Universality in Space and Time
Modern Treebanking for Ancient Languages

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

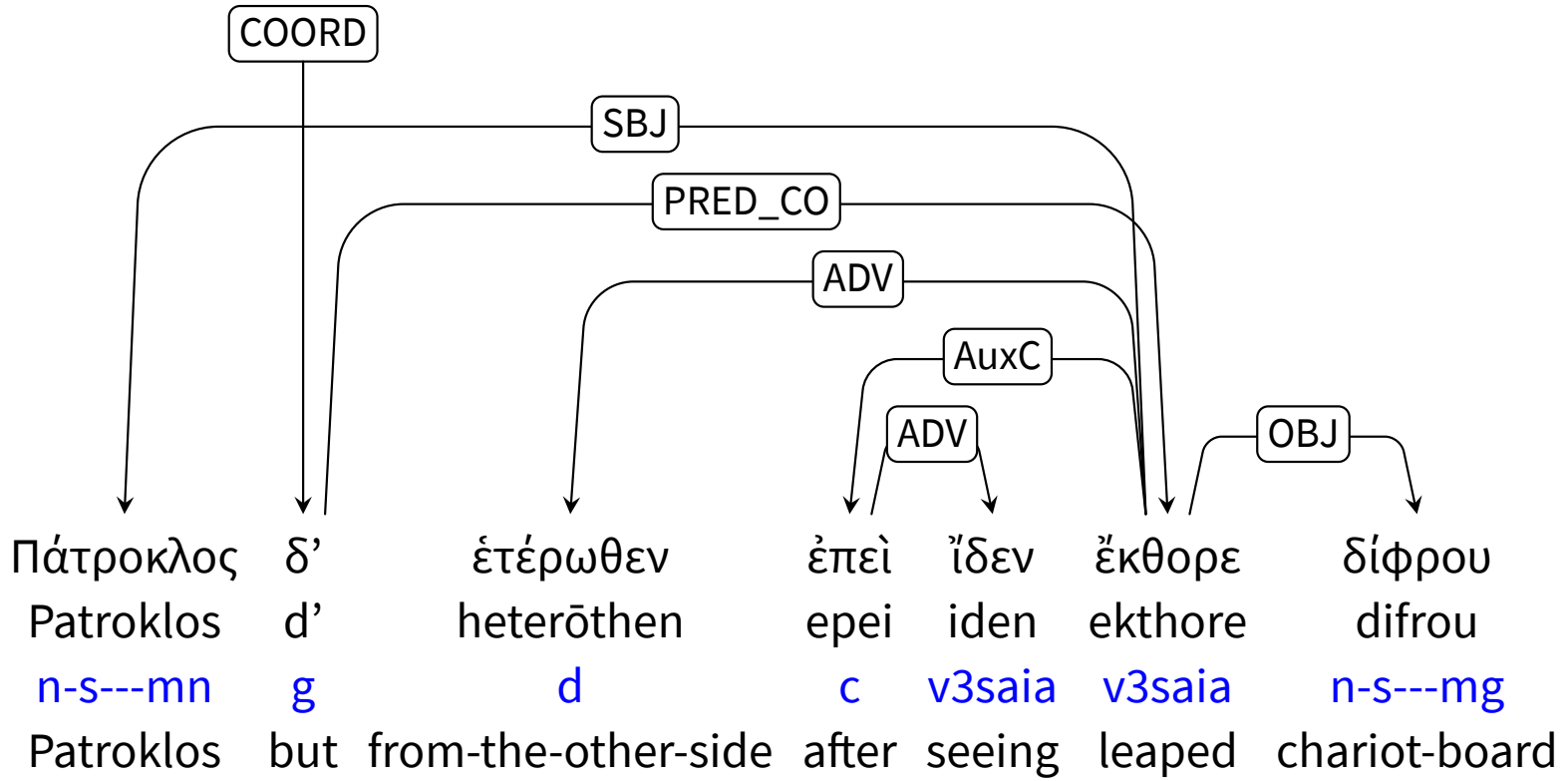
Dan Zeman

zeman@ufal.mff.cuni.cz
http://universaldependencies.org/

Dependency Treebanks

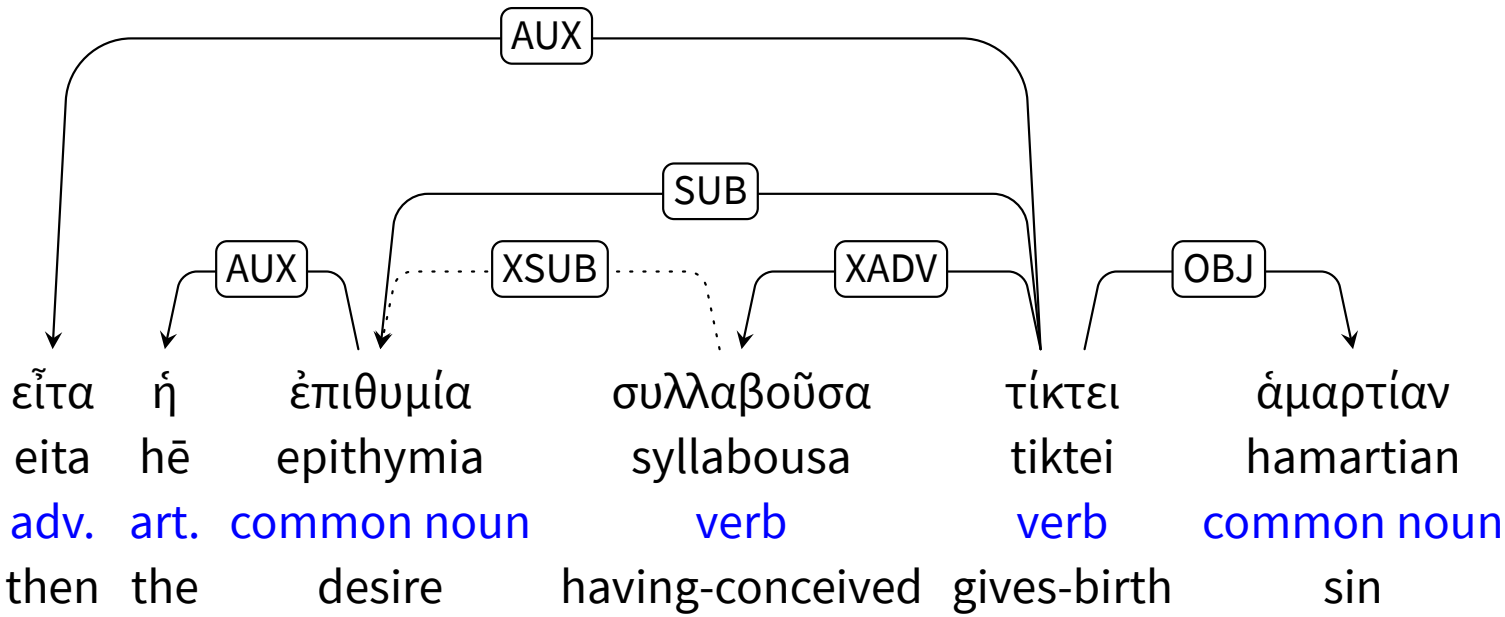
- Annotation available for growing number of languages
- Useful for
  - Natural language processing (learning models)
  - Linguistic research
  - *Jonathan and Marco made the case for treebanks yesterday*
- Cross-language/corpus comparison???

Ancient Greek: Perseus (Prague Style)



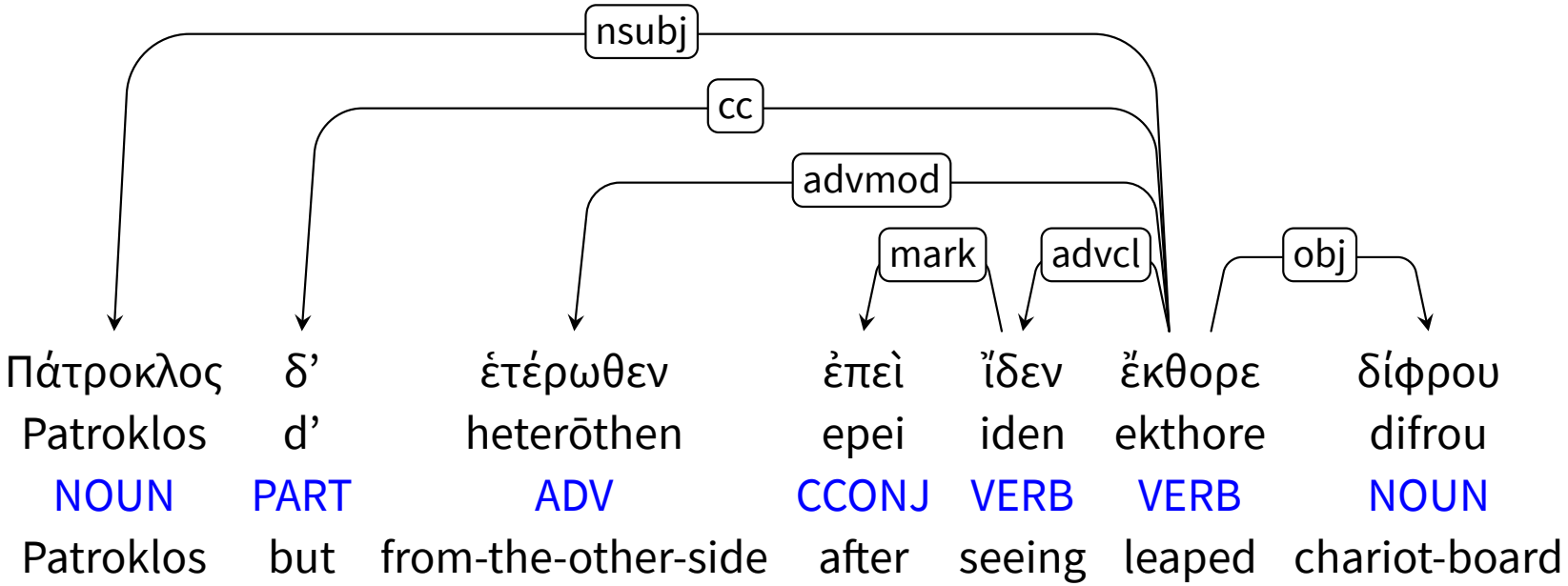
“and Patroklos, when he saw this, leaped on to the ground also”
(Homer, Iliad 16.427)

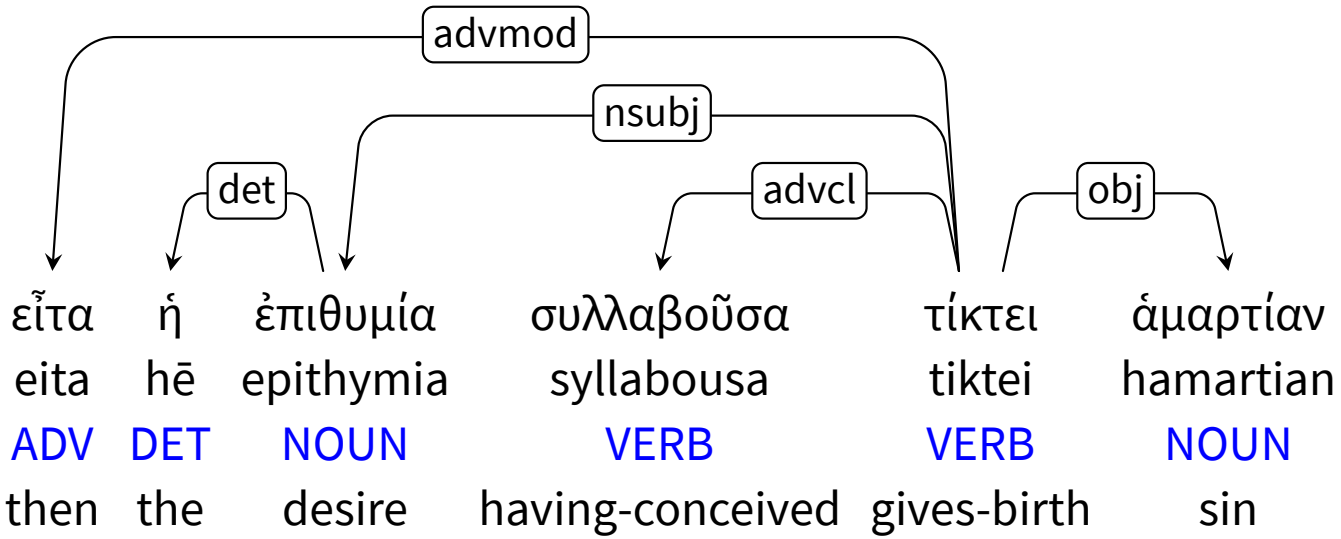
Ancient Greek: PROIEL



“then desire when it has conceived gives birth to sin” (James 1:15)

UD Ancient Greek





In Other Words…

- Recall Dirk's "programming theologist" demo from yesterday?
  - Hebrew: gn m f
  - Greek: Gender Masculine Feminine

In Other Words…

- Recall Dirk's "programming theologist" demo from yesterday?
  - ~~Hebrew: gn m f~~
  - ~~Greek: Gender Masculine Feminine~~

- UD (Hebrew, Greek or any other language):
  - Gender=Masc | Gender=Fem

Goals and Requirements

- Cross-linguistically consistent grammatical annotation

Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Based on common usage and existing de-facto standards
  - Originally NLP parsing community (not classical studies)
  - $\Rightarrow$ CoNLL-U table format (not XML)

Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Based on common usage and existing de-facto standards
  - Originally NLP parsing community (not classical studies)
  - $\Rightarrow$ CoNLL-U table format (not XML)
  - $\Rightarrow$ No unified reference to the canonical text
  - … but absolutely no problem with using the existing ones!
  - All sorts of extra info can be stored in CoNLL-U

Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Based on common usage and existing de-facto standards
- Caveats:
  - Not a new linguistic theory –
    but linguistically informed and relevant

Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Based on common usage and existing de-facto standards
- Caveats:
  - Not a new linguistic theory –
    but linguistically informed and relevant
  - Not the ultimate annotation scheme –
    but a lightweight lingua franca
  - Allow language-specific extensions

Not "Universal" in the strictly typological sense!

Golden Rules

- Maximize parallelism
  - Don’t annotate the same thing in different ways
  - Don’t make different things look the same

Golden Rules

- Maximize parallelism
  - Don't annotate the same thing in different ways
  - Don't make different things look the same
- But don't overdo it
  - Balance: is it still the same thing?
  - We want to compare different strategies across languages

Morphology

ι	вьсн	людне	вндѣвъше	въздашѧ	хвалѫ	б̄вн
i	vьsi	ljudie	viděvъše	vъzdašę	chvalǫ	b.vi
and	*all*	*people*	*having-seen*	*paid*	*praise*	*to-God*

Morphology

ι	вьсн	людне	вндѣвъше	въздашѧ	хвалѫ	б҃вн
i	vьsi	ljudie	viděvъše	vъzdašę	chvalǫ	b.vi
and	*all*	*people*	*having-seen*	*paid*	*praise*	*to-God*
н	вьсь	людиє	вндѣти	въздати	хвала	богъ

- Lemma representing the semantic content of the word

Morphology

ι	вьсн	людне	вндѣвъше	въздашѧ	хвалѫ	б҃вн
i	vьsi	ljudie	viděvъše	vъzdašę	chvalǫ	b.vi
and	*all*	*people*	*having-seen*	*paid*	*praise*	*to-God*
и	вьсь	людиє	вндѣти	въздати	хвала	богъ
CCONJ	DET	NOUN	VERB	VERB	NOUN	NOUN

- Lemma representing the semantic content of the word
- Part-of-speech tag representing the abstract lexical category associated with the word

Morphology

ι	вьсн	людне	вндѣвъше	въздашѧ	хвалѫ	б҃вн
i	vьsi	ljudie	viděvъše	vъzdašę	chvalǫ	b.vi
and	*all*	*people*	*having-seen*	*paid*	*praise*	*to-God*
н	вьсь	людиє	вндѣти	въздати	хвала	богъ
CCONJ	DET	NOUN	VERB	VERB	NOUN	NOUN
	PronType=Tot Gender=Masc Number=Plur Case=Nom	Gender=Masc Number=Plur Case=Nom	VerbForm=Part Voice=Act Tense=Past Gender=Masc Number=Plur Case=Nom Strength=Strong	VerbForm=Fin Voice=Act Tense=Past Aspect=Perf Mood=Ind Number=Plur Person=3	Gender=Fem Number=Sing Case=Acc	Gender=Masc Number=Sing Case=Dat

- Lemma representing the semantic content of the word
- Part-of-speech tag representing the abstract lexical category associated with the word
- Features representing lexical and grammatical properties associated with the lemma or the particular word form

Part-of-Speech Tags

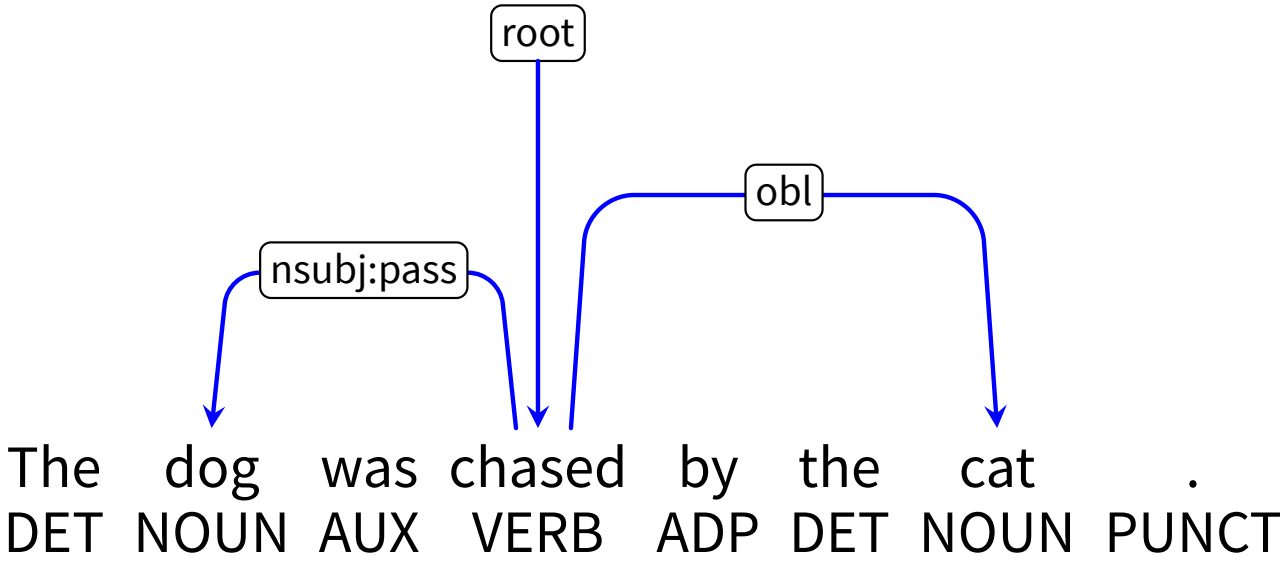
Open	***Closed***	***Other***
ADJ	*ADP*	*PUNCT*
ADV	*AUX*	*SYM*
INTJ	*CCONJ*	*X*
NOUN	*DET*	
PROPN	*NUM*	
VERB	*PART*	
	PRON	
	SCONJ	

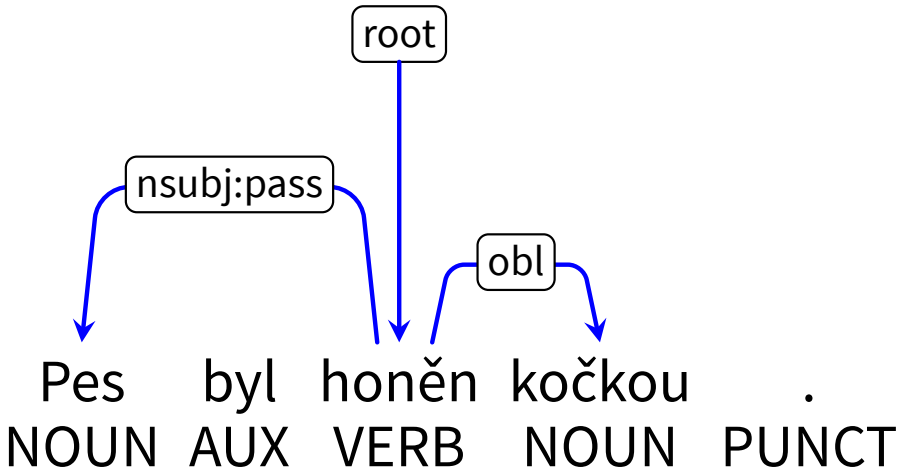
- Taxonomy of 17 universal part-of-speech tags, based on the Google Universal Tagset (Petrov et al., 2012)
- All languages use the same inventory, but not all tags have to be used by all languages

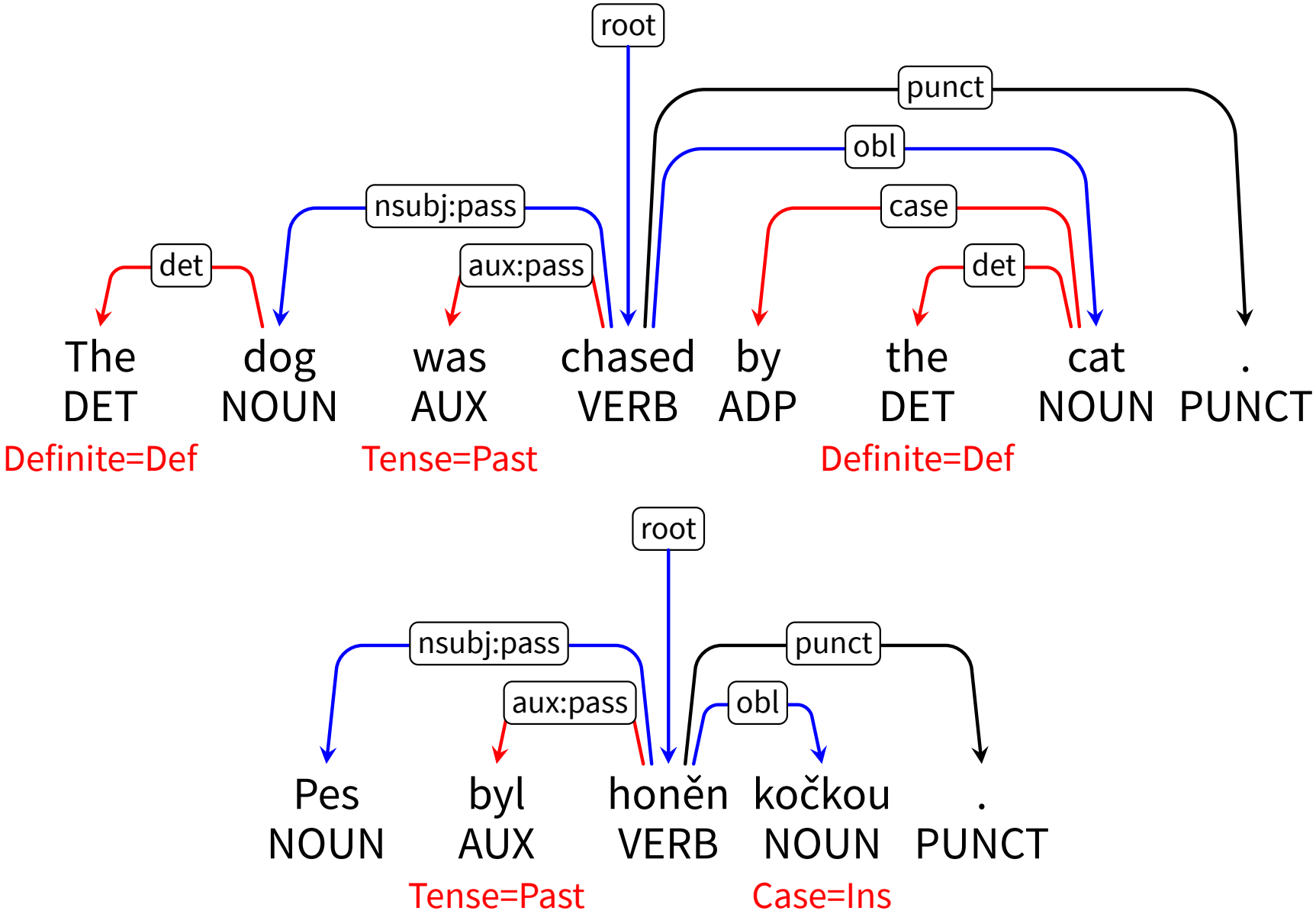
Features

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

- Standardized inventory of morphological features, based on Interset (Zeman, 2008)
- Languages select relevant features and can add language-specific features or values with documentation


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


root
nsubj:pass
obl
Pes byl honěn kočkou .
NOUN AUX VERB NOUN PUNCT

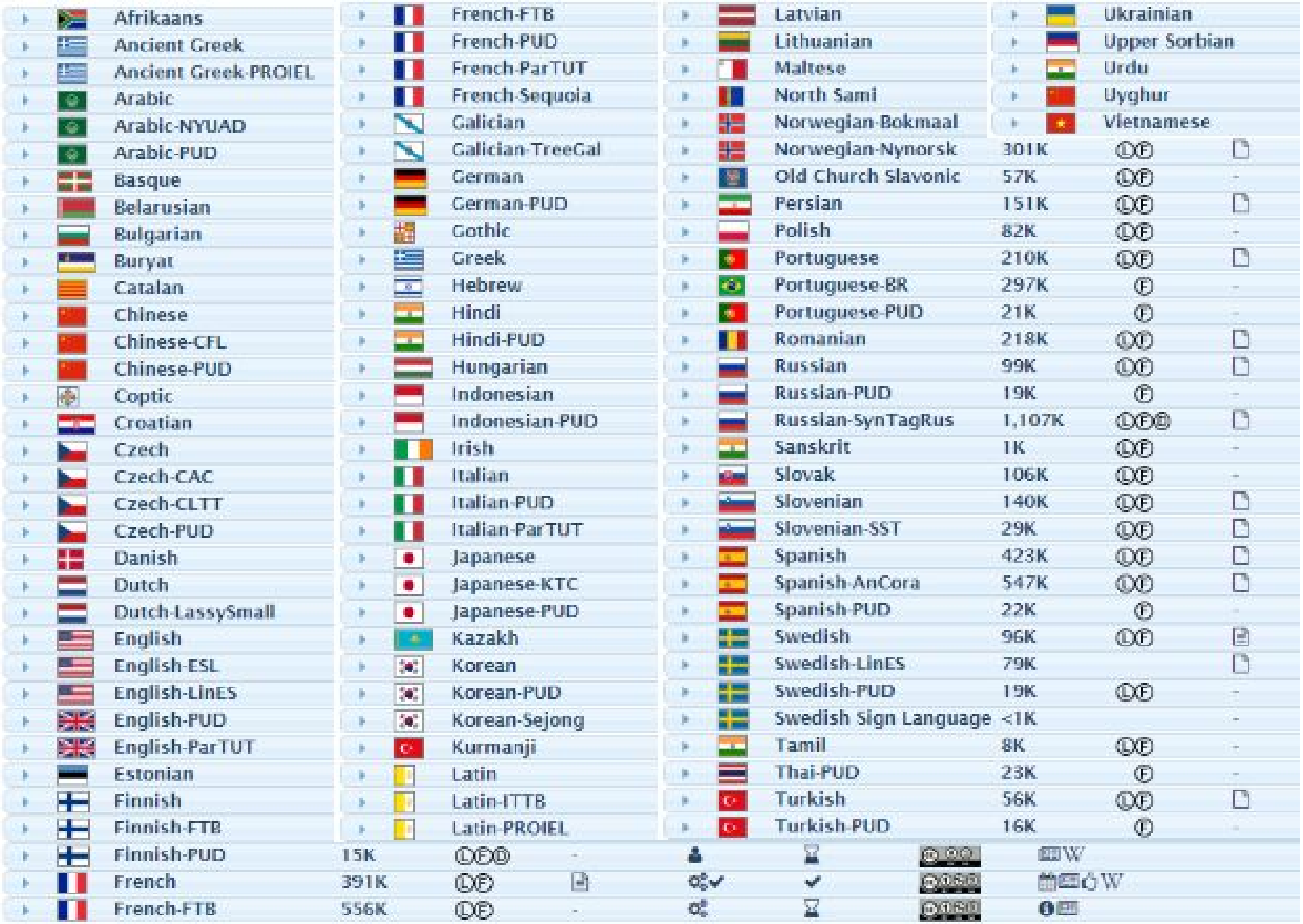

root
punct
obl
nsubj:pass
case
det
aux:pass
det
The DET
dog NOUN
was AUX
chased VERB
by ADP
the DET
cat NOUN
. PUNCT
Definite=Def
Tense=Past
Definite=Def
root
nsubj:pass
punct
aux:pass
obl
Pes NOUN
byl AUX
honěn VERB
kočkou NOUN
. PUNCT
Tense=Past
Case=Ins

Where Are We Now?

Where Are We Now?

- Three years since EACL 2014
- 6 treebank releases (every 6 months)
- 70 (95) treebanks
- 50 (57) languages (over 50% world's population)
- Over 13M tokens; treebanks range from <1K to 1.5M
- Over 200 contributors
  - language group consistency SIGs
- Version 2 guidelines
- CoNLL Shared Task in parsing UD

57 Languages and Growing

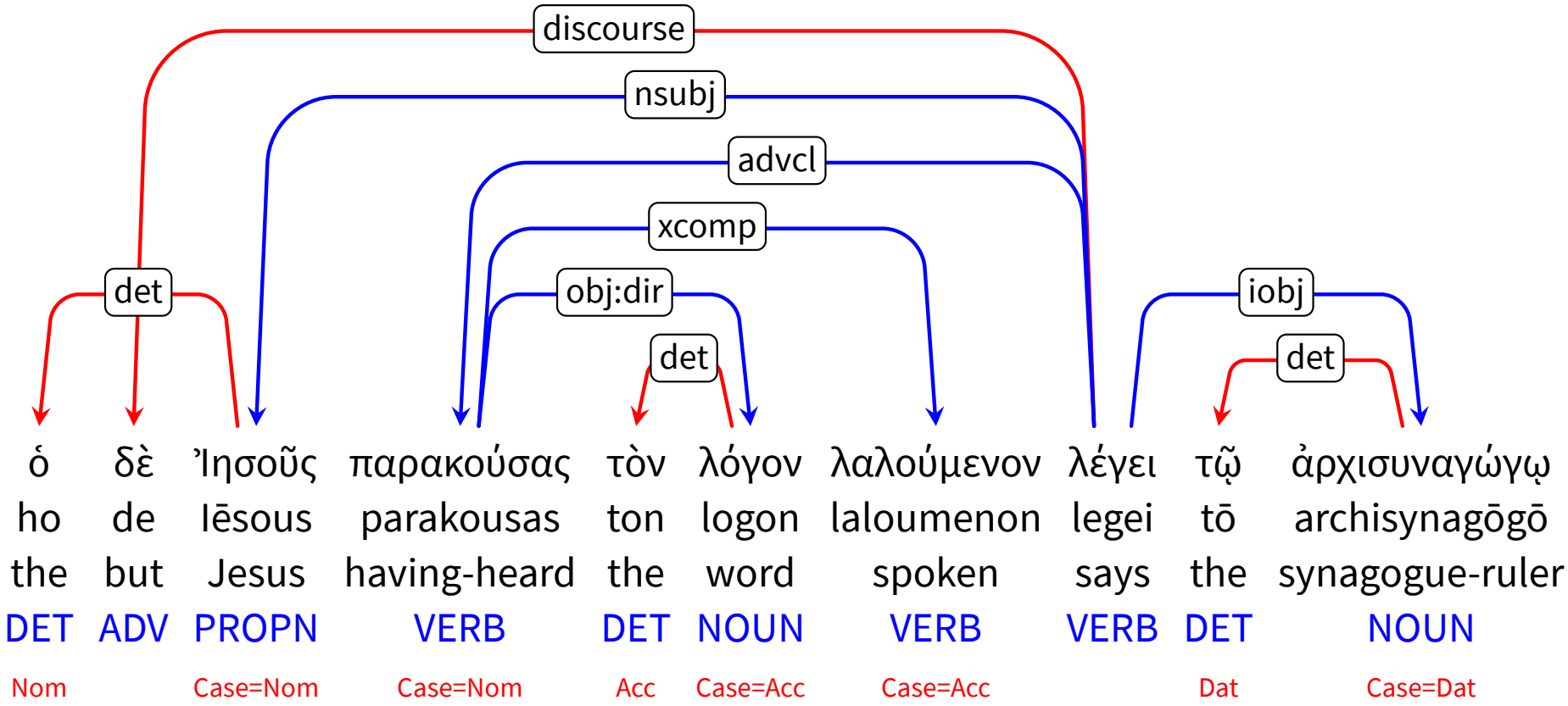


57 Languages and Growing

- Bottom-up process
- Majority: Big languages from Europe and Asia
- Some small languages: Buryat, Kurmanji, Sámi, Upper Sorbian, Uyghur
- Fieldwork – Mehweb Dargwa?

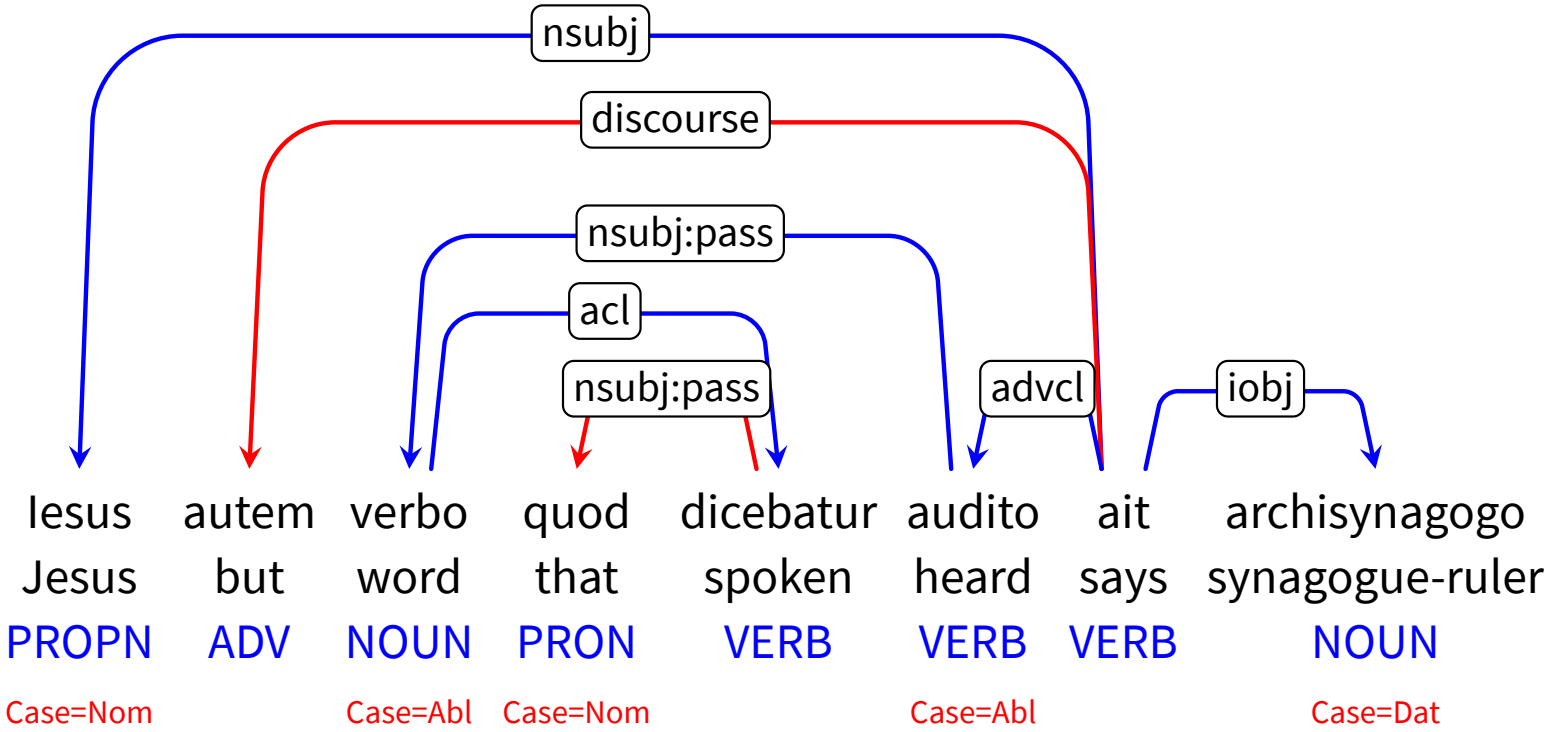
- Extinct / classical: Ancient Greek, Latin, Coptic, Gothic, Old Church Slavonic, Sanskrit
  - Only morphology: Old Hungarian
  - More coming?
    - ★ Old French
    - ★ PROIEL: Old Russian, Old Armenian
    - ★ Guglielmo Inglese: Hittite?
    - ★ Josef Válek: Akkadian?

UD Ancient Greek PROIEL



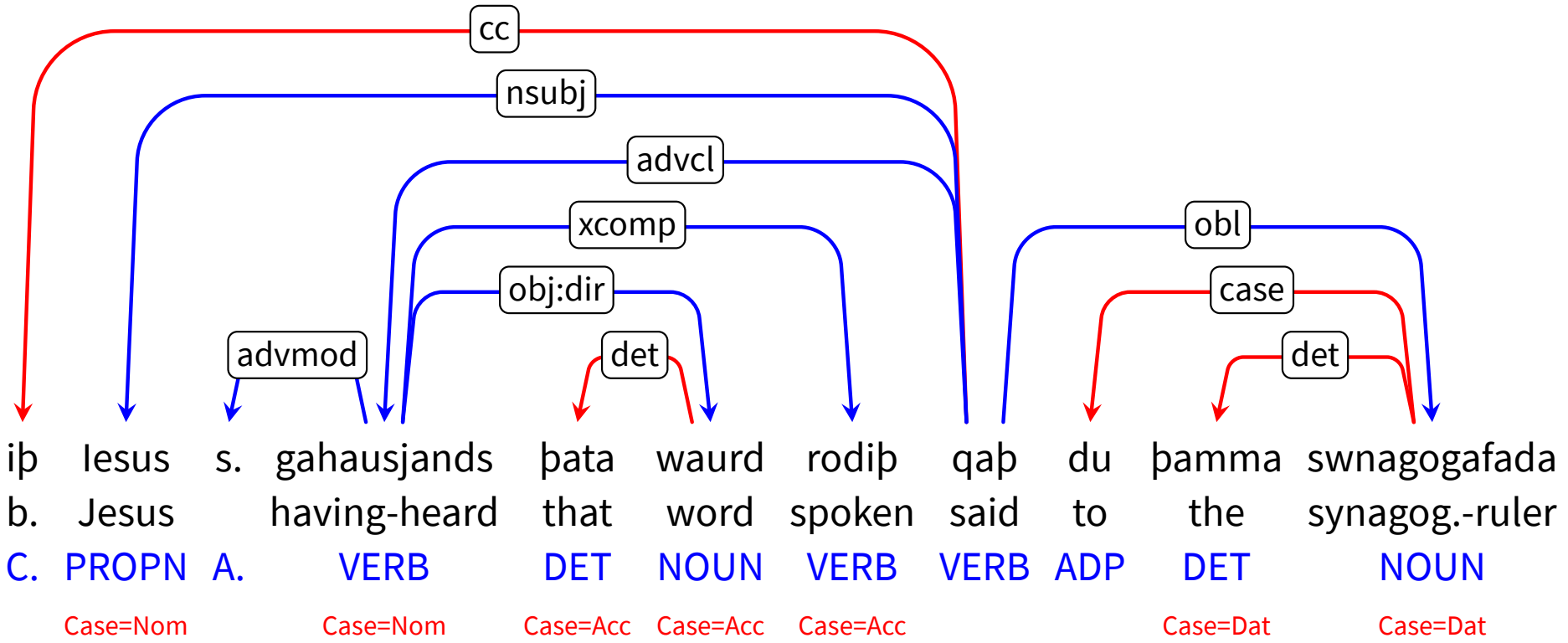
“But overhearing what they said, Jesus said to the ruler of the synagogue”
(Mark 5:36)

UD Latin PROIEL



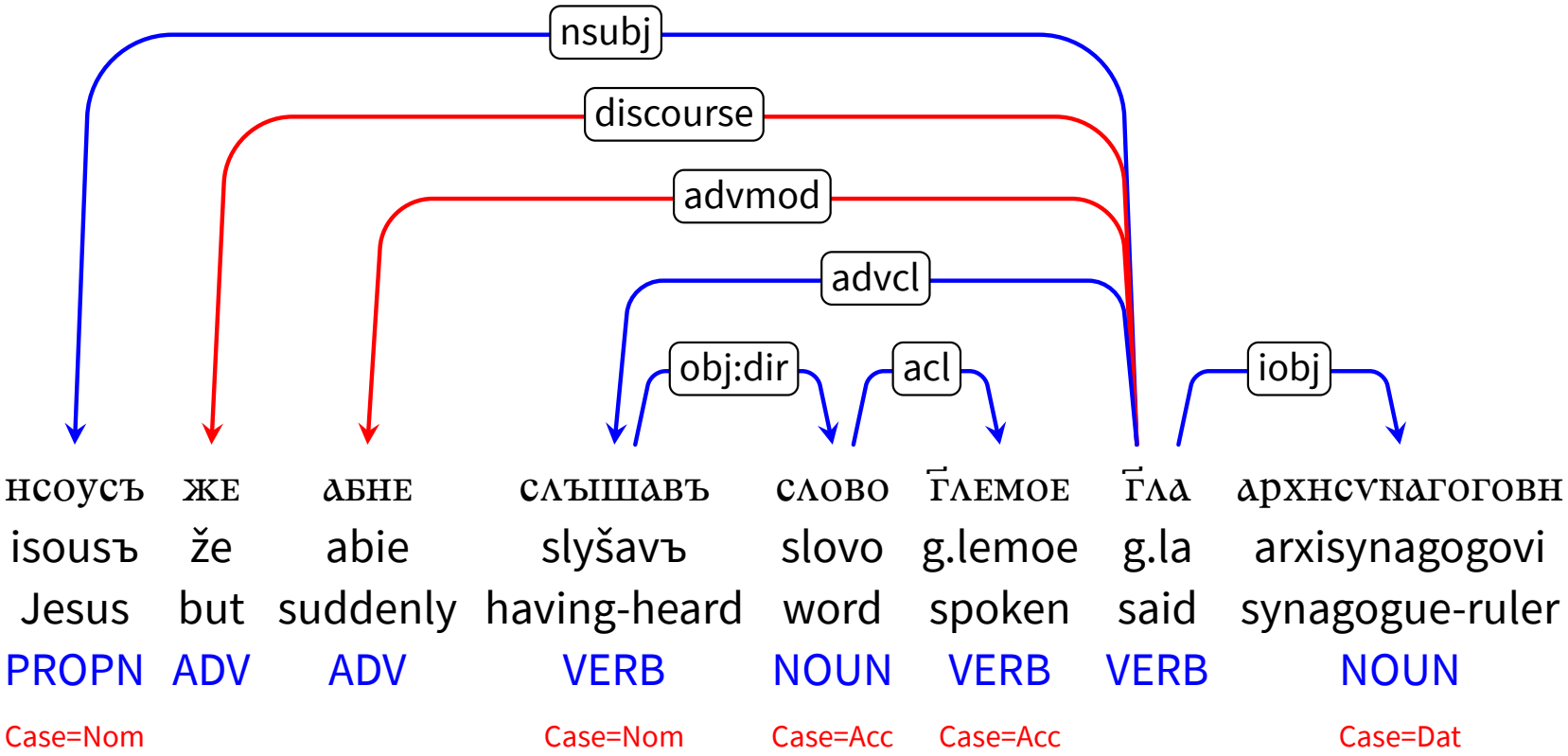
"But overhearing what they said, Jesus said to the ruler of the synagogue"
(Mark 5:36)

UD Gothic (PROIEL)



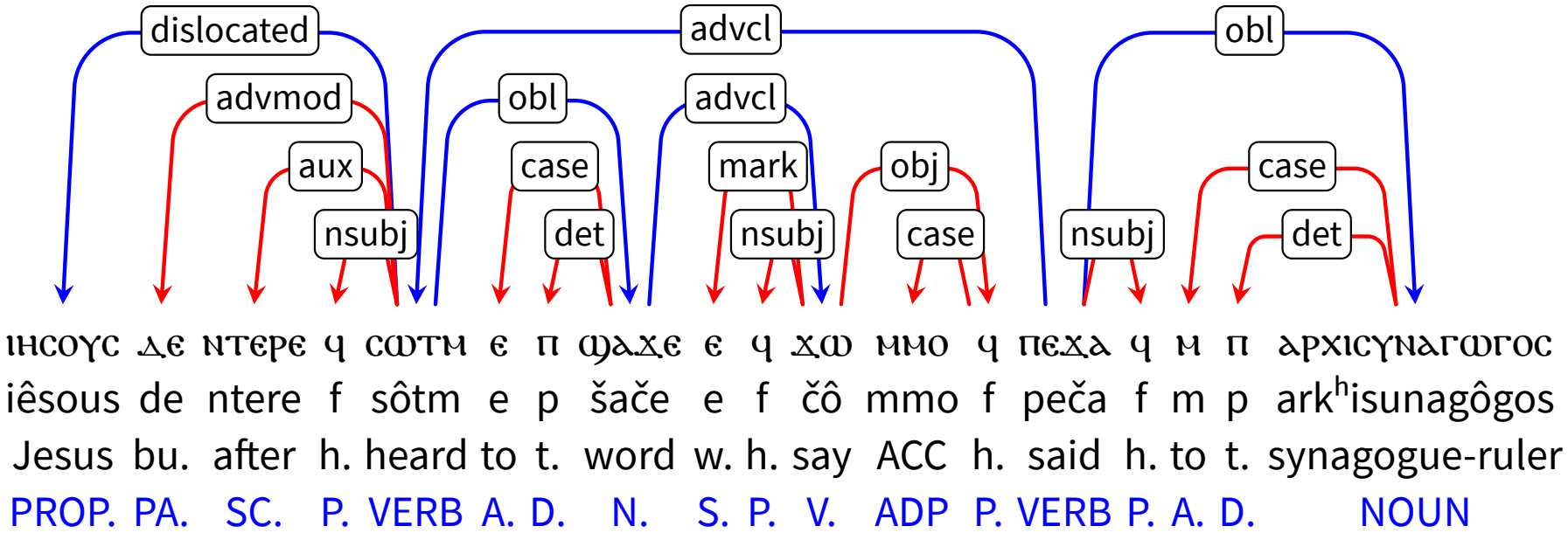
“But overhearing what they said, Jesus said to the ruler of the synagogue”
(Mark 5:36)

UD Old Church Slavonic (PROIEL)



“But overhearing what they said, Jesus said to the ruler of the synagogue”
(Mark 5:36)

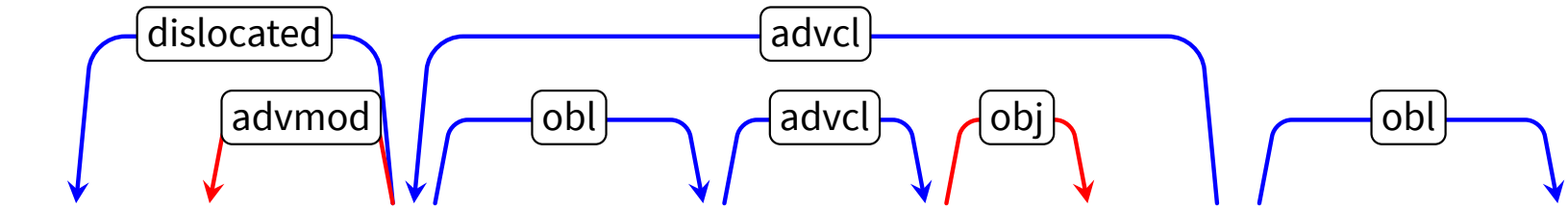
UD Coptic



iêsous de nterefsôtm epšače efčô mmof pečaf mpark^hisunagôgos

“But overhearing what they said, Jesus said to the ruler of the synagogue”
(Mark 5:36)

UD Coptic

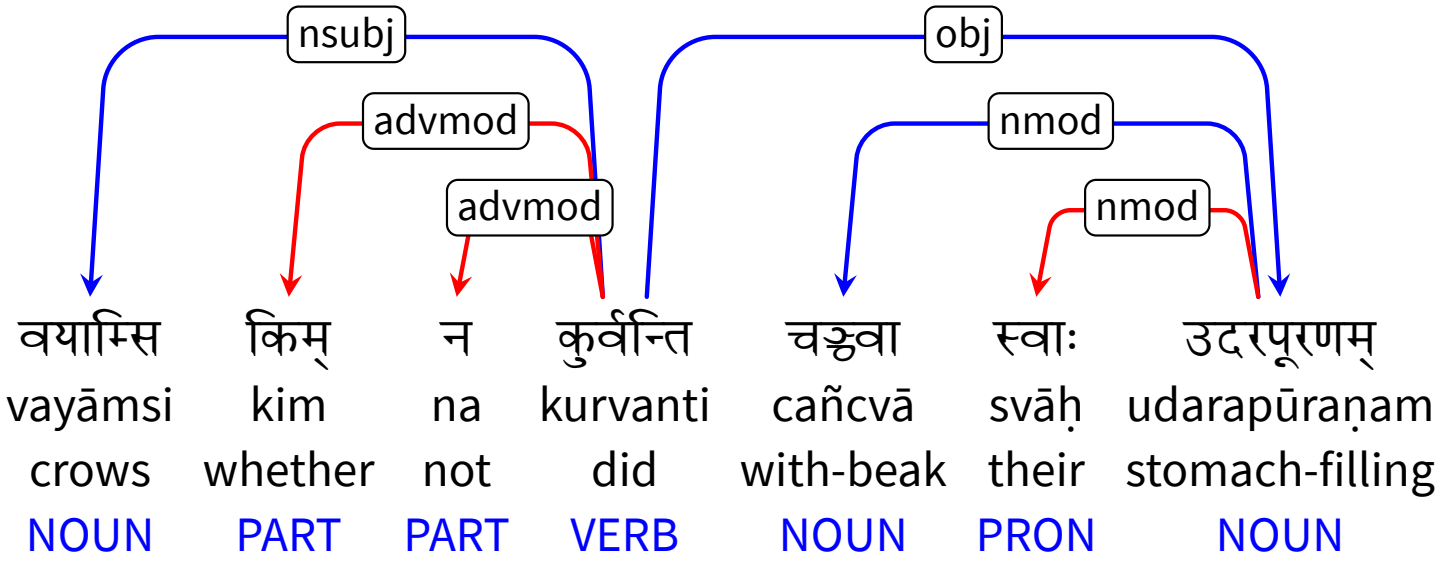


ⲓⲏⲥⲟⲩⲥ	ⲇⲉ	ⲛⲧⲉⲣⲉⲩⲥⲱⲧⲙ	ⲉⲡϣⲁϫⲉ	ⲉⲩϫⲱ	ⲙⲙⲟⲩ	ⲡⲉϫⲁⲩ	ⲙⲡⲁⲣⲭⲓⲥⲩⲛⲁⲅⲱⲅⲟⲥ
iêsous	de	nterefsôtm	epšače	efčô	mmof	pečaf	mpark^{h}isunagôgos
Jesus	but	after-he-heard	to-the-word	he-says	him	said	in-the-synagogue-ruler
PROPN	PA.	VERB	NOUN	VERB	PRON	VERB	NOUN

iêsous de nterefsôtm epšače efčô mmof pečaf mpark^{h}isunagôgos

“But overhearing what they said, Jesus said to the ruler of the synagogue”
(Mark 5:36)

UD Sanskrit



"Don't crows fill their bellies with their beaks?" (Panchatantra)

Conflicting Terminology & Language Change

- Past tense in modern East Slavic: -l / -la / -lo / -li, “finite” (no auxiliary).
- Past tense / active (“l”-)participle in West Slavic: auxiliary in 1st and 2nd persons. Also used in conditional mood.
- L-participle in Slovenian: auxiliary in all persons. Used for past, future and conditional.
- L-participle in old Slavic: perfect (resultative) aspect, and conditional.

она	пришла	к	нему
ona	prišla	k	nemu
she	came	to	him
PRON	VERB	ADP	PRON
	VerbForm=Fin		
	Tense=Past		

я	хотел	бы	поблагодарить
ja	hotel	by	poblagodarit’
I	like	would	to-thank
PRON	VERB	AUX	VERB
	VerbForm=Fin		VerbForm=Inf
	Tense=Past		

Conflicting Terminology & Language Change

- Past tense in modern East Slavic: -l / -la / -lo / -li, “finite” (no auxiliary).
- Past tense / active (“l”-)participle in West Slavic: auxiliary in 1st and 2nd persons. Also used in conditional mood.
- L-participle in Slovenian: auxiliary in all persons. Used for past, future and conditional.
- L-participle in old Slavic: perfect (resultative) aspect, and conditional.

přišel	jsi	nás	zabít
came	you-have	us	to-kill
VERB	AUX	PRON	VERB
VerbForm=Part	VerbForm=Fin		VerbForm=Inf
Tense=Past	Tense=Pres		

přišel	nás	zabít
he-came	us	to-kill
VERB	PRON	VERB
VerbForm=Part		VerbForm=Inf
Tense=Past		

Conflicting Terminology & Language Change

- Past tense in modern East Slavic: -l / -la / -lo / -li, “finite” (no auxiliary).
- Past tense / active (“l”-)participle in West Slavic: auxiliary in 1st and 2nd persons. Also used in conditional mood.
- L-participle in Slovenian: auxiliary in all persons. Used for past, future and conditional.
- L-participle in old Slavic: perfect (resultative) aspect, and conditional.

přišel	by	nás	zabít
came	he-would	us	to-kill
VERB	AUX	PRON	VERB
VerbForm=Part	VerbForm=Fin		VerbForm=Inf
Tense=Past	Mood=Cnd		

přišel	nás	zabít
he-came	us	to-kill
VERB	PRON	VERB
VerbForm=Part		VerbForm=Inf
Tense=Past		

Conflicting Terminology & Language Change

- Past tense in modern East Slavic: -l / -la / -lo / -li, “finite” (no auxiliary).
- Past tense / active (“l”-)participle in West Slavic: auxiliary in 1st and 2nd persons. Also used in conditional mood.
- L-participle in Slovenian: auxiliary in all persons. Used for past, future and conditional.
- L-participle in old Slavic: perfect (resultative) aspect, and conditional.

daleč	boste	přišli
far	you-will	come
ADV	AUX	VERB
	VerbForm=Fin	VerbForm=Part
	Tense=Fut	

prišel	je	do	ugotovitve
come	he-has	to	conclusion
VERB	AUX	ADP	NOUN
VerbForm=Part	VerbForm=Fin		
	Tense=Pres		

Conflicting Terminology & Language Change

- Past tense in modern East Slavic: -l / -la / -lo / -li, “finite” (no auxiliary).
- Past tense / active (“l”-)participle in West Slavic: auxiliary in 1st and 2nd persons. Also used in conditional mood.
- L-participle in Slovenian: auxiliary in all persons. Used for past, future and conditional.
- L-participle in old Slavic: perfect (resultative) aspect, and conditional.

пришелъ	есн	погоубнтъ	насъ	да	не	бн	отъшелъ	отъ	нихъ
prišelъ	jesi	pogubitъ	nasъ	da	ne	bi	otъšelъ	otъ	nichъ
come	you-have	to-kill	us	that	not	would	leave	from	them
VERB	AUX	VERB	PRON	SCONJ	PART	AUX	VERB	ADP	PRON
VerbForm=Part	VerbForm=Fin	VerbForm=Inf				Fin	VerbForm=Part		
Aspect=Res	Tense=Pres					Cnd	Aspect=Res		

Treebank Search Engines

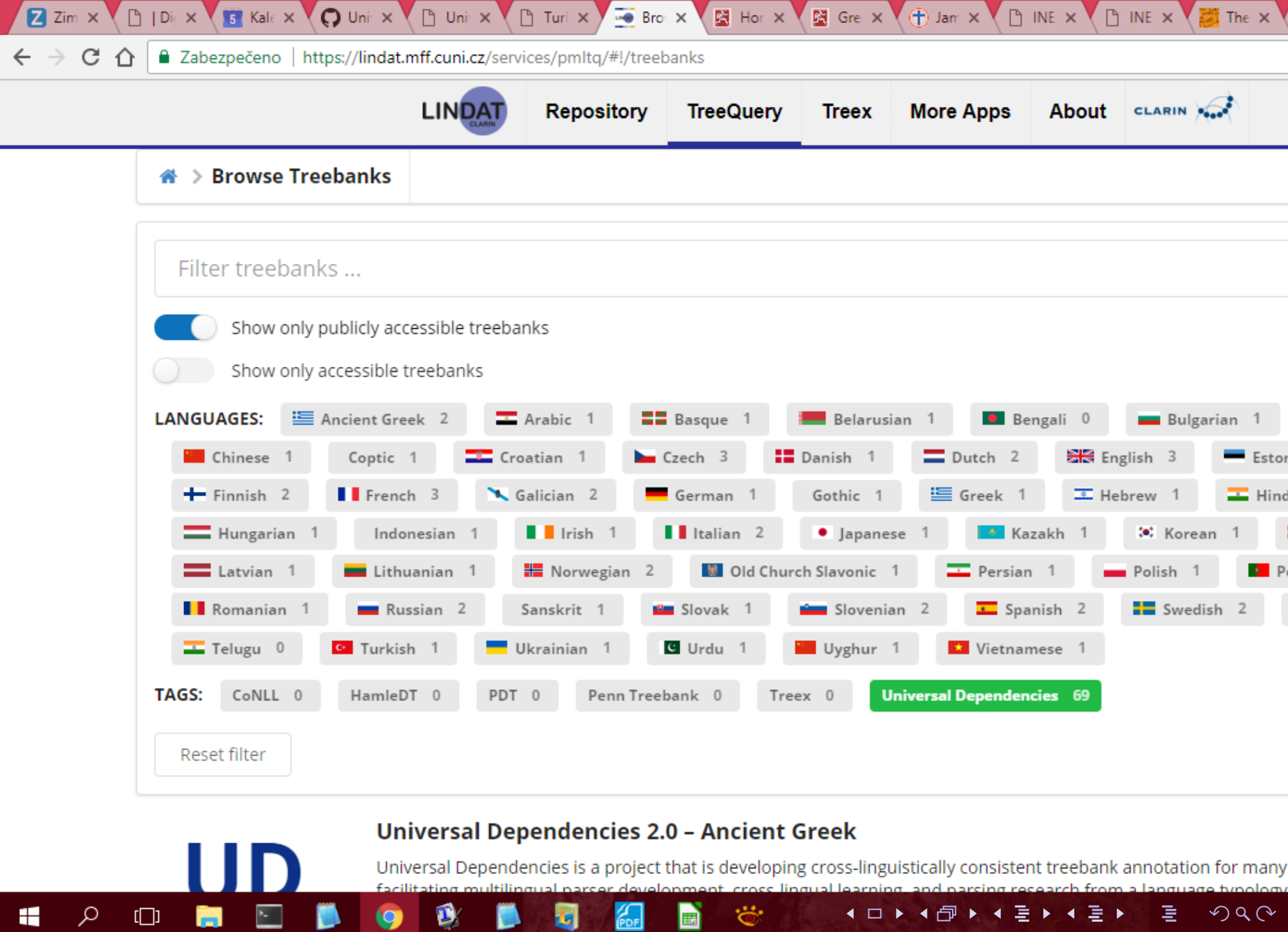
- PML-TQ
  (http://lindat.mff.cuni.cz/services/pmltq/)
- Turku dep search / SETS
  (http://bionlp-www.utu.fi/dep_search/)
- Kontext
  (http://lindat.mff.cuni.cz/services/kontext/)
- INESS
- Weblicht / Tündra

- UDPipe parsing demo
  (http://lindat.mff.cuni.cz/services/udpipe/)
- Udapi (Python, offline)
  (http://udapi.github.io/)

```
# sent_id = n01006011
# text = أخبر شاهد عيان الشرطة أن الضحية هاجم المشتبه به في شهر نيسان الماضي.
# original_text = أخبر شاهد عيان الشرطة أن الضحية هاجم المشتبه به في شهر نيسان الماضي.
# english_text = A witness told police that the victim had attacked the suspect in
```

أخبر >axobar_1 VERB VBC Aspect=Perf|Gender=Masc|Number=Sing|Person=3|Tense=Past
- شاهد $Ahid_1 NOUN NN Animacy=Hum|Case=Nom|Definite=Ind|Gender=Masc|Number=Sin
  - عيان EiyAn_1 ADJ JJ Case=Nom|Definite=Ind|Gender=Masc|Number=Sing amod _
- الشرطة $uroTap_1 NOUN NN Animacy=Nhum|Case=Acc|Definite=Def|Gender=Fem|Number=
  - أن >an~a_1 SCONJ IN _ mark _
  - الضحية DaHiy~ap_1 NOUN NN Animacy=Hum|Case=Acc|Definite=Def|Gender=Fem|Numbe
- هاجم hAjam_1 VERB VBC Aspect=Perf|Gender=Masc|Number=Sing|Person=3|Tense=Past
  - المشتبه mu$otabah_1 NOUN NN Animacy=Hum|Case=Acc|Definite=Def|Gender=Masc|Num
    - ب bi_1 ADP IN _ case SpaceAfter=No
    - ه h PRON PRP Case=Gen|Gender=Masc|Number=Sing|Person=3 nmod _
  - في fiy_1 ADP IN _ case _
  - شهر $ahor_1 NOUN NN Animacy=Nhum|Case=Gen|Definite=Def|Gender=Masc|Number=S
    - نيسان niysAn_1 PROPN NNP Animacy=Nhum|Gender=Masc|Number=Sing nmod:gmod _
      - الماضي mADiy_1 ADJ JJ Case=Gen|Definite=Def|Gender=Masc|Number=Sing amod
- . ._0 PUNCT . _ punct _

```
# sent_id = n01008017
# text = الأصوات طرحت يوم الانتخابات عام 1996 وفق مكتب الإحصاء. ويبدو أنها من المرجح أن ترتفع مجدداً هذا العام
# original_text = ت يوم الانتخابات عام 1996 وفق مكتب الاحصاء . ويبدو أنها من المرجح أن ترتفع مجددا هذا العام
```


Zabezpečeno | https://lindat.mff.cuni.cz/services/pmltq/#!/treebanks
LINDAT
Repository
TreeQuery
Treex
More Apps
About
CLARIN
Browse Treebanks
Filter treebanks ...
Show only publicly accessible treebanks
Show only accessible treebanks
LANGUAGES:
Ancient Greek 2
Arabic 1
Basque 1
Belarusian 1
Bengali 0
Bulgarian 1
Chinese 1
Coptic 1
Croatian 1
Czech 3
Danish 1
Dutch 2
English 3
Finnish 2
French 3
Galician 2
German 1
Gothic 1
Greek 1
Hebrew 1
Hungarian 1
Indonesian 1
Irish 1
Italian 2
Japanese 1
Kazakh 1
Korean 1
Latvian 1
Lithuanian 1
Norwegian 2
Old Church Slavonic 1
Persian 1
Polish 1
Romanian 1
Russian 2
Sanskrit 1
Slovak 1
Slovenian 2
Spanish 2
Swedish 2
Telugu 0
Turkish 1
Ukrainian 1
Urdu 1
Uyghur 1
Vietnamese 1
TAGS:
CoNLL 0
HamleDT 0
PDT 0
Penn Treebank 0
Treex 0
Universal Dependencies 69
Reset filter
Universal Dependencies 2.0 – Ancient Greek
Universal Dependencies is a project that is developing cross-linguistically consistent treebank annotation for many

Zabezpečeno | https://lindat.mff.cuni.cz/services/pmltq/#!/treebank/ud20_la/query/IYWgdg9gJgpgBAEgA5wFwF44G0DGALASwBso5RJZEc1MsBdA

LINDAT CLARIN | Repository | TreeQuery | Treex | More Apps | About | CLARIN

Browse Treebanks > Universal Dependencies 2.0 – Latin | My queries | Public queries | Query | Help

Undo | Repeat

Query list: **UD demo queries** | 3/7 | Query: **Nonprojectivity** | + Save as... | Rename | Save

Relations | Node Types | Attributes | Operators | Functions

```
a-node $p :=
[
  child a-node $c := [],
  same-tree-as a-node $x :=
  [
    !ancestor $p,
    (order-follows $p and order-precedes $c) or (order-follows $c and order-precedes $p)
  ]
];
```

Execute query | w/o Filters

Previous | 1 | of 100 | Next | 1 a-node $p | 2 a-node $c | 3 a-node $x

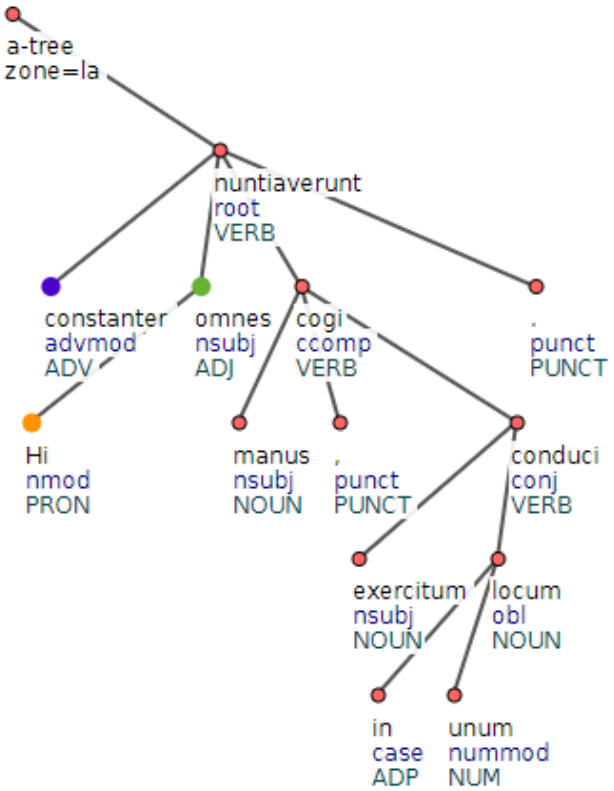
[la] Hi constanter omnes nuntiaverunt manus cogi, exercitum in unum locum conduci.

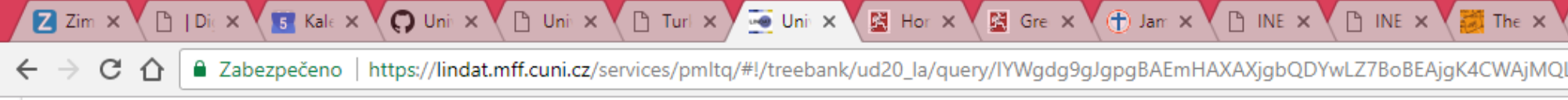
Zabezpečeno | https://lindat.mff.cuni.cz/services/pmltq/#!/treebank/ud20_la/query/IYWgdg9gJgpgBAEgA5wFwF44G0DGALASwBso5RJZEc1MsBdA

Execute query | w/o Filters

Previous | 1 | of 100 | Next | 1 a-node $p | 2 a-node $c | 3 a-node $x

[la] Hi constanter omnes nuntiaverunt manus cogi, exercitum in unum locum conduci.



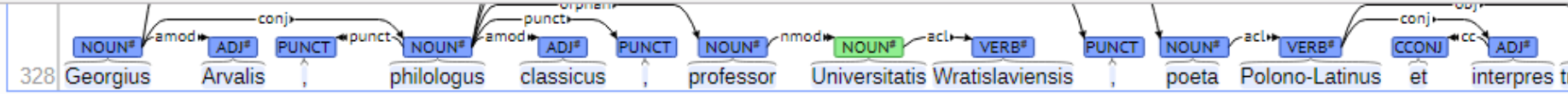


Relations ▾ Node Types ▾ Attributes ▾ Operators ▾ Functions ▾

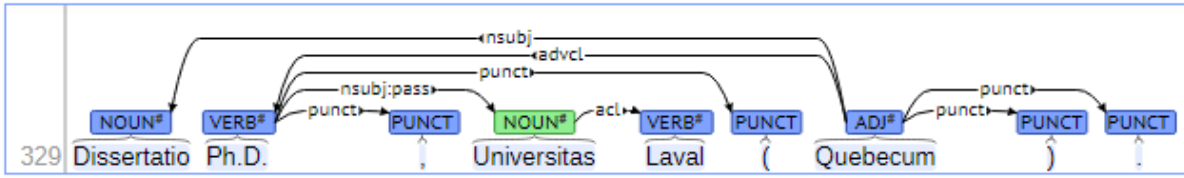
```
a-node $a := [lemma="qui1"]
>> for $a.lemma, $a.iset/case, lower($a.form) give $1, $2, $3, count() sort by $4 desc;
```

Execute query | w/o Filters

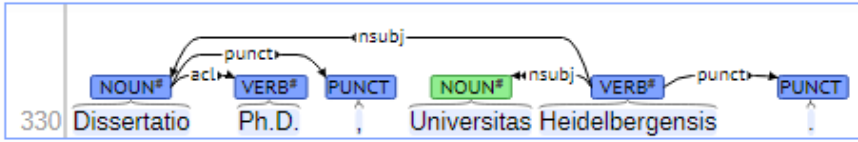
qui1	nom	qui	91
qui1	nom	quae	41
qui1	acc	quae	23
qui1	acc	quem	21
qui1	acc	quod	21
qui1	acc	quam	19
qui1	acc	quos	15
qui1	nom	quod	13
qui1	acc	quas	9
qui1	abl	quo	9



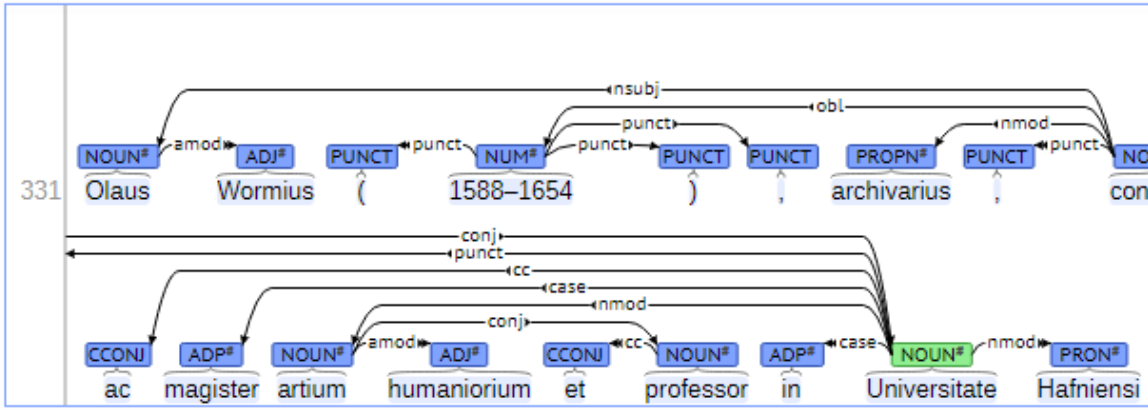
[context] [conllu]



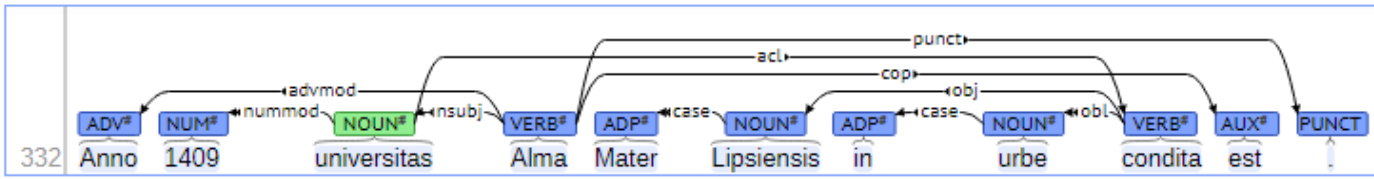
[context] [conllu]



[context] [conllu]



[context] [conllu]



Wrapping Up

Wrapping Up

- UD tries to cover morphology and surface syntax

Wrapping Up

- UD tries to cover morphology and surface syntax
- You need extra stuff?
- Everyone does! Please go ahead!

Wrapping Up

- UD tries to cover morphology and surface syntax

- You need extra stuff?
- Everyone does! Please go ahead!

- But if possible, think of convertibility to/from UD
- You get visibility
- The world gets your data

Thank You, Community!

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Aljoscha Burchardt, Marie Candito, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Silvie Cinková, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Eric de la Clergerie, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Tomaž Erjavec, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, 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, Jan Hajič jr., Linh Hà Mỹ, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Pinkey Nainwani, Anna Nedoluzhko, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Larissa Rinaldi, Laura Rituma, Rudolf Rosa, Mike Rosner, Davide Rovati, Benoit Sagot, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Hanzhi Zhu

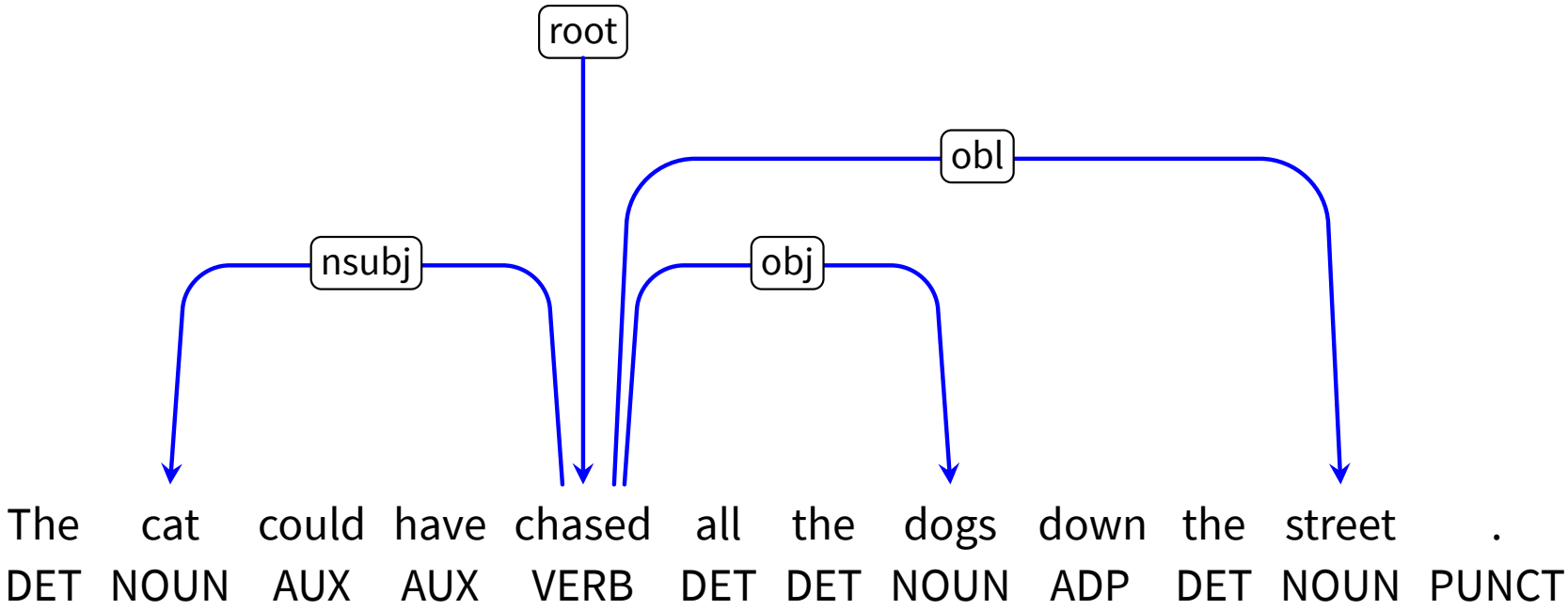
Thank You! Questions?

- PML-TQ
  (http://lindat.mff.cuni.cz/services/pmltq/)
- Turku dep search / SETS
  (http://bionlp-www.utu.fi/dep_search/)
- UDPipe parsing demo
  (http://lindat.mff.cuni.cz/services/udpipe/)
- Udapi (Python, offline)
  (http://udapi.github.io/)

Syntax

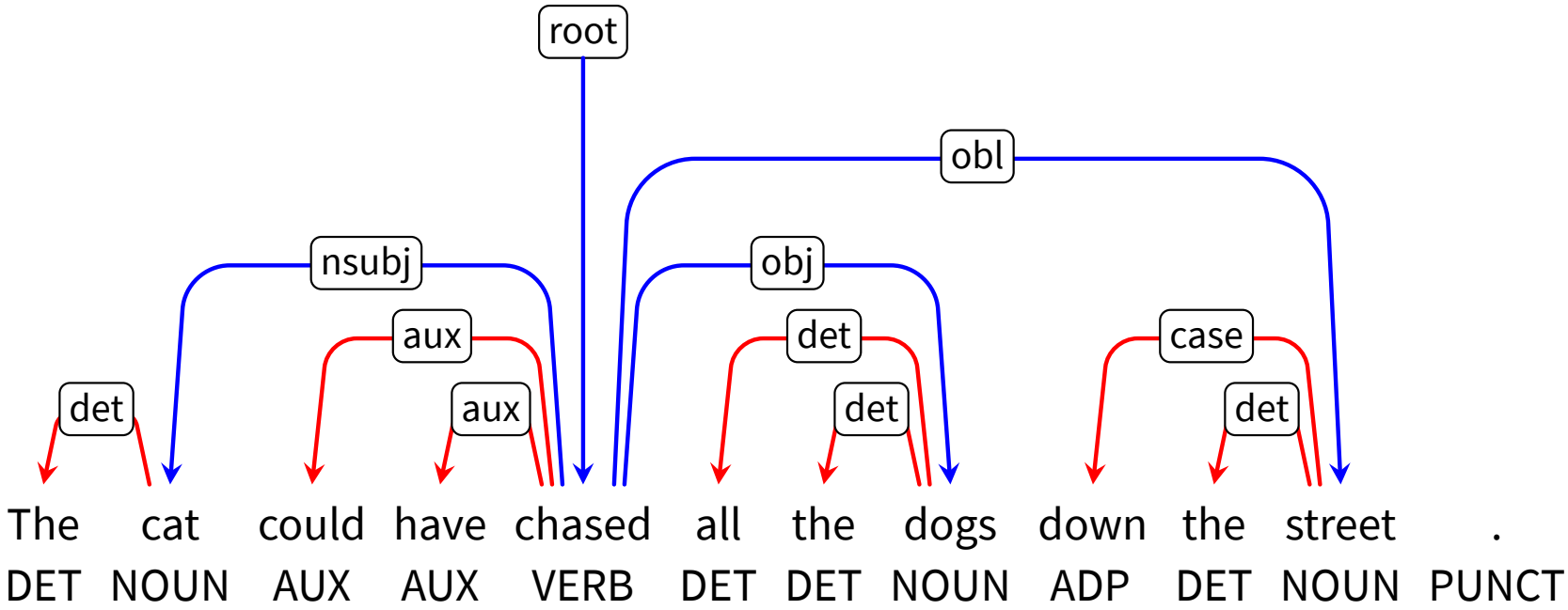
The	cat	could	have	chased	all	the	dogs	down	the	street	.
DET	NOUN	AUX	AUX	VERB	DET	DET	NOUN	ADP	DET	NOUN	PUNCT

Syntax



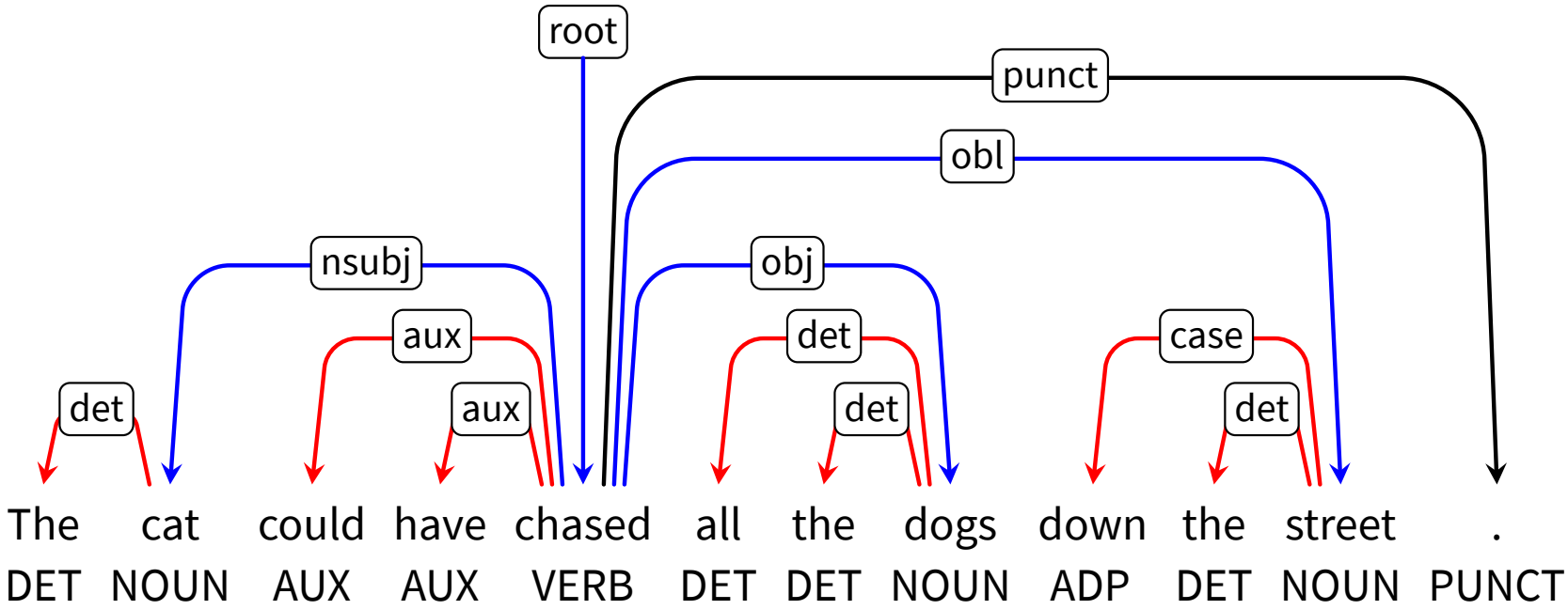
- Content words are related by dependency relations

Syntax



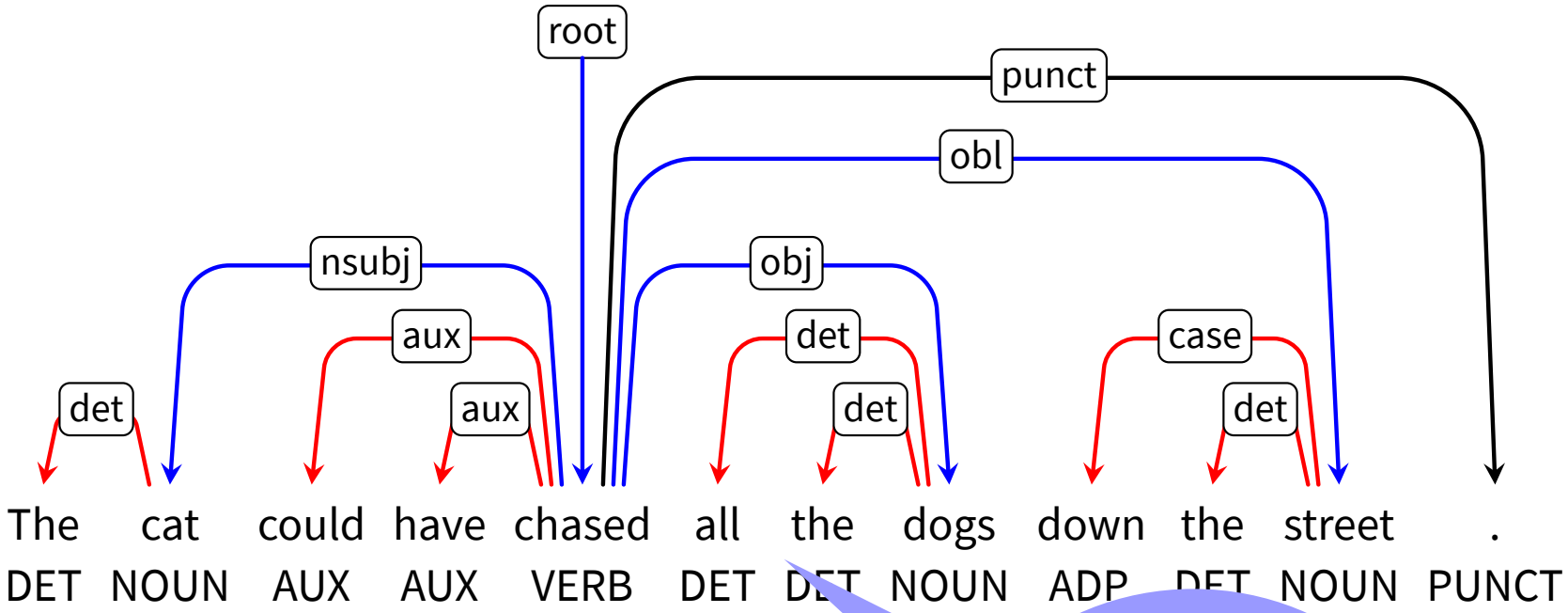
- Content words are related by dependency relations
- Function words attach to closest content words

Syntax

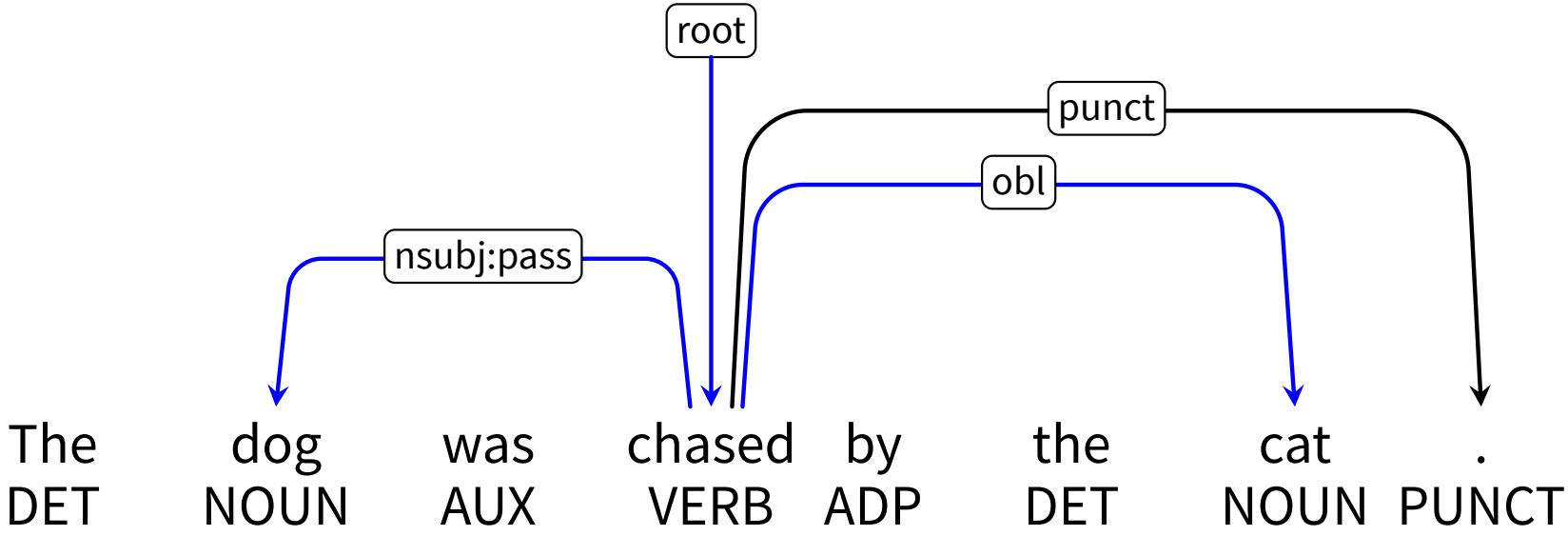


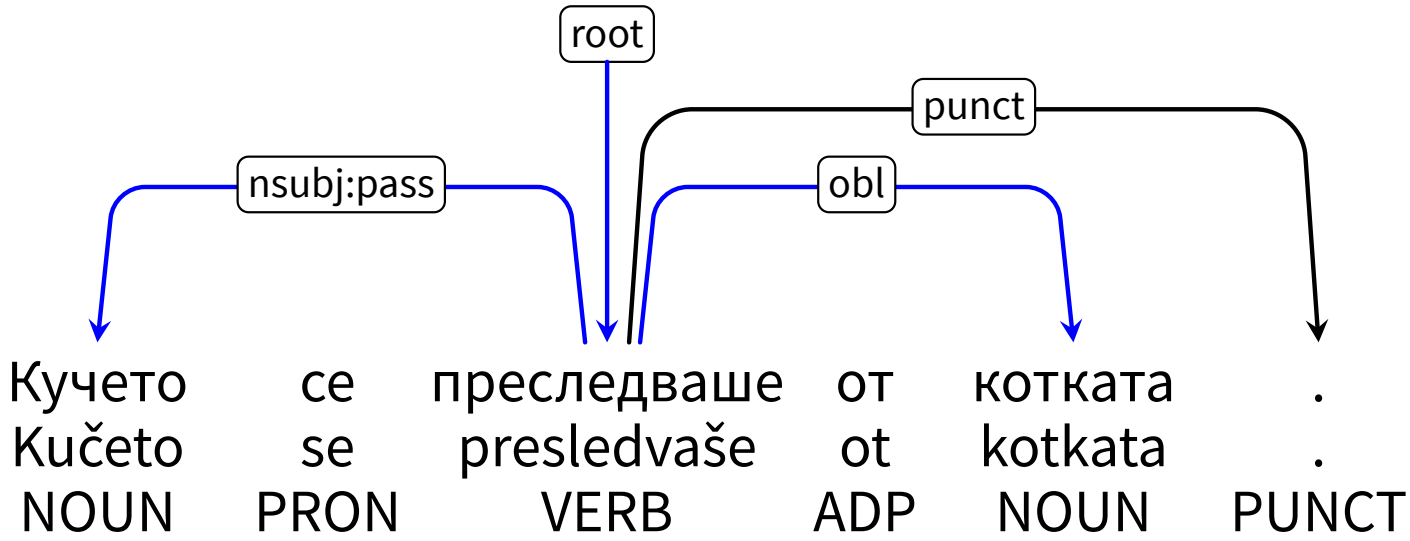
- Content words are related by dependency relations
- Function words attach to closest content words
- Punctuation attach to head of phrase or clause

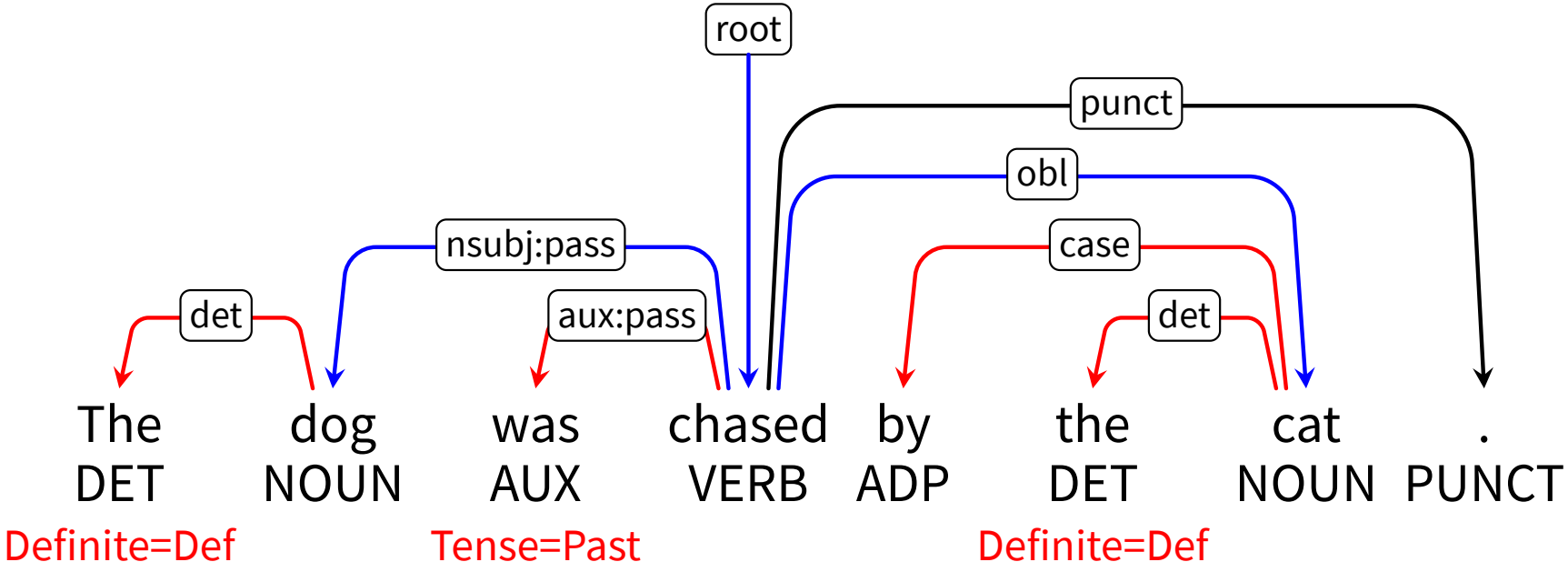
Syntax

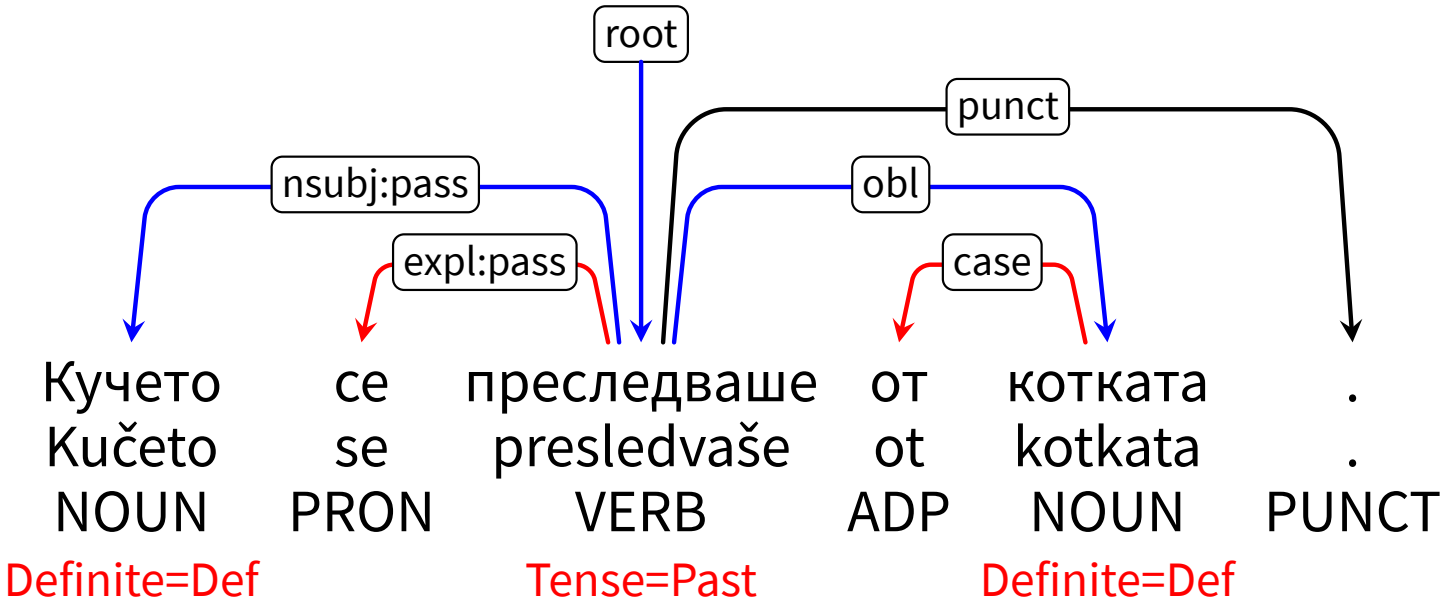


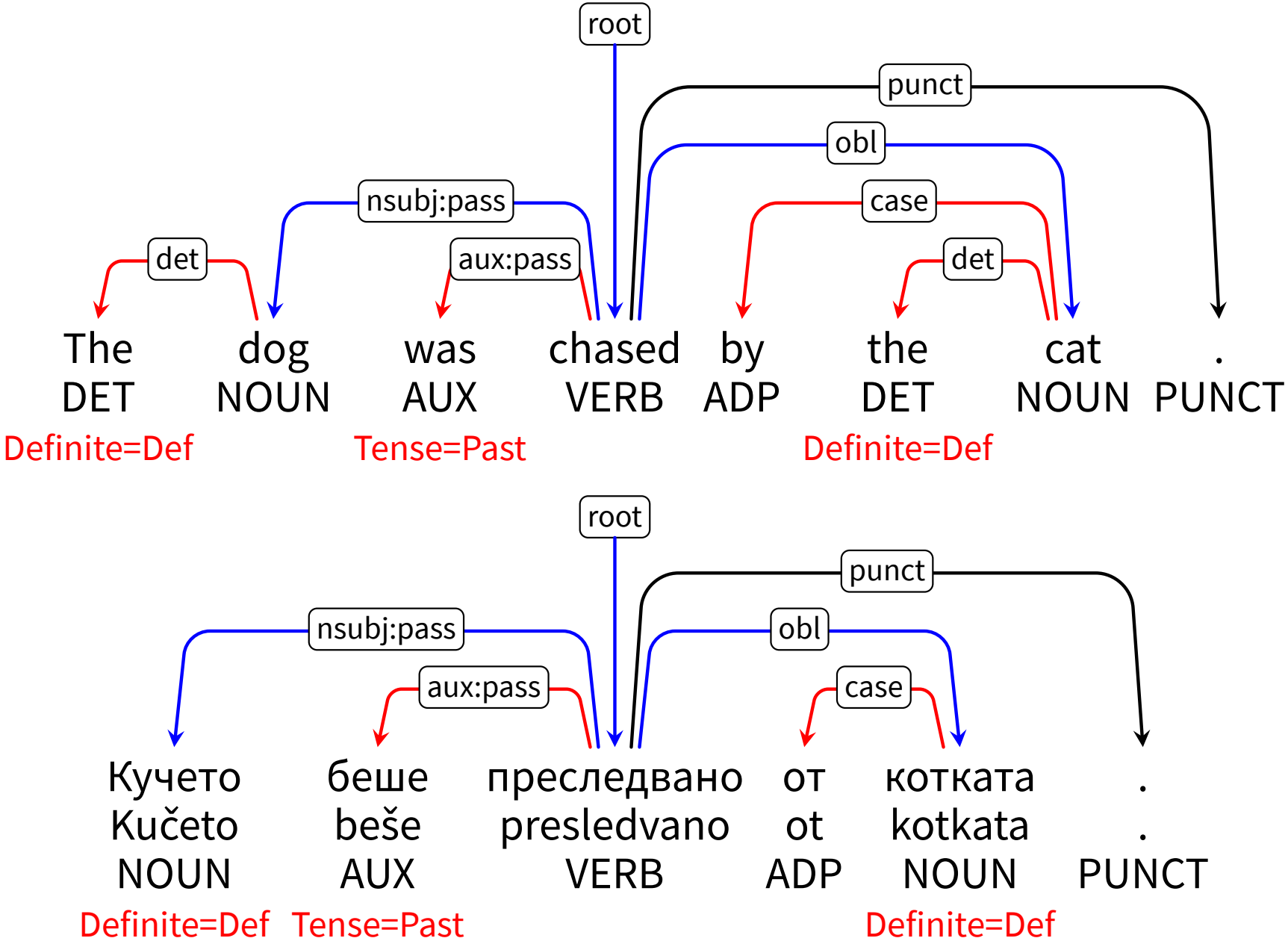
Not "dependency" in the strictly syntactic sense!

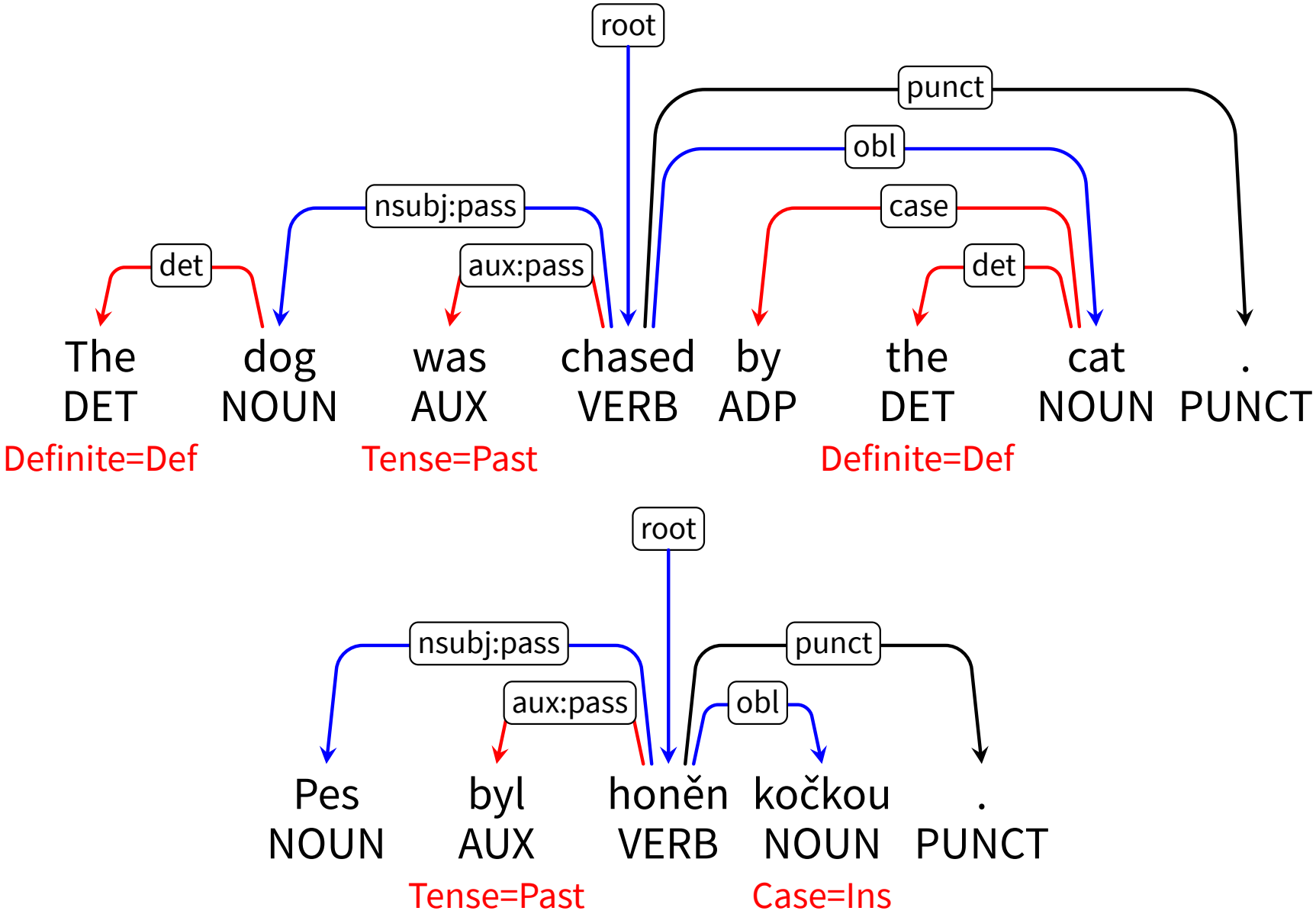

root
punct
obl
nsubj:pass
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT


root
punct
obl
nsubj:pass
Кучето се преследваше от котката .
Kučeto se presledvaše ot kotkata .
NOUN PRON VERB ADP NOUN PUNCT


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


root
punct
nsubj:pass
obl
expl:pass
case
Кучето Kučeto NOUN
се se PRON
преследваше presledvaše VERB
от ot ADP
котката kotkata NOUN
. . PUNCT
Definite=Def
Tense=Past
Definite=Def


root
punct
obl
nsubj:pass
case
det
aux:pass
det
The DET
dog NOUN
was AUX
chased VERB
by ADP
the DET
cat NOUN
. PUNCT
Definite=Def
Tense=Past
Definite=Def
root
punct
nsubj:pass
obl
aux:pass
case
Кучето
Kučeto
NOUN
беше
beše
AUX
преследвано
presledvano
VERB
от
ot
ADP
котката
kotkata
NOUN
.
.
PUNCT
Definite=Def
Tense=Past
Definite=Def


root
punct
obl
nsubj:pass
case
det
aux:pass
det
The DET
dog NOUN
was AUX
chased VERB
by ADP
the DET
cat NOUN
. PUNCT
Definite=Def
Tense=Past
Definite=Def
root
nsubj:pass
punct
aux:pass
obl
Pes NOUN
byl AUX
honěn VERB
kočkou NOUN
. PUNCT
Tense=Past
Case=Ins

Dependency Relations

- Taxonomy of 37 universal grammatical relations, broadly attested in language typology (de Marneffe et al., 2014)
  - Language-specific subtypes may be added

Dependency Relations

- Taxonomy of 37 universal grammatical relations, broadly attested in language typology (de Marneffe et al., 2014)
  - Language-specific subtypes may be added
- Organizing principles
  - Three types of structures: nominals, clauses, modifiers
  - Core arguments vs. other dependents (not arguments vs. adjuncts)

Dependents of Clausal Predicates

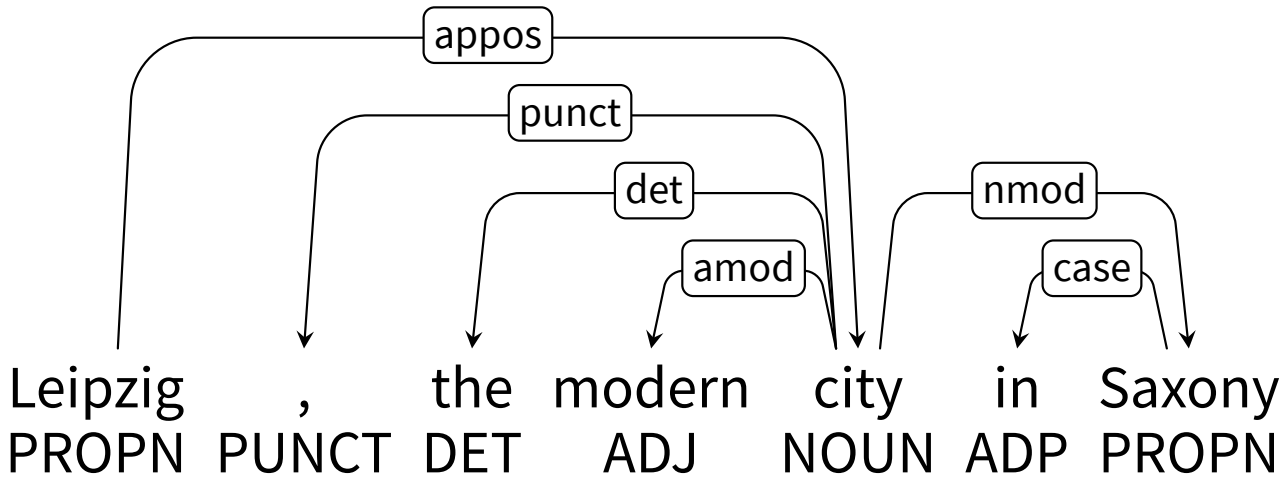
	Nominal	**Clausal**	**Modifier**	**Function**
Core	*nsubj*	*csubj*		
	obj	*ccomp*		
	iobj	*xcomp*		
Non-Core	*obl*	*advcl*	*advmod*	*aux*
	vocative		*discourse*	*cop*
	dislocated			*mark*
	expl			

Dependents of Non-predicative Adjectives and Adverbs

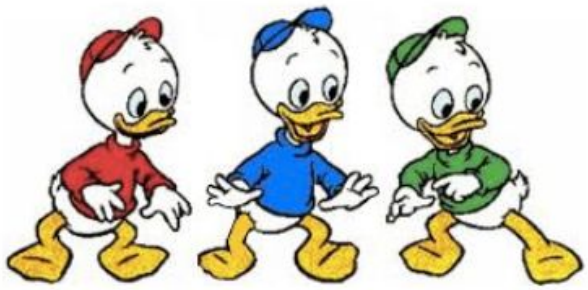
	Nominal	**Clausal**	**Modifier**
Core	*obj*	*ccomp*	
	iobj	*xcomp*	
Non-Core	*obl*	*advcl*	*advmod*

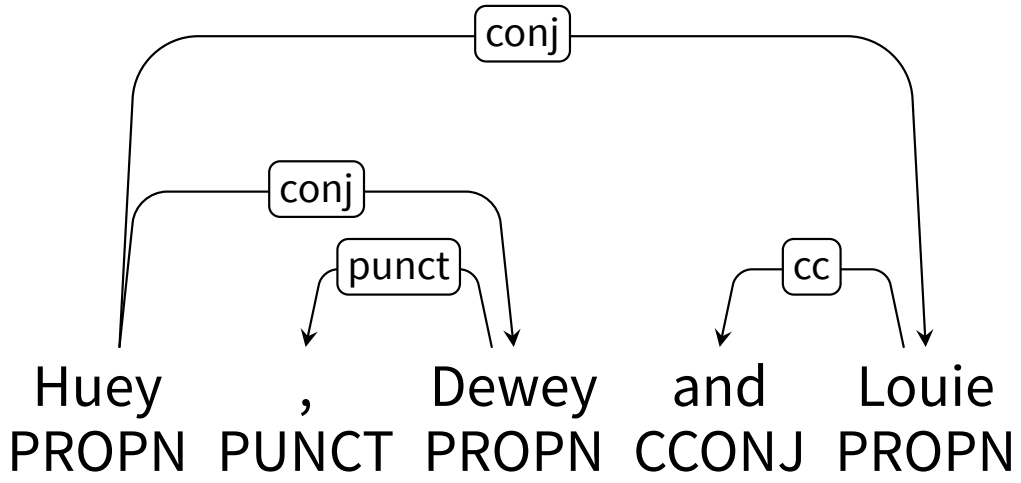
Dependents of Nominals

Nominal	***Clausal***	***Other***
nmod	*acl*	*amod*
appos		*det*
clf		*nummod*
		case



Coordination





- Coordinate structures are headed by the first conjunct
  - Subsequent conjuncts depend on it via the conj relation
  - Conjunctions depend on the next conjunct via the cc relation
  - Punctuation marks depend on the next conjunct via the punct relation

Multiword Expressions

Relation	***Examples***
fixed	*in spite of, as well as, ad hoc*
flat	*president Havel, New York, four thousand*
compound	*phone book, dress up*
goeswith	*notwith standing, with out*

- UD annotation almost does not permit "words with spaces"
  - Multiword expressions are analyzed using special relations
  - The fixed, flat and goeswith relations are always head-initial
  - The compound relation reflects the internal structure
- Words with spaces may be allowed in v2:
  - Vietnamese (spaces delimit syllables, not words)
  - Numbers ("1 000 000")
  - Possibly other approved cases, e.g. multi-word abbreviations

Other Relations

Relation	***Explanation***
parataxis	*Loosely linked clauses of same rank*
list	*Lists without syntactic structure*
orphan	*Orphans in ellipsis linked together*
reparandum	*Disfluency linked to (speech) repair*
dep	*Unspecified dependency*
root	*Syntactically independent element of clause/phrase*

Language-Specific Relations

- Language-specific relations are subtypes of universal relations added to capture important phenomena
- Subtyping permits us to "back off" to universal relations

Language-Specific Relations

Relation	***Explanation***
acl:relcl	*Relative clause*
compound:prt	*Verb particle (dress up)*
nmod:poss	*Possessive nominal (Mary 's book)*
obl:agent	*Agent in passive (saved by the bell)*
cc:preconj	*Preconjunction (both … and)*
det:predet	*Predeterminer (all those …)*

Word Segmentation

- Must be reproducible on new data
- Surface tokens vs. syntactic words
- Chinese, Vietnamese etc.: no clues, non-trivial algorithm
- Arabic, Sanskrit etc.: part of morphological analysis
- Spanish, German etc.: rather limited cases of contractions
- Others: only punctuation (low-level tokenization)

Word Segmentation

Vamos	*nos*	*a*	*el*	*mar*	*.*
VERB	*PRON*	*ADP*	*DET*	*NOUN*	*PUNCT*

Vámonos	*al*	*mar*	*.*
VERB+PRON	*ADP+DET*	*NOUN*	*PUNCT*

Manning's Law

The secret to understanding the design and current success of UD is to realize that the design is a very subtle compromise between approximately 6 things:

- UD needs to be satisfactory on linguistic analysis grounds for individual languages.
- UD needs to be good for linguistic typology, i.e., providing a suitable basis for bringing out cross-linguistic parallelism across languages and language families.
- UD must be suitable for rapid, consistent annotation by a human annotator.
- UD must be suitable for computer parsing with high accuracy.
- UD must be easily comprehended and used by a non-linguist, whether a language learner or an engineer with prosaic needs for language processing. … it leads us to favor traditional grammar notions and terminology.
- UD must support well downstream language understanding tasks (relation extraction, reading comprehension, machine translation, …)

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