the the `obj` relation despite the fact that it carries the same case marking as the `nsubj` argument (S) of an intranstive verb.

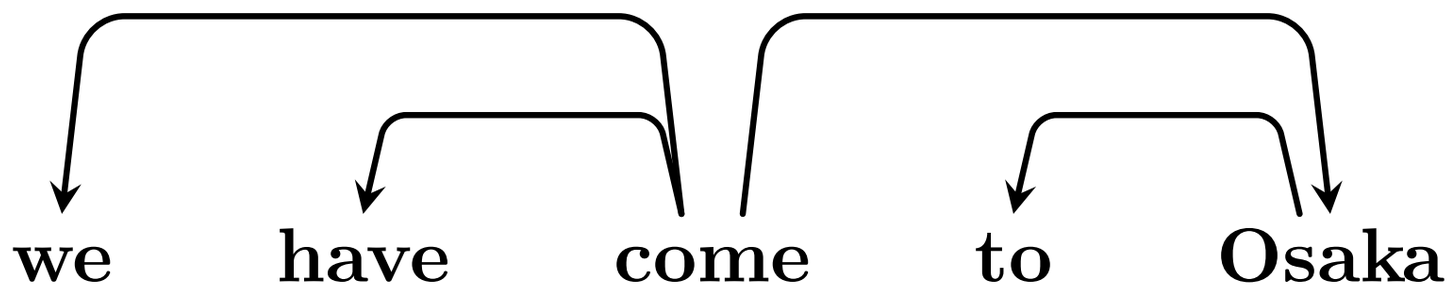
Some languages allow extended transitive clauses, where more than two dependents are realized as core arguments. The additional core arguments then receive the `iobj` relation (for "indirect object"), while the `obj` relation is reserved for the argument most patient-like non-subject argument. The criterion for deciding whether an additional dependent is a core argument is whether it has the typical encoding of a core argument with respect to case-marking, agreement and word order. For example, the English double object construction qualifies as an extended transitive clause because all three nominals appear without prepositions:



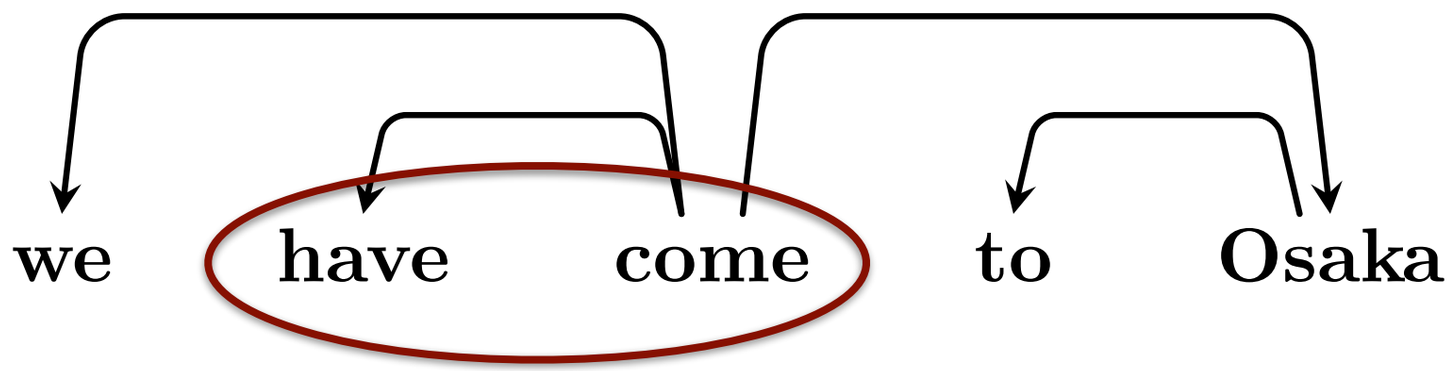
This is the online documentation for Universal Dependencies, version 2 (2016-12-01). **Note:** The treebanks listed below still follow the v1 guidelines available here.

Why such weird dependency trees?

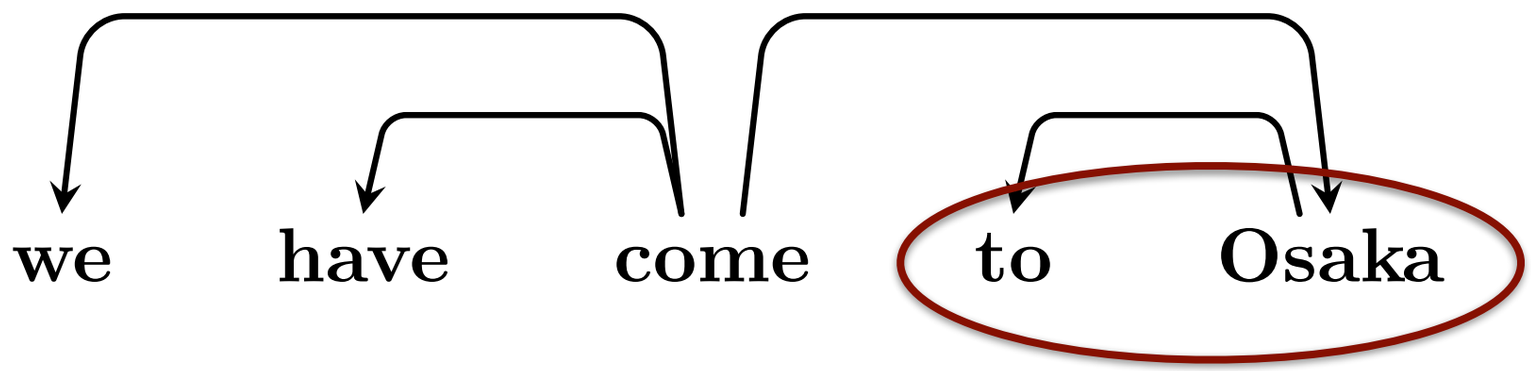
“Content-Head Dependencies”



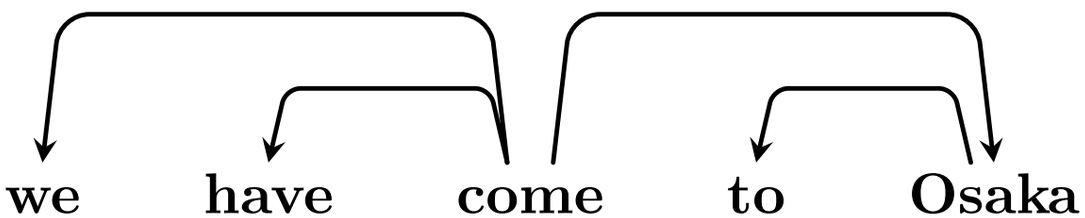
“Content-Head Dependencies”



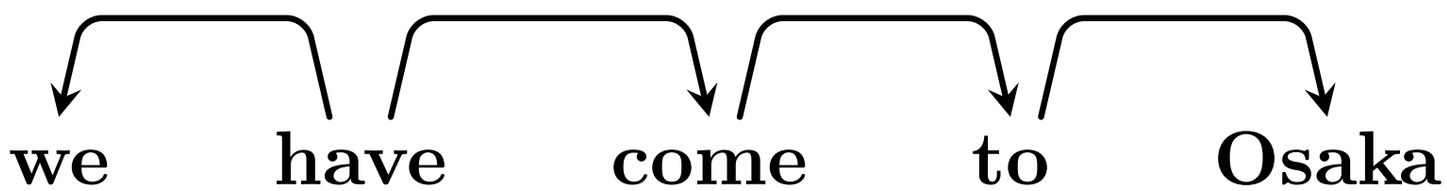
“Content-Head Dependencies”



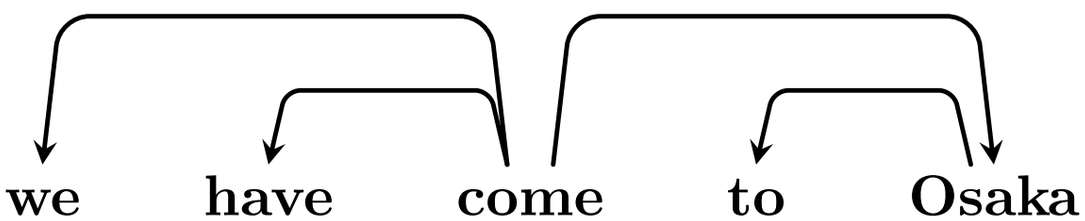
“Content-Head Dependencies”



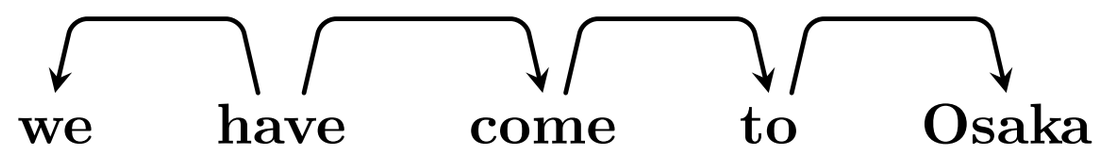
“Function-Head Dependencies”



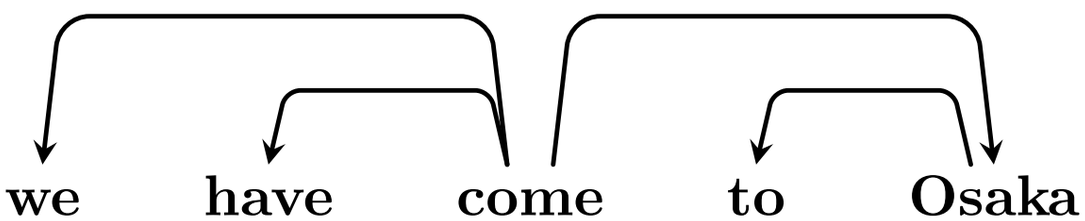
"Content-Head Dependencies"



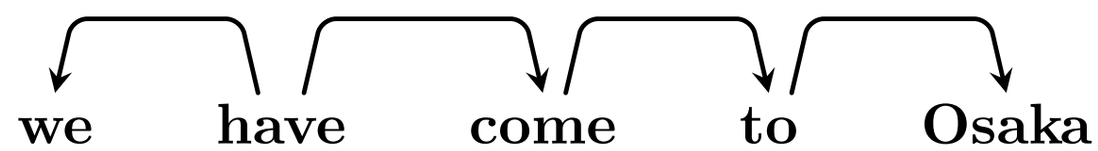
"Function-Head Dependencies"



“Content-Head Dependencies”



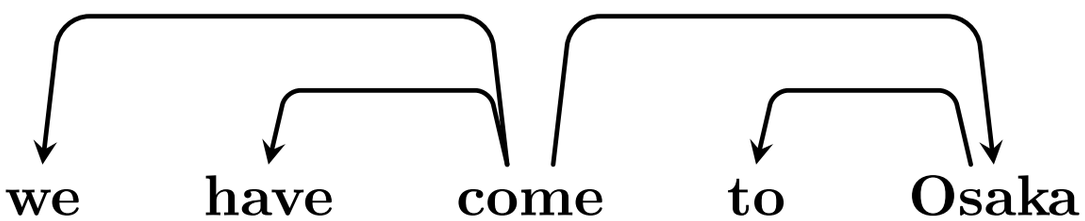
“Function-Head Dependencies”



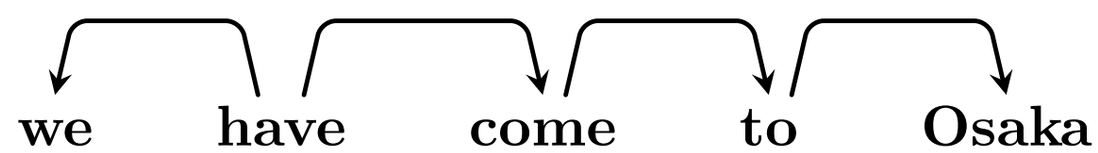
Dubious Linguistics?

“Such an approach to the syntax of natural languages is contrary to most work in theoretical syntax in the past 35 years, regardless of whether this work is constituency- or dependency-based.” (Groß and Osborne, 2015)

"Content-Head Dependencies"



"Function-Head Dependencies"



Dubious Linguistics?

"Such an approach to the syntax of natural languages is contrary to most work in theoretical syntax in the past 35 years, regardless of whether this work is constituency- or dependency-based." (Groß and Osborne, 2015)

Crappy Parsing?

"It is now fairly well known that, while dependency representations in which content words are made heads tend to help semantically oriented downstream applications, dependency parsing numbers are higher if you make auxiliary verbs heads […] and if you make prepositions the head of prepositional phrases." (De Marneffe et al., 2014)

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.

What is a head?

Semantic functor . . .	V+NP (V)	P+NP (P)	NP+VP (VP)	Det+N (Det)	Aux+VP (Aux)	Comp+S (Comp)
(A) Semantic argument	*	*	*	*	*	*
(B) Determinant of concord	(*)	·	*	*	·	·
(C) Morphosyntactic locus	=	=	=	*	=	*
(D) Subcategorizand	=	·	·	=	·	=
(E) Governor	=	=	=	·	=	·
(F) Distributional equivalent	=	·	·	*	*	*
(G) Obligatory	=	=	=	*	*	*
(H) Ruler	=	·	·	*	*	=

Key: = same as entry for 'Semantic functor'
* different from entry for 'Semantic functor'

Zwicky (1985), summarised by Hudson (1987)

Why choose one?

Why choose one?

Head properties may be shared by several elements

- So neither content-head nor function-head can be quite right

Why choose one?

Head properties may be shared by several elements

- So neither content-head nor function-head can be quite right

Linguistic theories capture this in different ways

- Lexical vs. functional heads (Chomsky, 1995)
- Surface syntax vs. deep syntax (Sgall et al., 1986; Mel'čuk, 1988)
- Dissociated nucleus (Tesnière, 1959)

Why choose one?

Head properties may be shared by several elements

- So neither content-head nor function-head can be quite right

Linguistic theories capture this in different ways

- Lexical vs. functional heads (Chomsky, 1995)
- Surface syntax vs. deep syntax (Sgall et al., 1986; Mel'čuk, 1988)
- Dissociated nucleus (Tesnière, 1959)

What about UD?

UD Syntax

UD Syntax

UD representations are mono-stratal – single tree

- Facilitates annotation, parsing and downstream tasks

UD Syntax

UD representations are mono-stratal – single tree

- Facilitates annotation, parsing and downstream tasks

Tree structure primarily reflects lexical dependencies

- Brings out parallelism between typologically diverse languages
- Reveals predicate-argument structure for downstream tasks

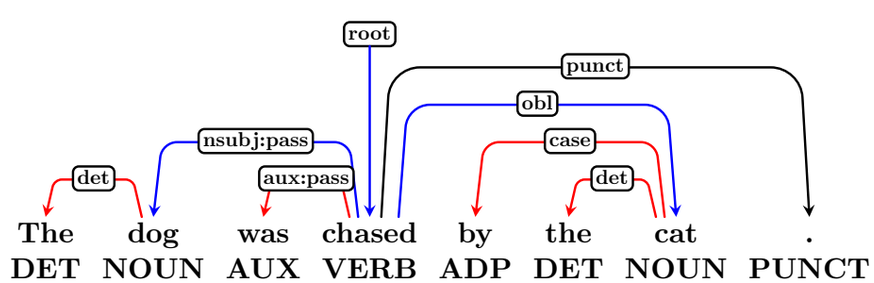
UD Syntax

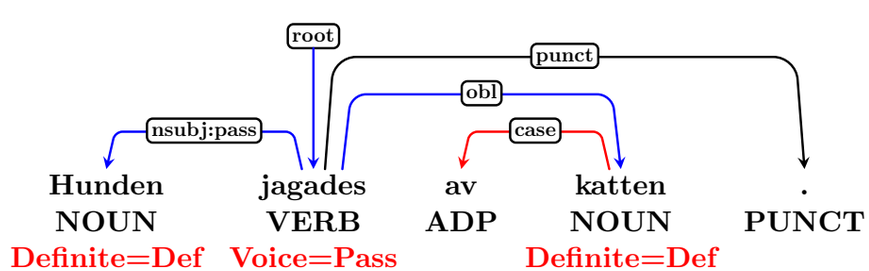
UD representations are mono-stratal – single tree

- Facilitates annotation, parsing and downstream tasks

Tree structure primarily reflects lexical dependencies

- Brings out parallelism between typologically diverse languages
- Reveals predicate-argument structure for downstream tasks





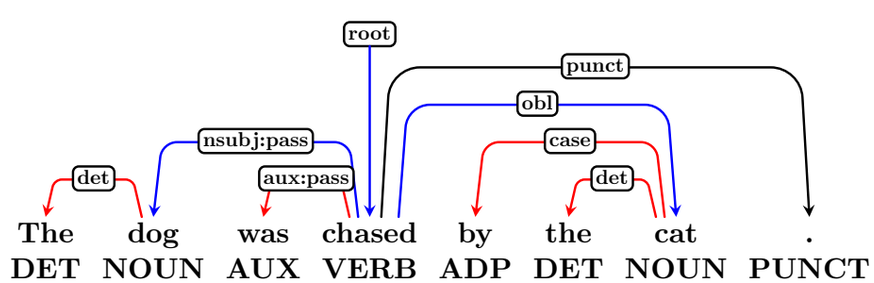
UD Syntax

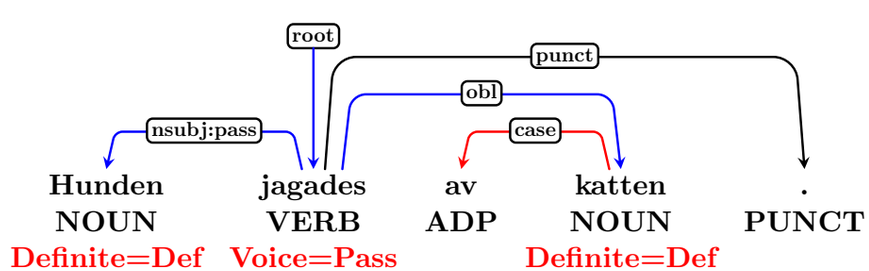
UD representations are mono-stratal – single tree

- Facilitates annotation, parsing and downstream tasks

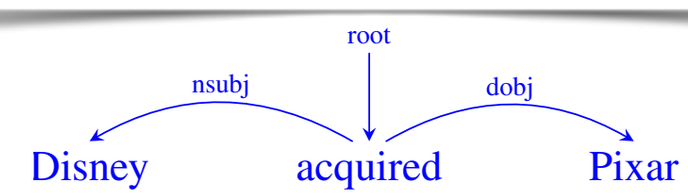
Tree structure primarily reflects lexical dependencies

- Brings out parallelism between typologically diverse languages
- Reveals predicate-argument structure for downstream tasks





Reddy et al. (2016) Transforming Dependency Structures to Logical Forms for Semantic Parsing



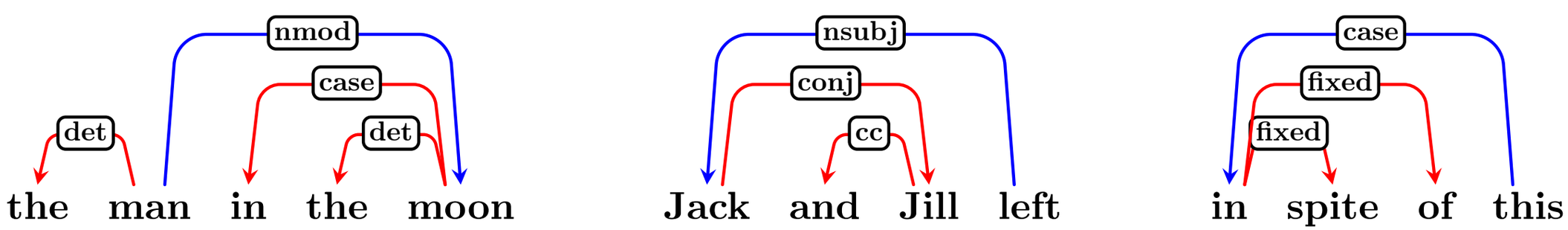
(**nsubj** (**dobj** **acquired** **Pixar**) **Disney**)

$$\lambda z.\exists xy.\text{acquired}(z_e) \wedge \text{Pixar}(y_a) \wedge \text{Disney}(x_a) \wedge \text{arg}_1(z_e, x_a) \wedge \text{arg}_2(z_e, y_a)$$

UD Syntax

Other relations encoded in labels – not tree structure

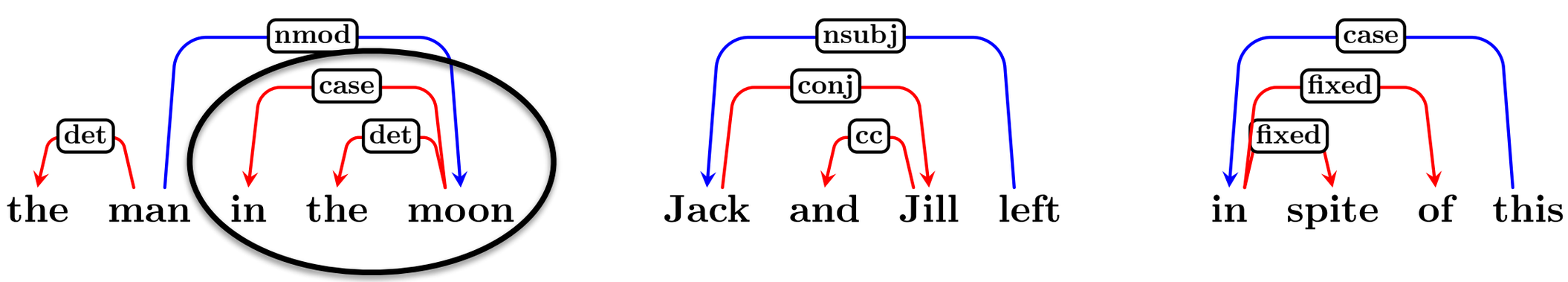
- Functional relations link functional heads to lexical heads
- Coordination relations link equivalent heads/dependents
- Multiword relations link elements of lexicalized expressions



UD Syntax

Other relations encoded in labels – not tree structure

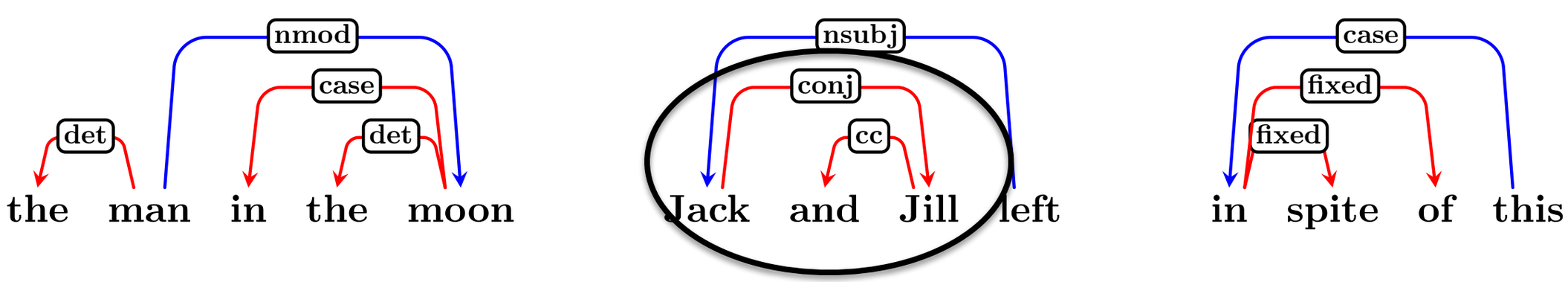
- Functional relations link functional heads to lexical heads
- Coordination relations link equivalent heads/dependents
- Multiword relations link elements of lexicalized expressions



UD Syntax

Other relations encoded in labels – not tree structure

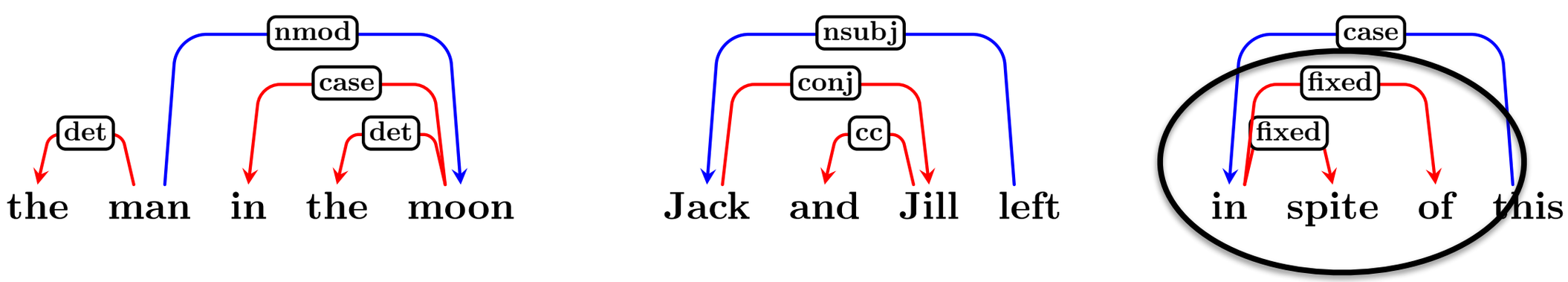
- Functional relations link functional heads to lexical heads
- Coordination relations link equivalent heads/dependents
- Multiword relations link elements of lexicalized expressions

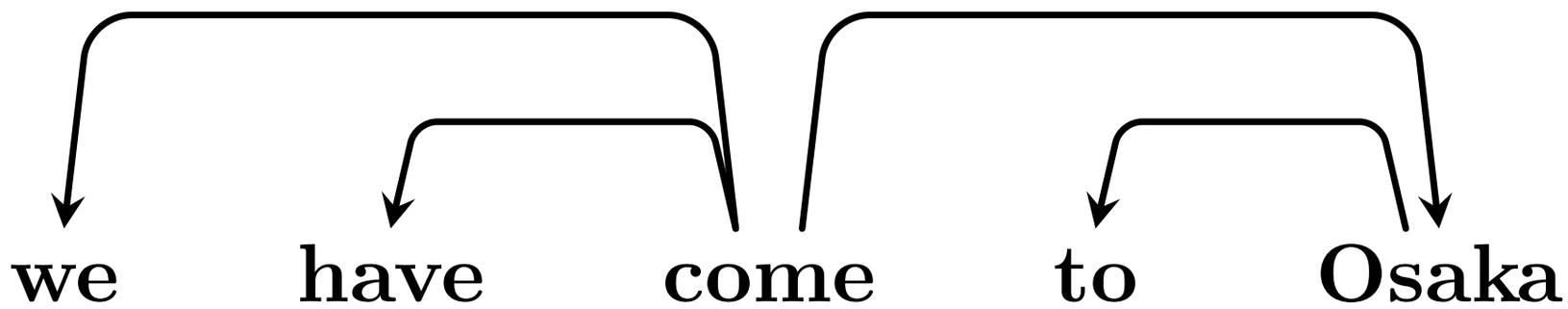


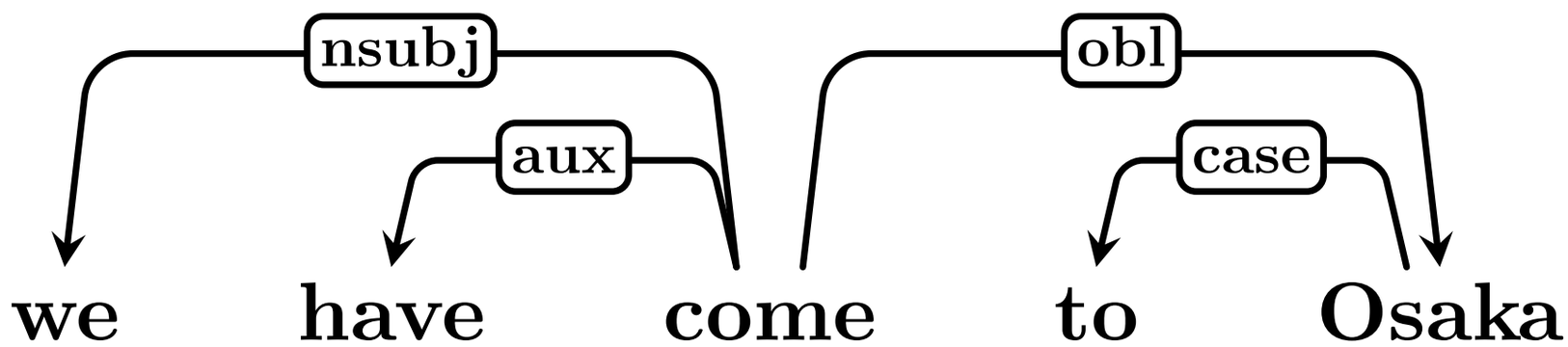
UD Syntax

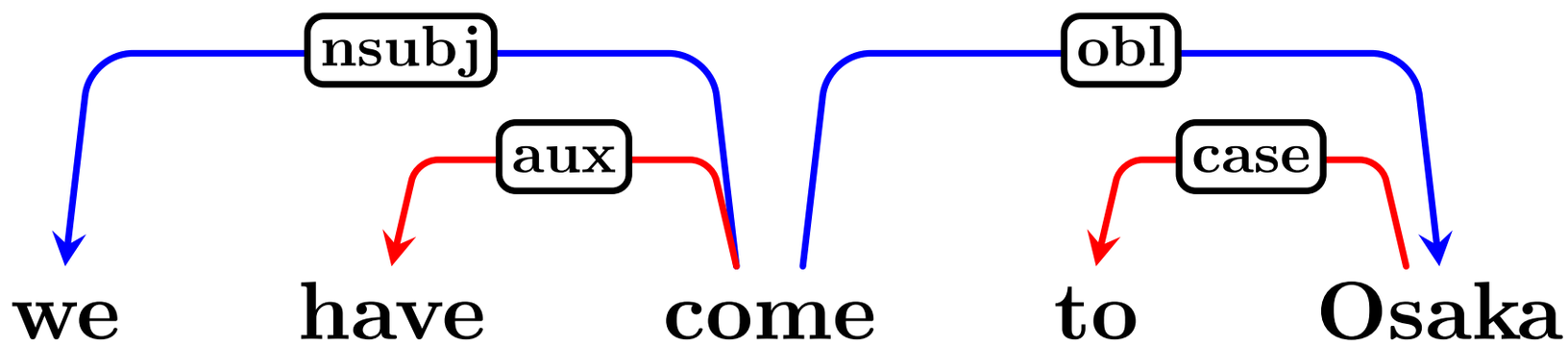
Other relations encoded in labels – not tree structure

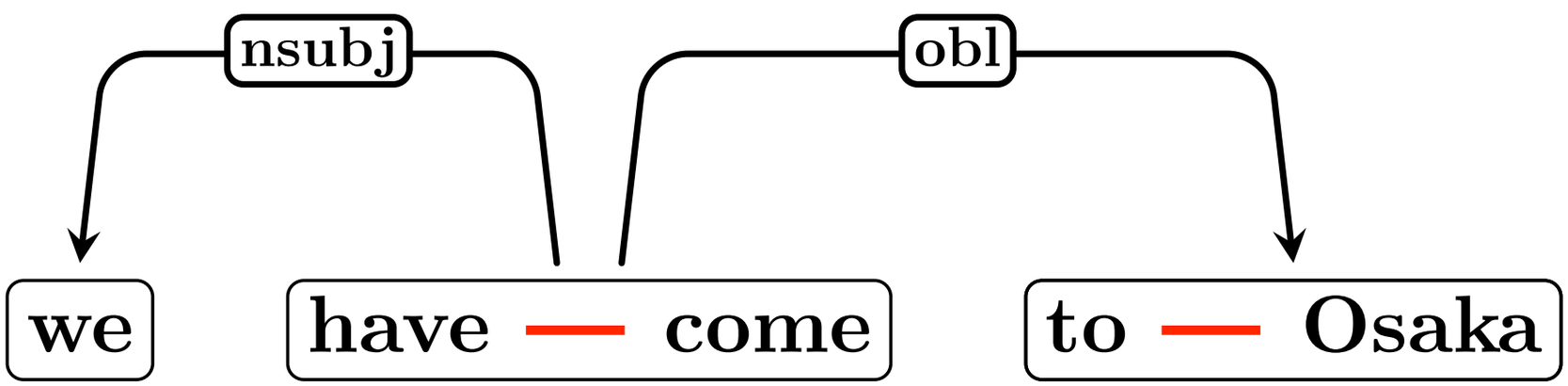
- Functional relations link functional heads to lexical heads
- Coordination relations link equivalent heads/dependents
- Multiword relations link elements of lexicalized expressions



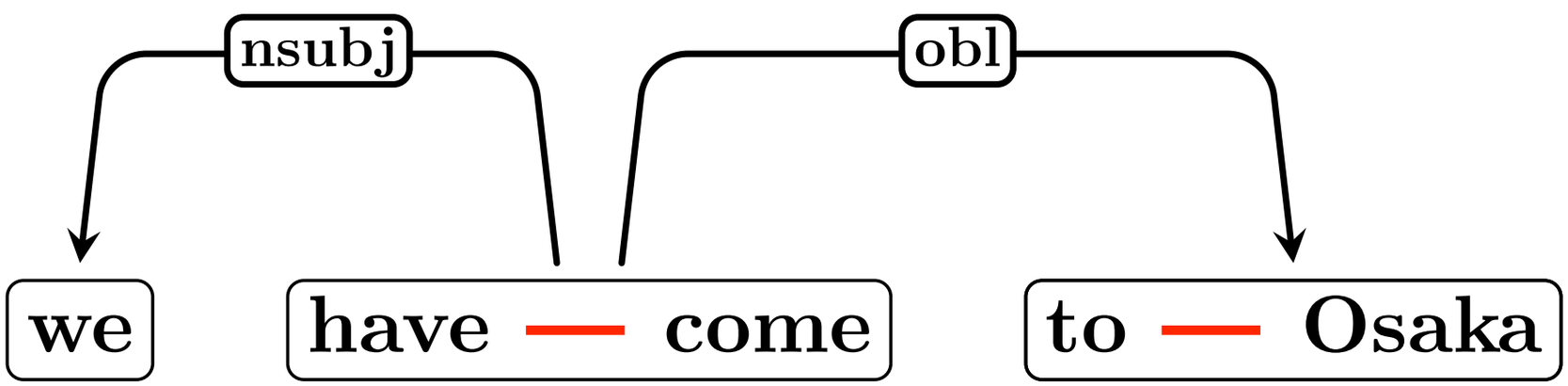

we
have
come
to
Osaka


nsubj
obl
aux
case
we
have
come
to
Osaka

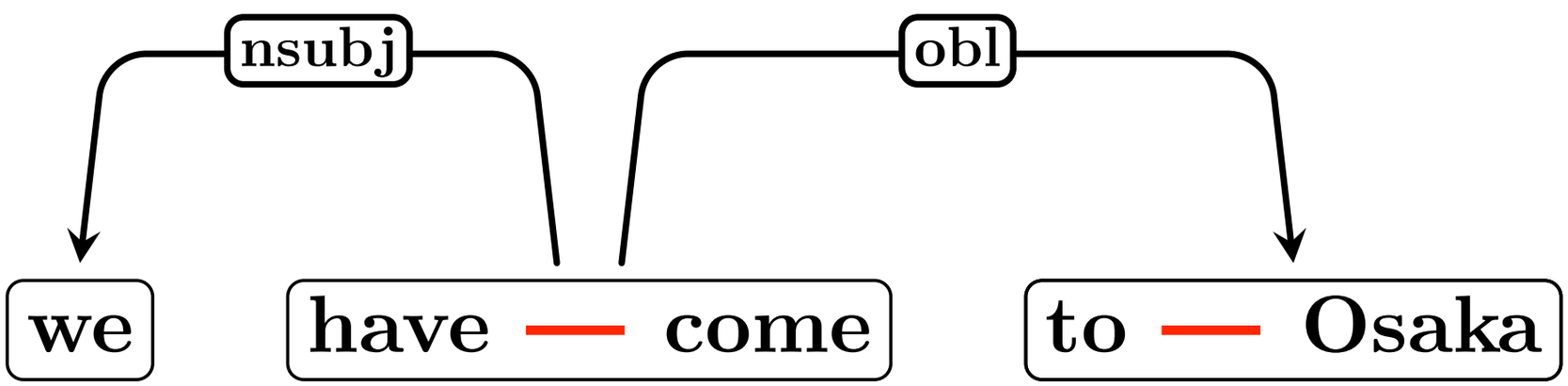

nsubj
obl
aux
case
we
have
come
to
Osaka



dependency	nucleus

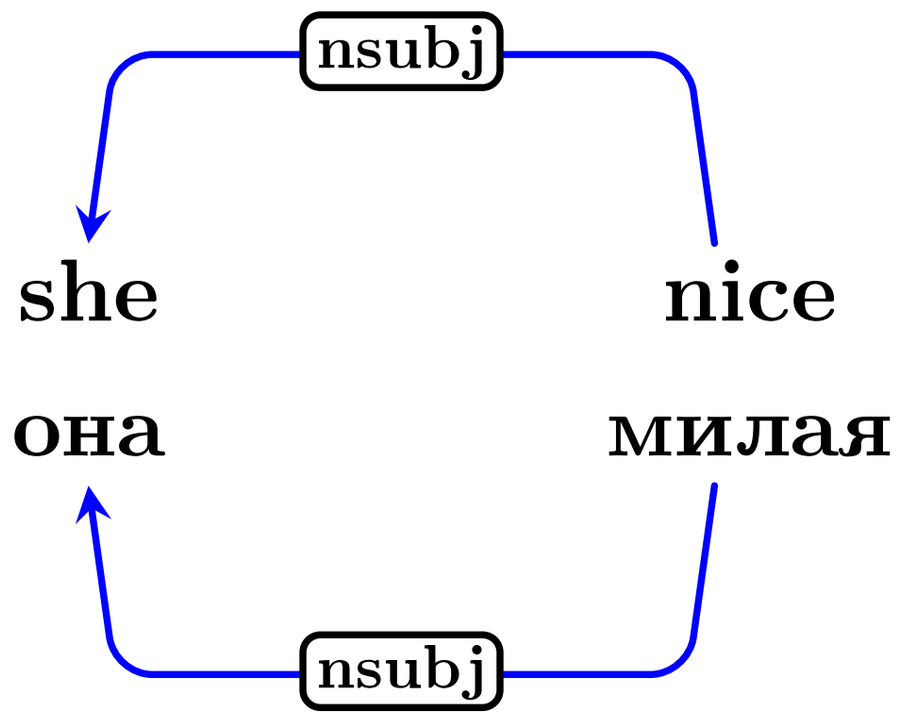


dependency	nucleus
karaka	vibhakti

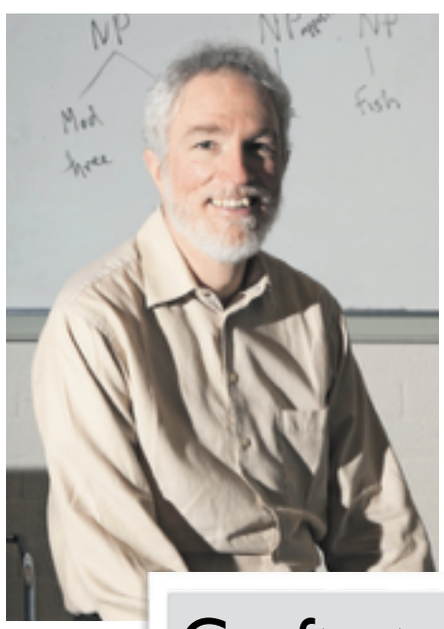


dependency	nucleus
karaka	vibhakti
kakariuke	bunsetsu

Linguistic Typology

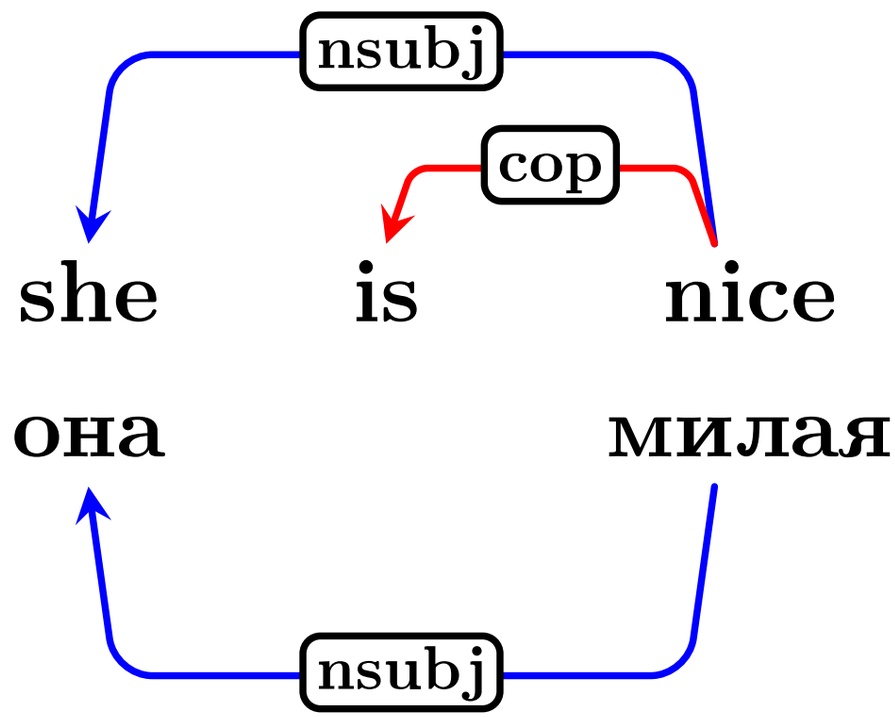


nonverbal predication



Croft et al. (2017) Linguistic Typology Meets Universal Dependencies

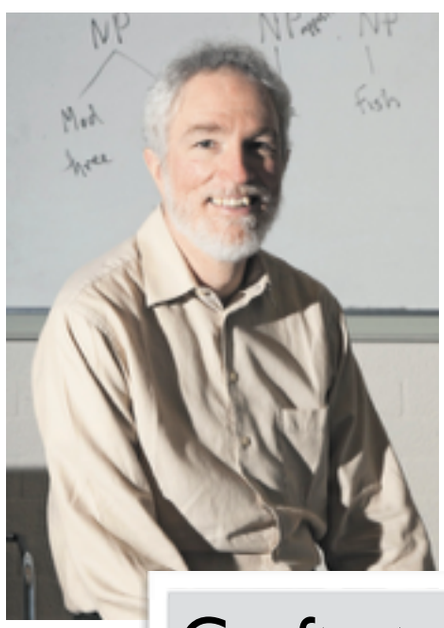
Linguistic Typology



copula strategy

nonverbal predication

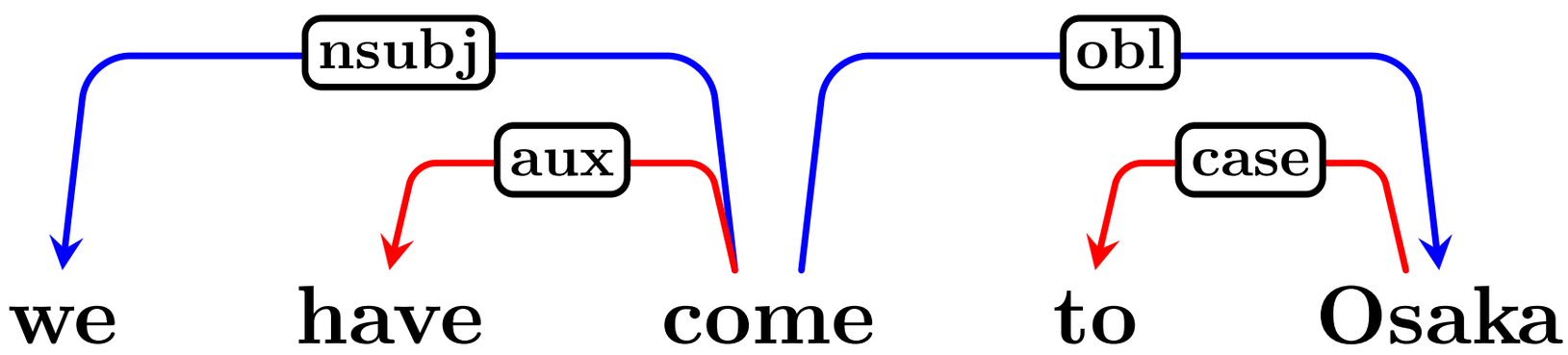
null strategy



Croft et al. (2017) Linguistic Typology Meets Universal Dependencies

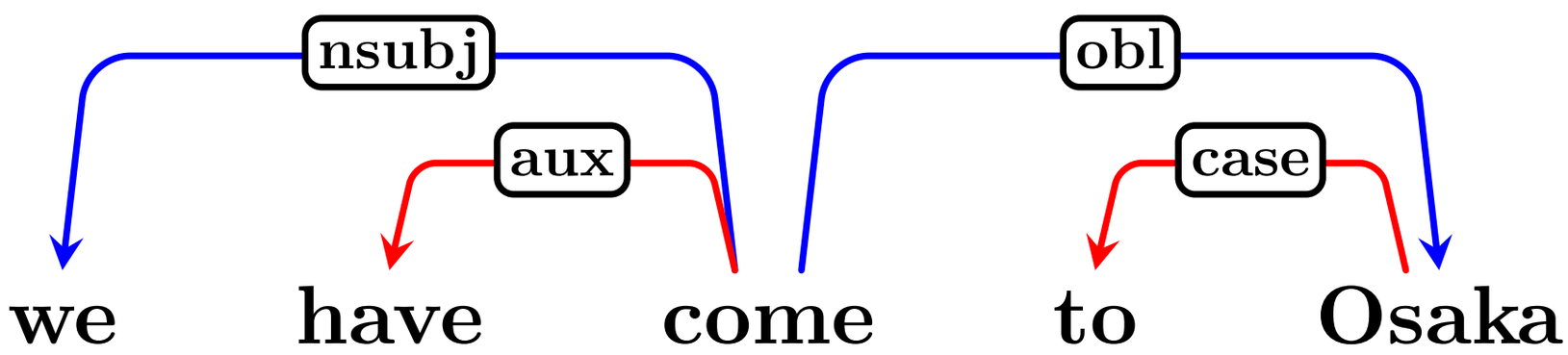
Linguistics vs. Parsing

- Mono-stratal but multi-relational representations
- Both lexical and functional heads can be extracted

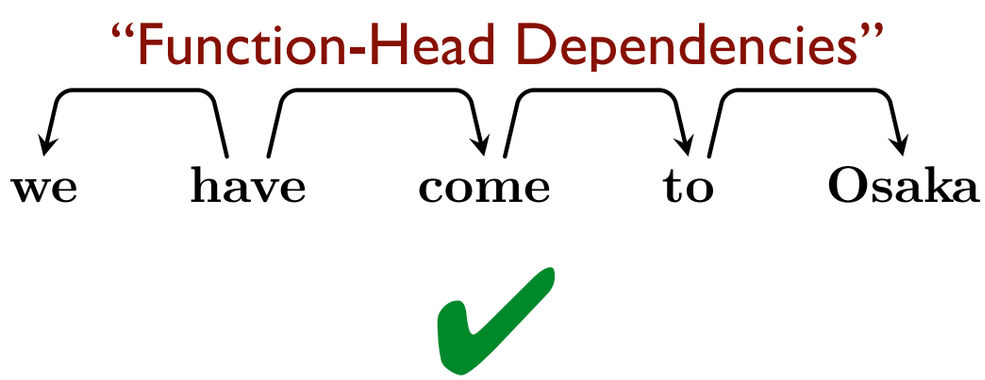


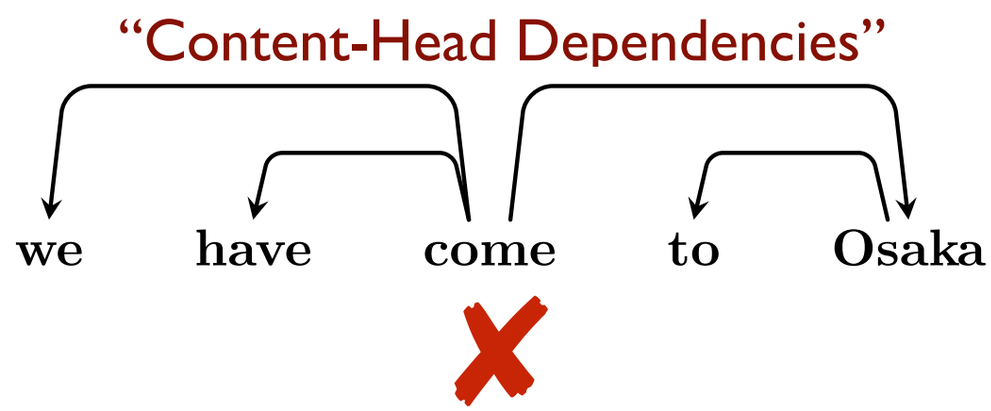
Linguistics vs. Parsing

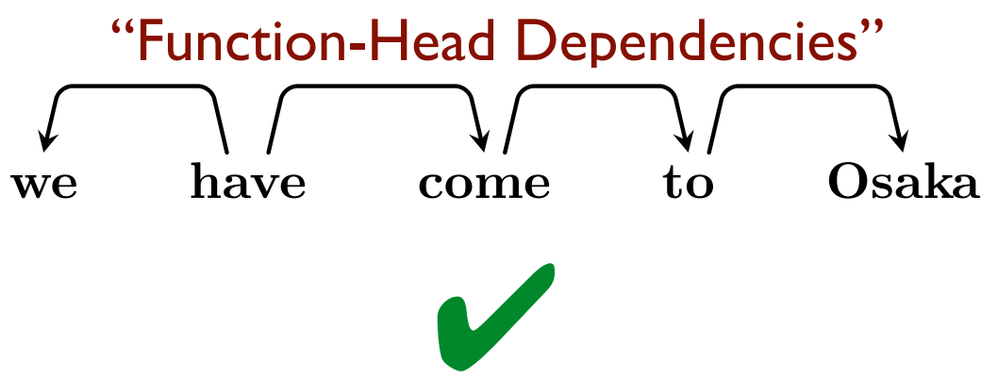
- Mono-stratal but multi-relational representations
- Both lexical and functional heads can be extracted

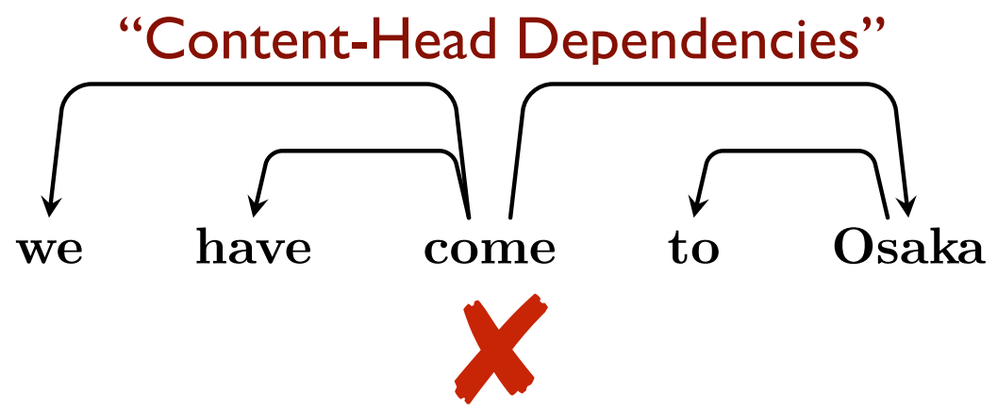


But syntactic parsers don't know this!?

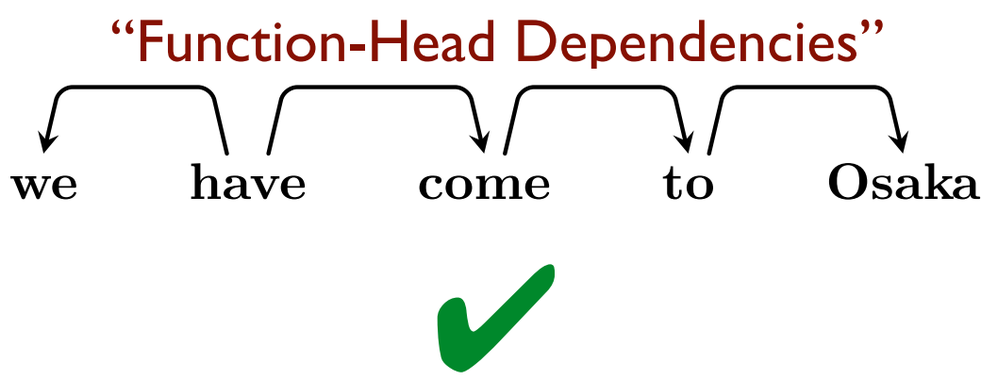

"Function-Head Dependencies"
we
have
come
to
Osaka

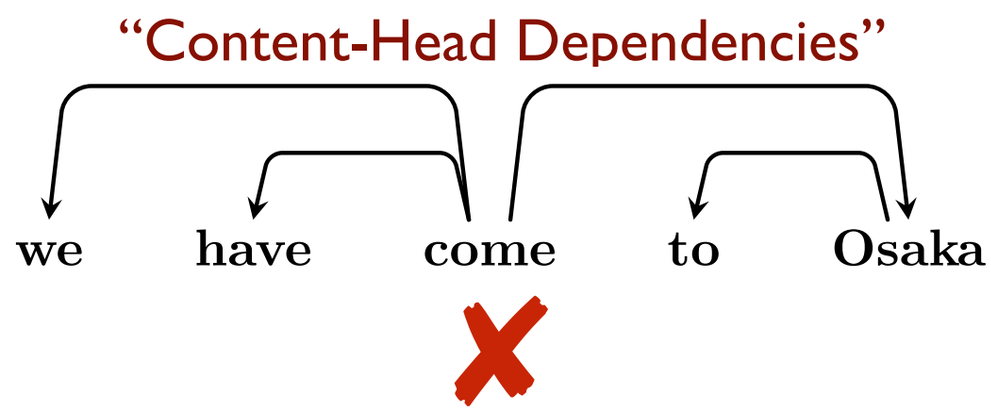

"Content-Head Dependencies"
we
have
come
to
Osaka





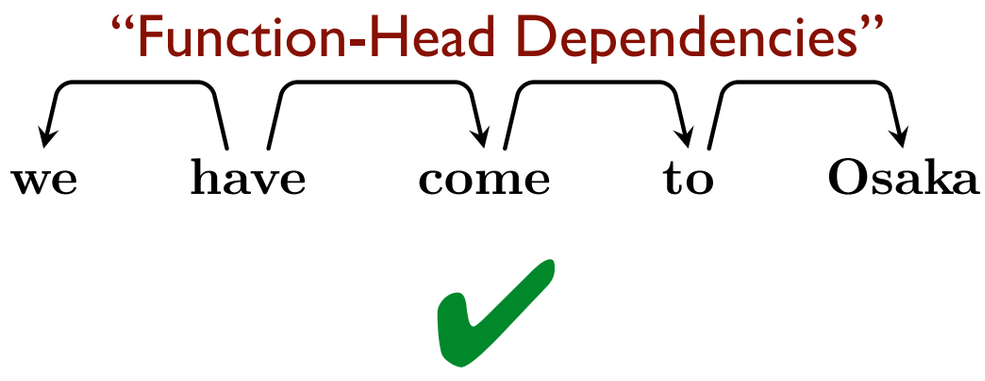
Schwartz et al. (2012) Learnability-Based Syntactic Annotation Design

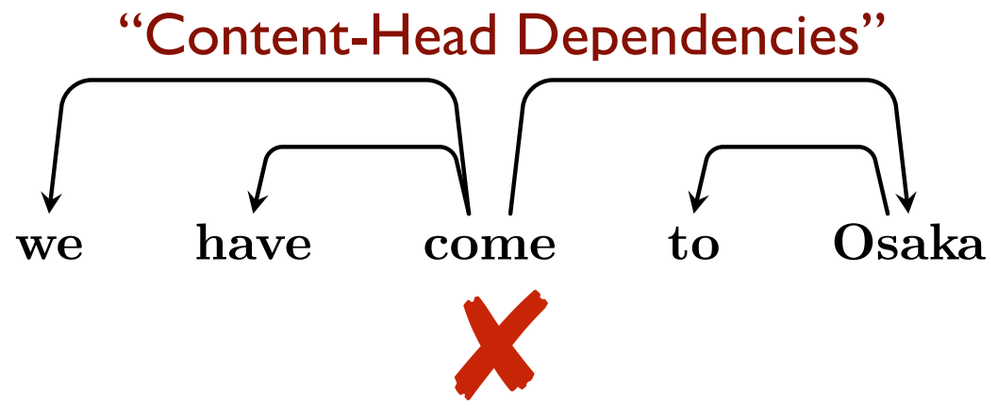




Schwartz et al. (2012) Learnability-Based Syntactic Annotation Design

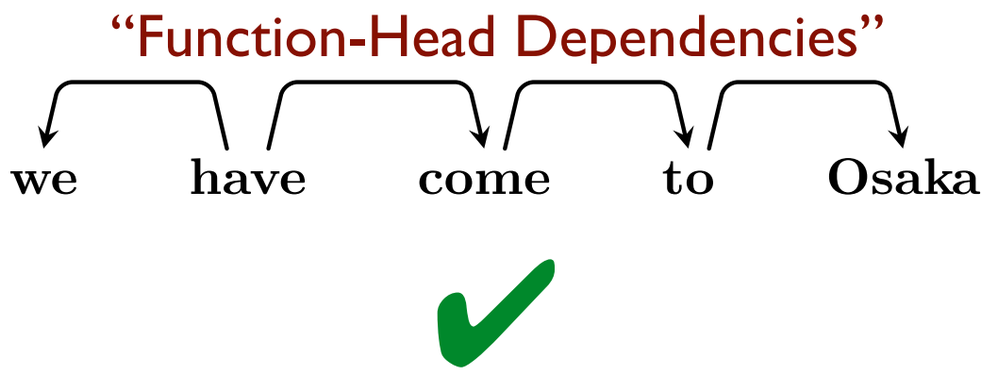
	Function head	Content head

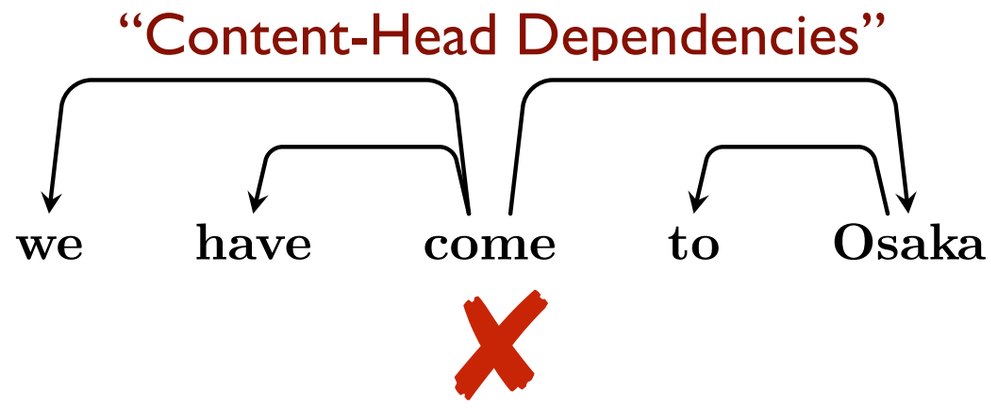




Schwartz et al. (2012) Learnability-Based Syntactic Annotation Design

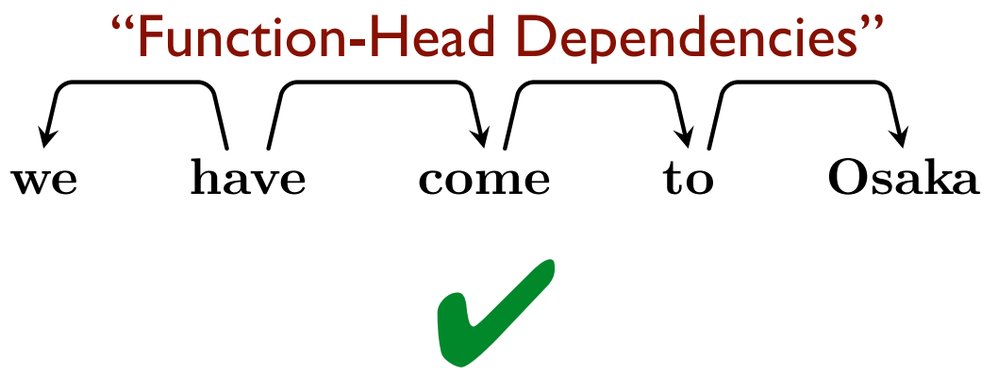
	Function head	Content head
Prep – Noun	✔	✘

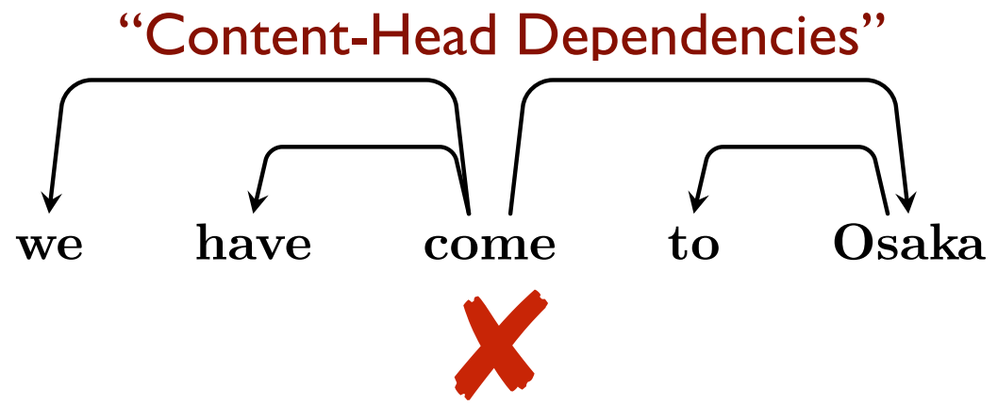




Schwartz et al. (2012) Learnability-Based Syntactic Annotation Design

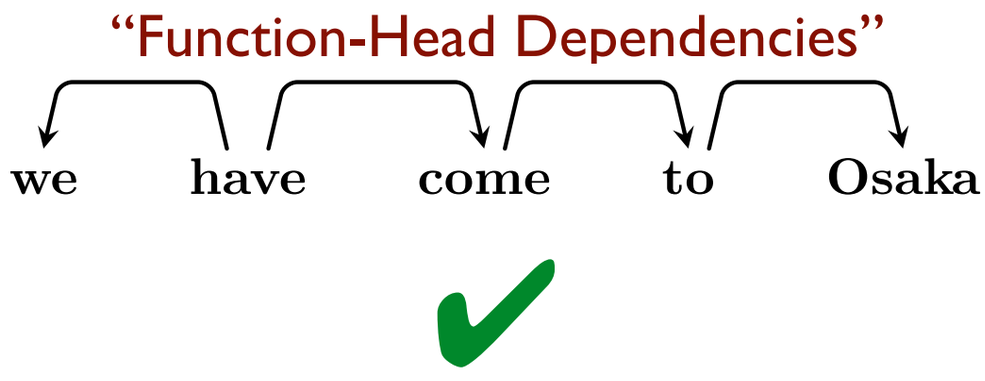
	Function head	Content head
Prep – Noun	✔	✘
Det – Noun	✘	✔

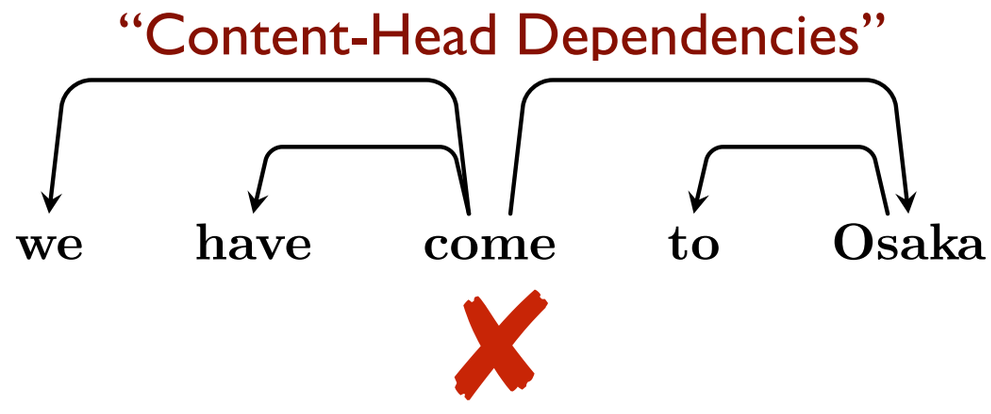




Schwartz et al. (2012) Learnability-Based Syntactic Annotation Design

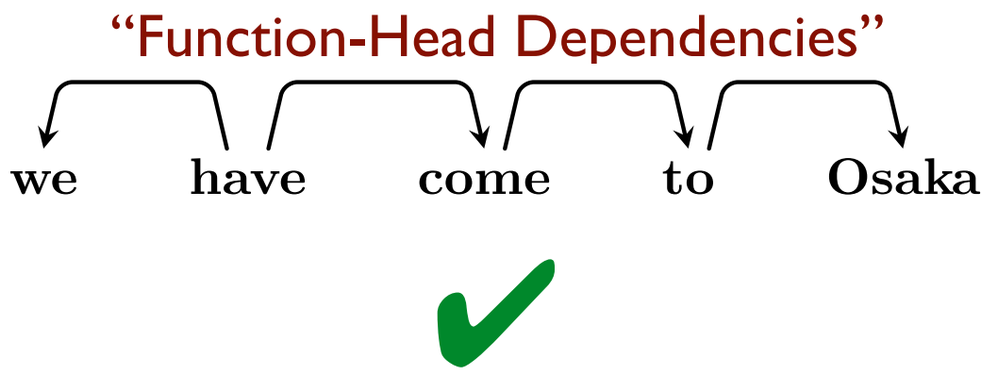
	Function head	Content head
Prep – Noun	✔	✘
Det – Noun	✘	✔
CC – Conj	✘	✔

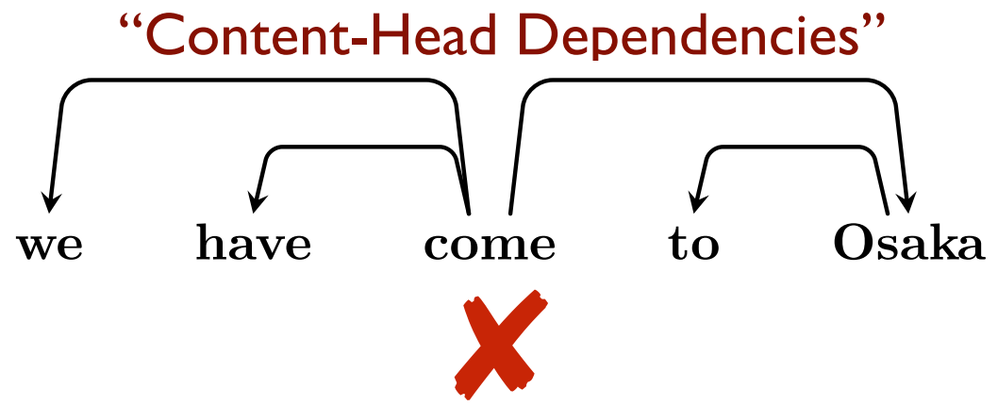




Schwartz et al. (2012) Learnability-Based Syntactic Annotation Design

	Function head	Content head
Prep – Noun	✔	✘
Det – Noun	✘	✔
CC – Conj	✘	✔
Aux – Verb	?	?





Schwartz et al. (2012) Learnability-Based Syntactic Annotation Design

	Function head	Content head
Prep – Noun	✔	✘
Det – Noun	✘	✔
CC – Conj	✘	✔
Aux – Verb	?	?
Mark – Infinitive	?	?

UD Parsing

UD Parsing

Silveira and Manning (2015) Monolingual parsing using transform-detransform	English	aux case cop	Inconclusive results

UD Parsing

Silveira and Manning (2015) Monolingual parsing using transform-detransform	English	aux case cop	Inconclusive results
De Lhoneux and Nivre (2016) Monolingual parsing using transform-detransform	All	aux	Negative results

UD Parsing

Silveira and Manning (2015) Monolingual parsing using transform-detransform	English	aux case cop	Inconclusive results
De Lhoneux and Nivre (2016) Monolingual parsing using transform-detransform	All	aux	Negative results
Attardi et al. (2015) Monolingual parsing using different representations	Italian	case cop	UD > ISDT

UD Parsing

Silveira and Manning (2015) Monolingual parsing using transform-detransform	English	aux case cop	Inconclusive results
De Lhoneux and Nivre (2016) Monolingual parsing using transform-detransform	All	aux	Negative results
Attardi et al. (2015) Monolingual parsing using different representations	Italian	case cop	UD > ISDT
Rosa (2015) Multi-source delexicalized transfer parsing	All	case	UD > PDT

UD Parsing

UD Parsing

Not so bad after all?

- No clear evidence that “content-head” is harder to parse in general
- In the cross-lingual setting, it even seems to work better

UD Parsing

Not so bad after all?

- No clear evidence that "content-head" is harder to parse in general
- In the cross-lingual setting, it even seems to work better

Can we do better?

- Exploit the full representation – lexical and functional heads
- Use typology of syntactic relations as a bias for learning

A Historical Perspective

A Historical Perspective

Constituency parsing – largely driven by PTB

- Perhaps too much emphasis on English (until recently)
- But deep analysis of categories and representations

A Historical Perspective

Constituency parsing – largely driven by PTB

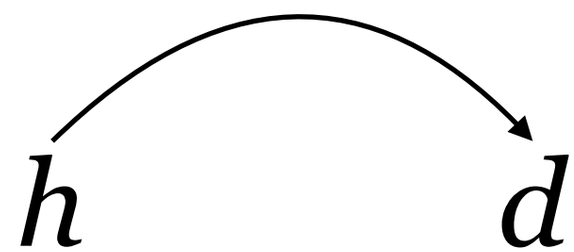
- Perhaps too much emphasis on English (until recently)
- But deep analysis of categories and representations

Dependency parsing – largely driven by CoNLL data

- More attention to typological diversity from the start
- But parsers had to remain agnostic about linguistic categories

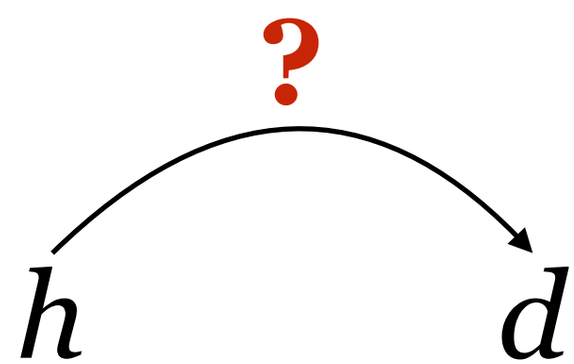
Dependency Parsing

Dependency Parsing



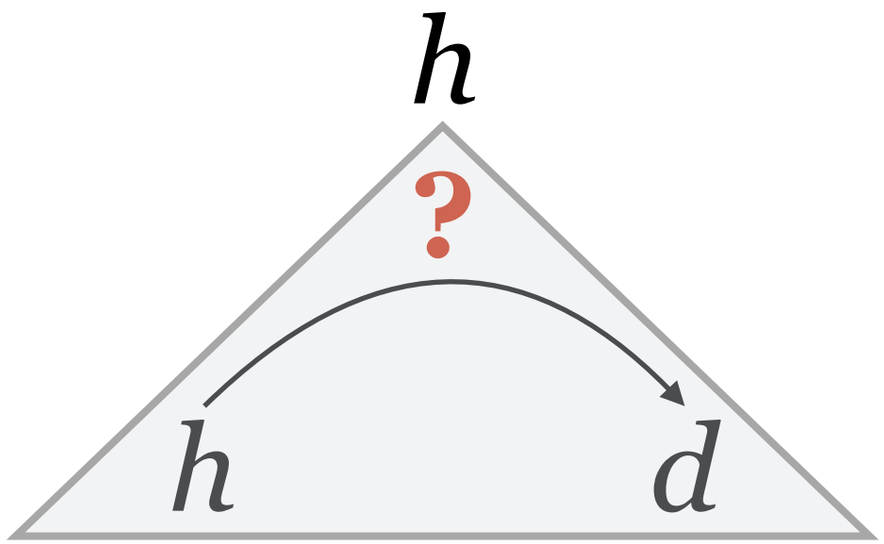
- Parsers know only one type of syntactic relation

Dependency Parsing



- Parsers know only one type of syntactic relation
- Parsers do not interpret dependency labels

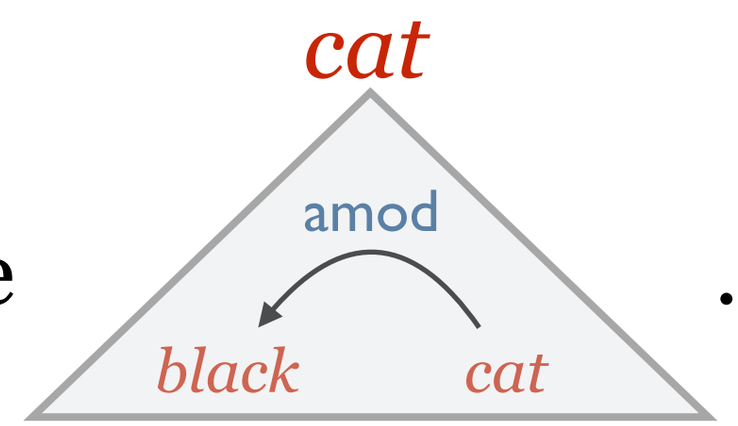
Dependency Parsing



- Parsers know only one type of syntactic relation
- Parsers do not interpret dependency labels
- Parsers represent every construction by its “head”

Dependency Parsing

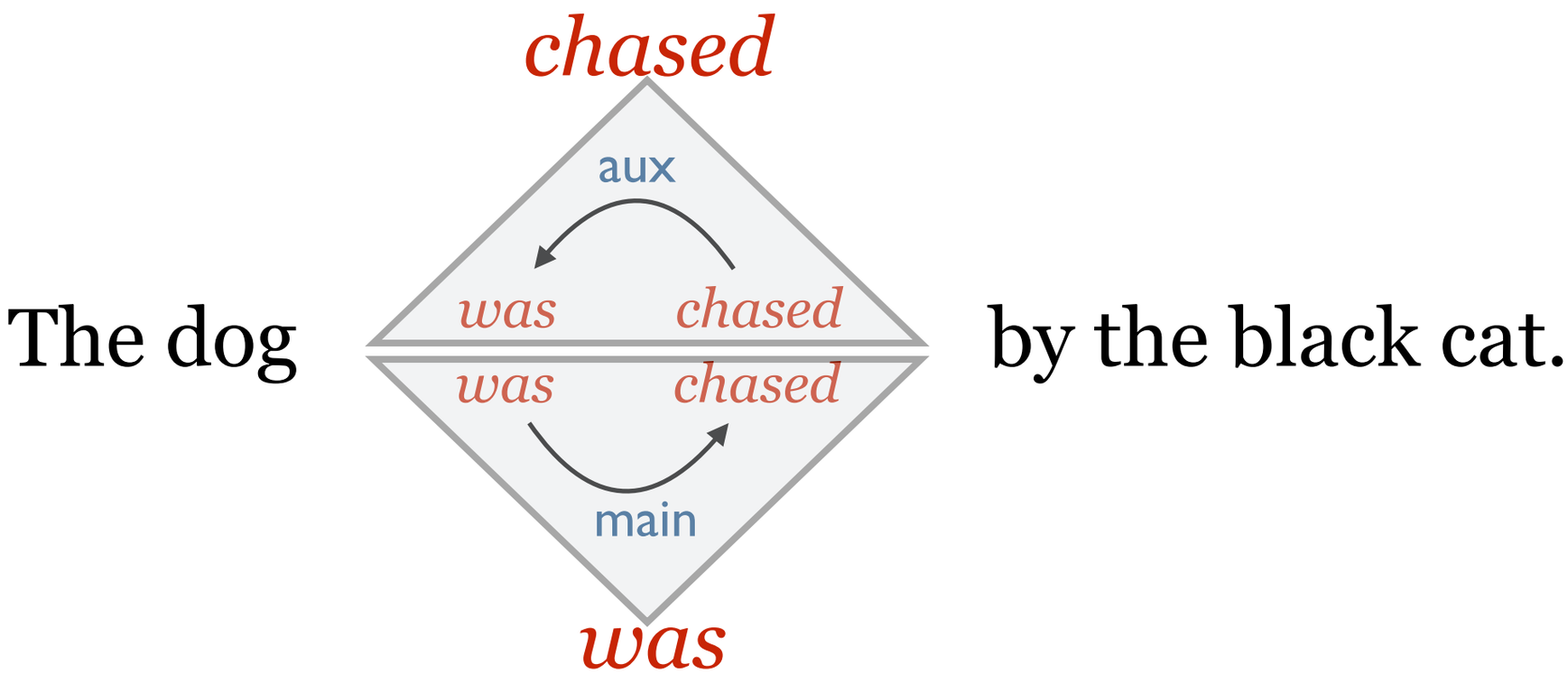
The dog was chased by the



.

- Endocentric construction with *cat* as head
- Little (syntactic) information is lost by dropping *black*

Dependency Parsing



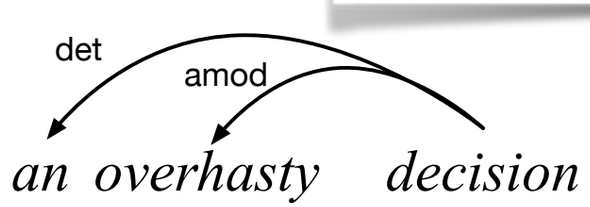
- Dissociated nucleus consisting of *was* and *chased*
- Neither content-head nor function-head is right!

UD and Deep Learning

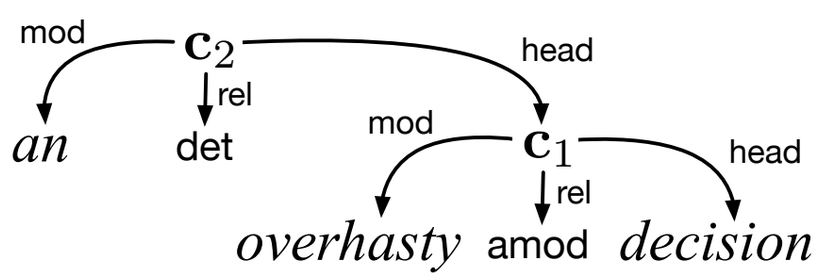
UD and Deep Learning

Stenetorp (2013) Transition-Based Dependency Parsing Using Recursive Neural Networks

Dyer et al. (2015) Transition-Based Dependency Parsing with Stack Long Short-Term Memory



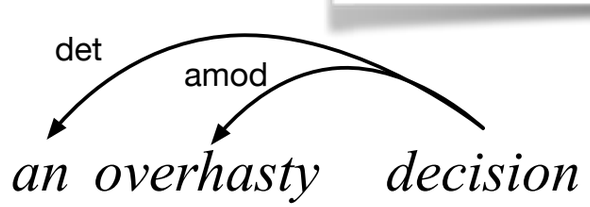
composition functions
+1–2% labeled accuracy



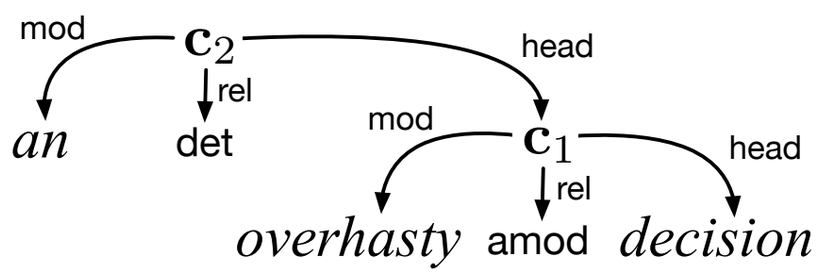
UD and Deep Learning

Stenetorp (2013) Transition-Based Dependency Parsing Using Recursive Neural Networks

Dyer et al. (2015) Transition-Based Dependency Parsing with Stack Long Short-Term Memory



composition functions
+1–2% labeled accuracy



Kuncoro et al. (2016) What Do Recurrent Neural Network Grammars Learn About Syntax?

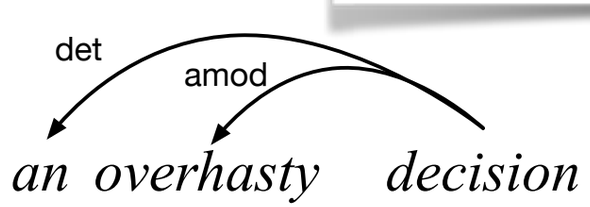
attention
graded endocentricity

Noun phrases	Verb phrases	Prepositional phrases
Canadian (0.09) **Auto (0.31)** Workers (0.2) union (0.22) president (0.18)	**buying (0.31)** and (0.25) selling (0.21) NP (0.23)	ADVP (0.14) **on (0.72)** NP (0.14)
no (0.29) major (0.05) **Eurobond (0.32)** or (0.01) foreign (0.01) bond (0.1) offerings (0.22)	ADVP (0.27) **show (0.29)** PRT (0.23) PP (0.21)	ADVP (0.05) **for (0.54)** NP (0.40)
Saatchi (0.12) client (0.14) Philips (0.21) Lighting (0.24) **Co. (0.29)**	**pleaded (0.48)** ADJP (0.23) PP (0.15) PP (0.08) PP (0.06)	ADVP (0.02) **because (0.73)** of (0.18) NP (0.07)
nonperforming (0.18) commercial (0.23) **real (0.25)** estate (0.1) **assets (0.25)**	**received (0.33)** PP (0.18) NP (0.32) PP (0.17)	such (0.31) **as (0.65)** NP (0.04)
the (0.1) Jamaica (0.1) Tourist (0.03) Board (0.17) ad (0.20) **account (0.40)**	cut (0.27) **NP (0.37)** PP (0.22) PP (0.14)	from (0.39) **NP (0.49)** PP (0.12)
the (0.0) final (0.18) **hour (0.81)**	**to (0.99)** VP (0.01)	**of (0.97)** NP (0.03)
their (0.0) first (0.23) **test (0.77)**	**were (0.77)** n't (0.22) VP (0.01)	**in (0.93)** NP (0.07)
Apple (0.62) , (0.02) Compaq (0.1) and (0.01) IBM (0.25)	did (0.39) **n't (0.60)** VP (0.01)	**by (0.96)** S (0.04)
both (0.02) stocks (0.03) and (0.06) **futures (0.88)**	handle (0.09) **NP (0.91)**	**at (0.99)** NP (0.01)
NP (0.01) , (0.0) **and (0.98)** NP (0.01)	VP (0.15) **and (0.83)** VP 0.02)	NP (0.1) **after (0.83)** NP (0.06)

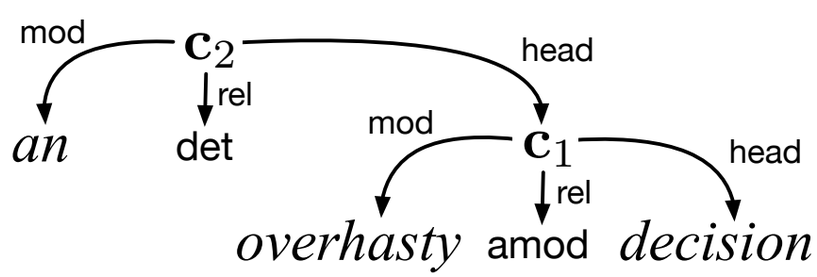
UD and Deep Learning

Stenetorp (2013) Transition-Based Dependency Parsing Using Recursive Neural Networks

Dyer et al. (2015) Transition-Based Dependency Parsing with Stack Long Short-Term Memory



composition functions
+1–2% labeled accuracy



UD as an inductive bias

Kuncoro et al. (2016) What Do Recurrent Neural Network Grammars Learn About Syntax?

attention
graded endocentricity

Noun phrases	Verb phrases	Prepositional phrases
Canadian (0.09) **Auto (0.31)** Workers (0.2) union (0.22) president (0.18)	**buying (0.31)** and (0.25) selling (0.21) NP (0.23)	ADVP (0.14) **on (0.72)** NP (0.14)
no (0.29) major (0.05) **Eurobond (0.32)** or (0.01) foreign (0.01) bond (0.1) offerings (0.22)	ADVP (0.27) **show (0.29)** PRT (0.23) PP (0.21)	ADVP (0.05) **for (0.54)** NP (0.40)
Saatchi (0.12) client (0.14) Philips (0.21) Lighting (0.24) **Co. (0.29)**	**pleaded (0.48)** ADJP (0.23) PP (0.15) PP (0.08) PP (0.06)	ADVP (0.02) **because (0.73)** of (0.18) NP (0.07)
nonperforming (0.18) commercial (0.23) **real (0.25)** estate (0.1) **assets (0.25)**	**received (0.33)** PP (0.18) NP (0.32) PP (0.17)	such (0.31) **as (0.65)** NP (0.04)
the (0.1) Jamaica (0.1) Tourist (0.03) Board (0.17) ad (0.20) **account (0.40)**	cut (0.27) **NP (0.37)** PP (0.22) PP (0.14)	from (0.39) **NP (0.49)** PP (0.12)
the (0.0) final (0.18) **hour (0.81)**	**to (0.99)** VP (0.01)	**of (0.97)** NP (0.03)
their (0.0) first (0.23) **test (0.77)**	**were (0.77)** n't (0.22) VP (0.01)	**in (0.93)** NP (0.07)
Apple (0.62) , (0.02) Compaq (0.1) and (0.01) IBM (0.25)	did (0.39) **n't (0.60)** VP (0.01)	**by (0.96)** S (0.04)
both (0.02) stocks (0.03) and (0.06) **futures (0.88)**	handle (0.09) **NP (0.91)**	**at (0.99)** NP (0.01)
NP (0.01) , (0.0) **and (0.98)** NP (0.01)	VP (0.15) **and (0.83)** VP 0.02)	NP (0.1) **after (0.83)** NP (0.06)

Conclusion

Conclusion

Dubious linguistics?

- Lexical dependencies and functional relations encoded in a single tree
- Grounded in linguistic typology and dependency grammar traditions

Conclusion

Dubious linguistics?

- Lexical dependencies and functional relations encoded in a single tree
- Grounded in linguistic typology and dependency grammar traditions

Crappy parsing?

- Not so bad with existing parsers, especially for cross-lingual parsing
- Learn richer parsing models grounded in linguistic typology

UD Events in 2017

CoNLL-2017 Shared Task

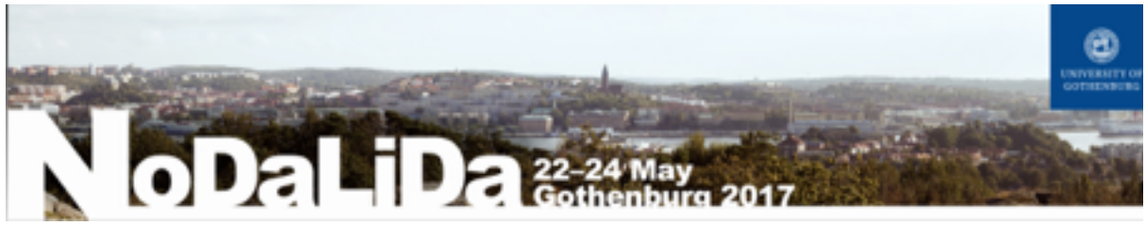
http://universaldependencies.org/conll17/

- Multilingual Parsing from Raw Text to Universal Dependencies
- Collocated with ACL, August 3–4, 2017, Vancouver, Canada
- Call for participation in December 2016, data release in March 2017

First Workshop on Universal Dependencies

http://universaldependencies.org/udw17/

- Collocated with NoDaLiDa, May 20, 2017, Gothenburg, Sweden
- Submission deadline: March 20, 2017



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, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, 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, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, 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, Keiko Sophie Mori, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Petya Osenova, Robert Östling, Lilja Øvrelid, Valeria Paiva, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, 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, Shadi Saleh, Baiba Saulīte, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Carolyn Spadine, Alane Suhr, Umut Sulubacak, Zsolt Szántó, 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