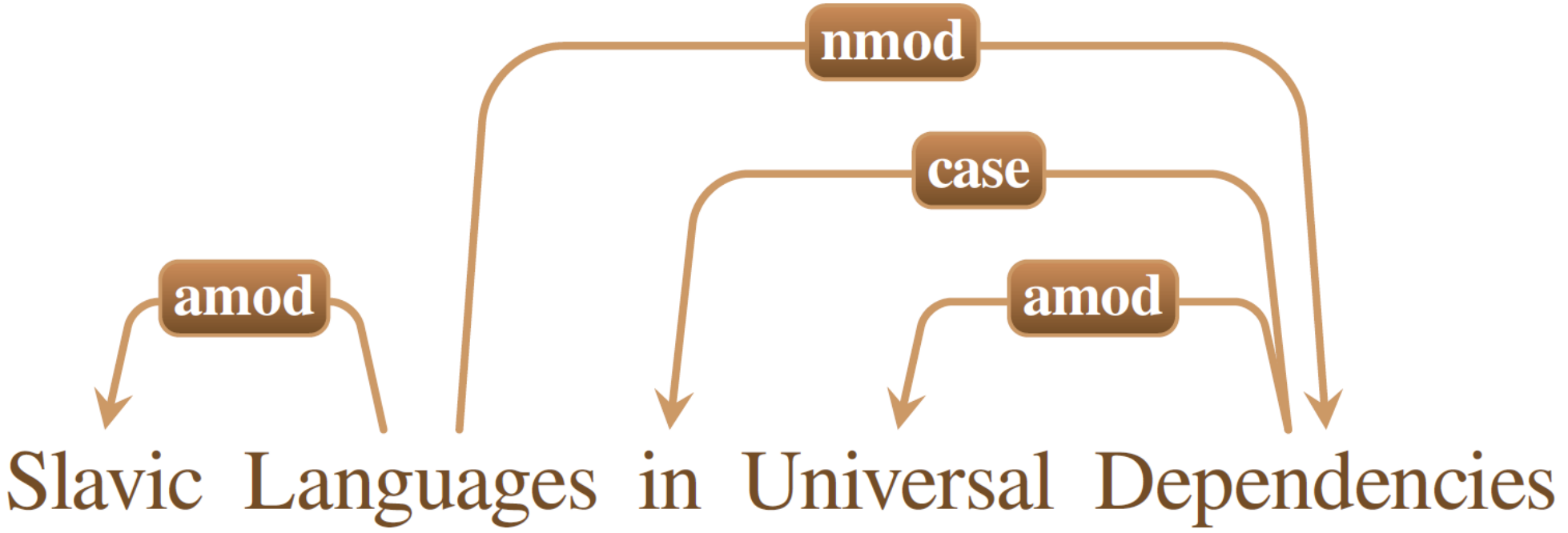
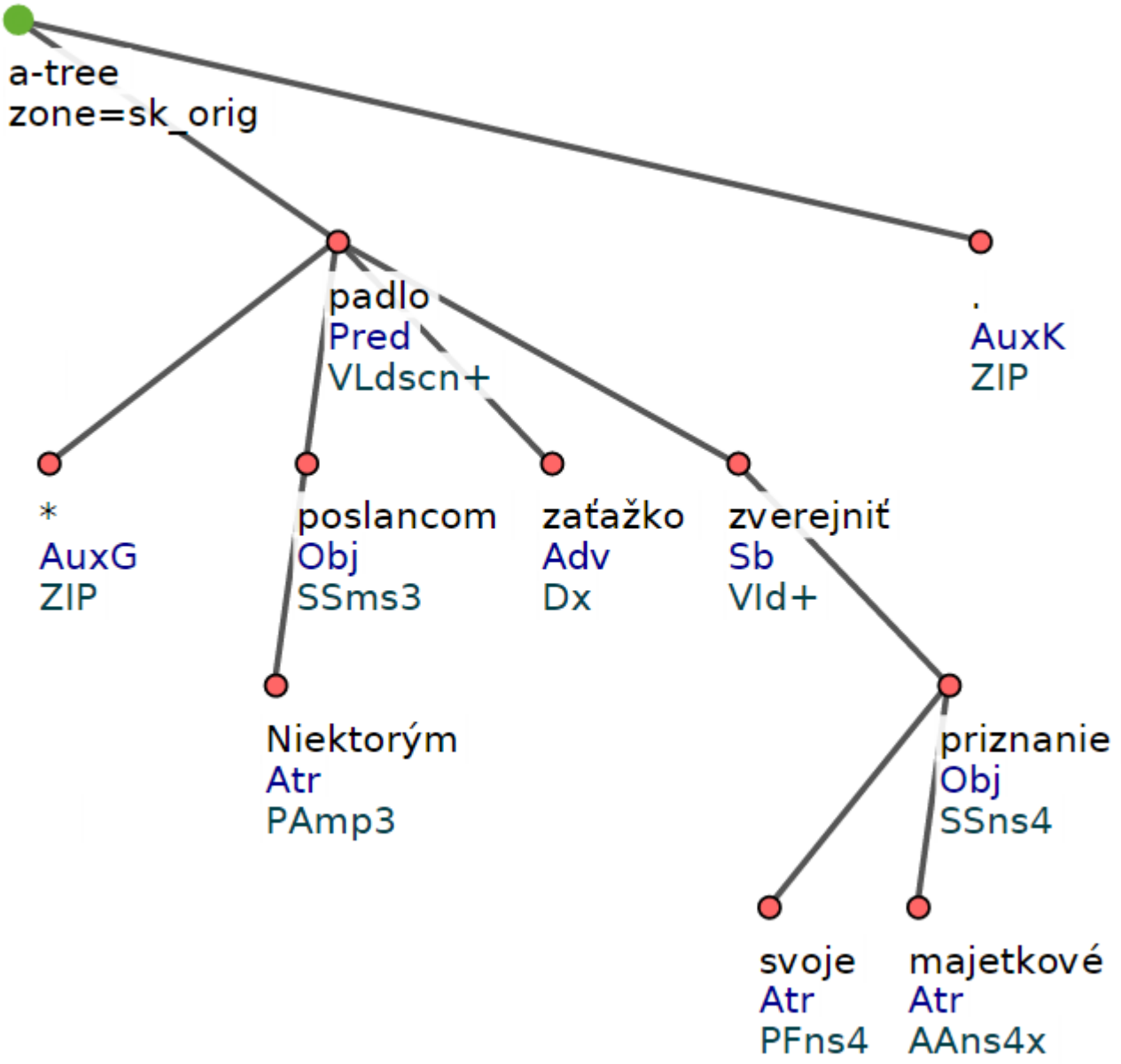


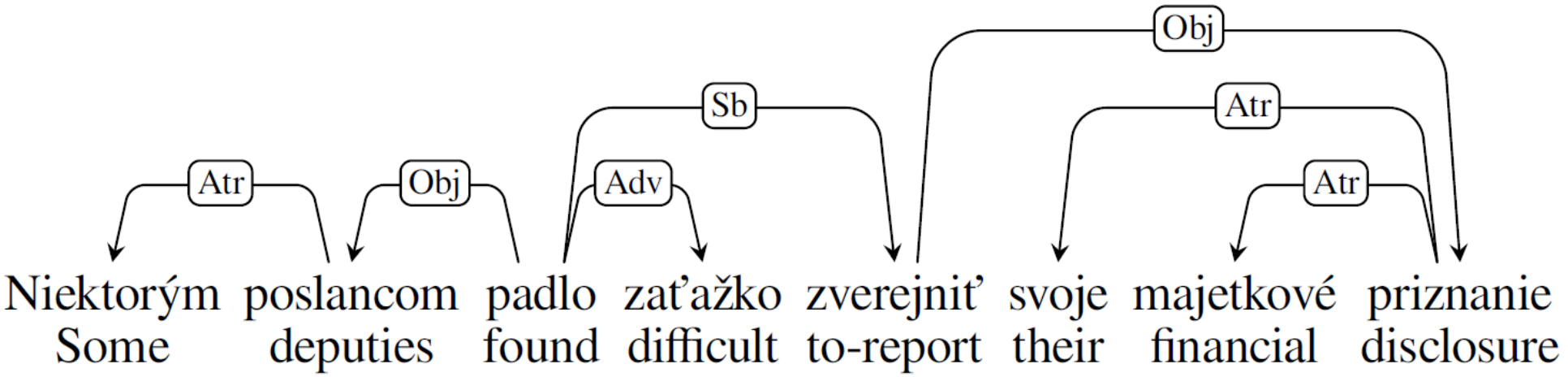
Daniel Zeman
Charles University in Prague

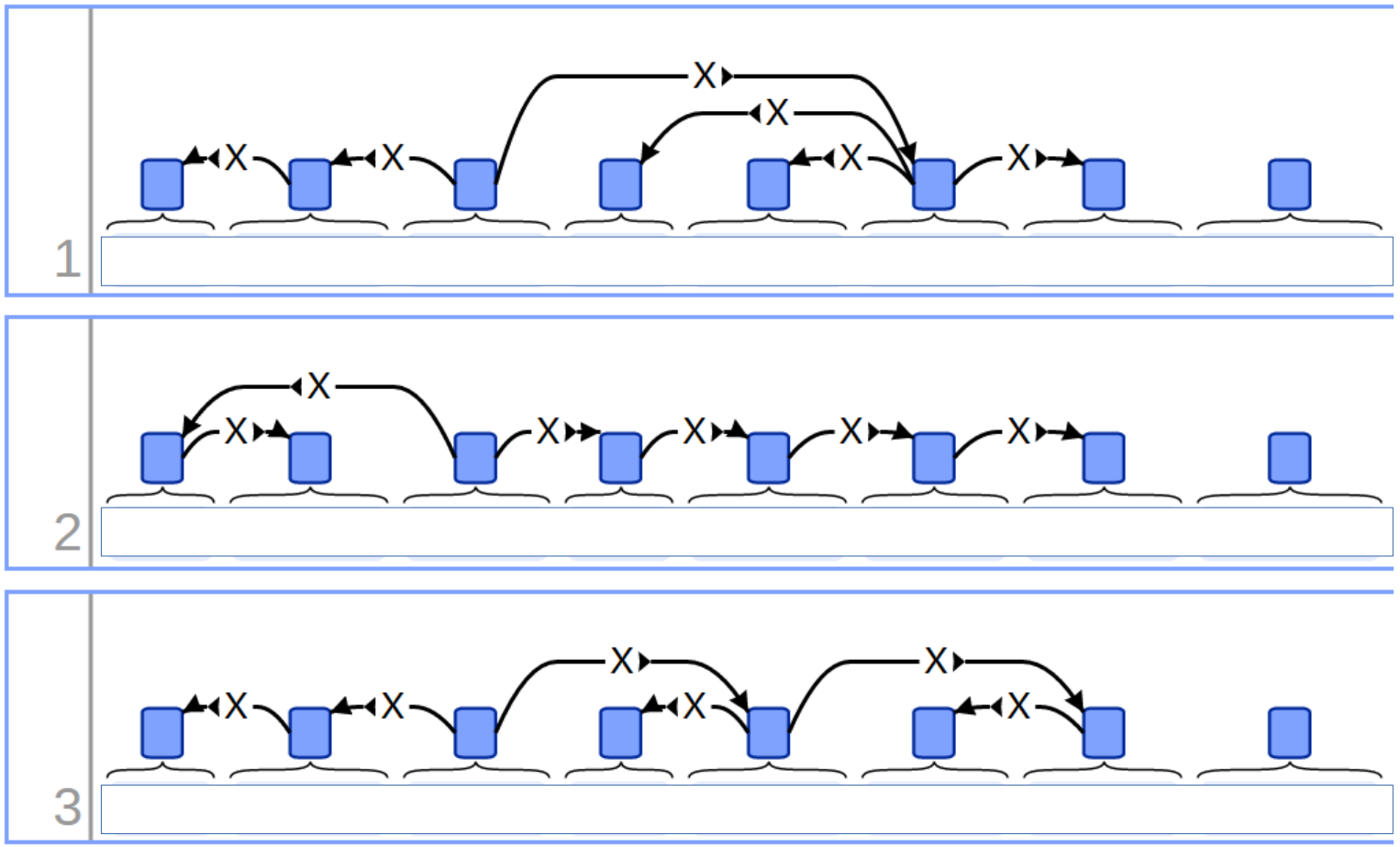
based on joint work with 50+ people

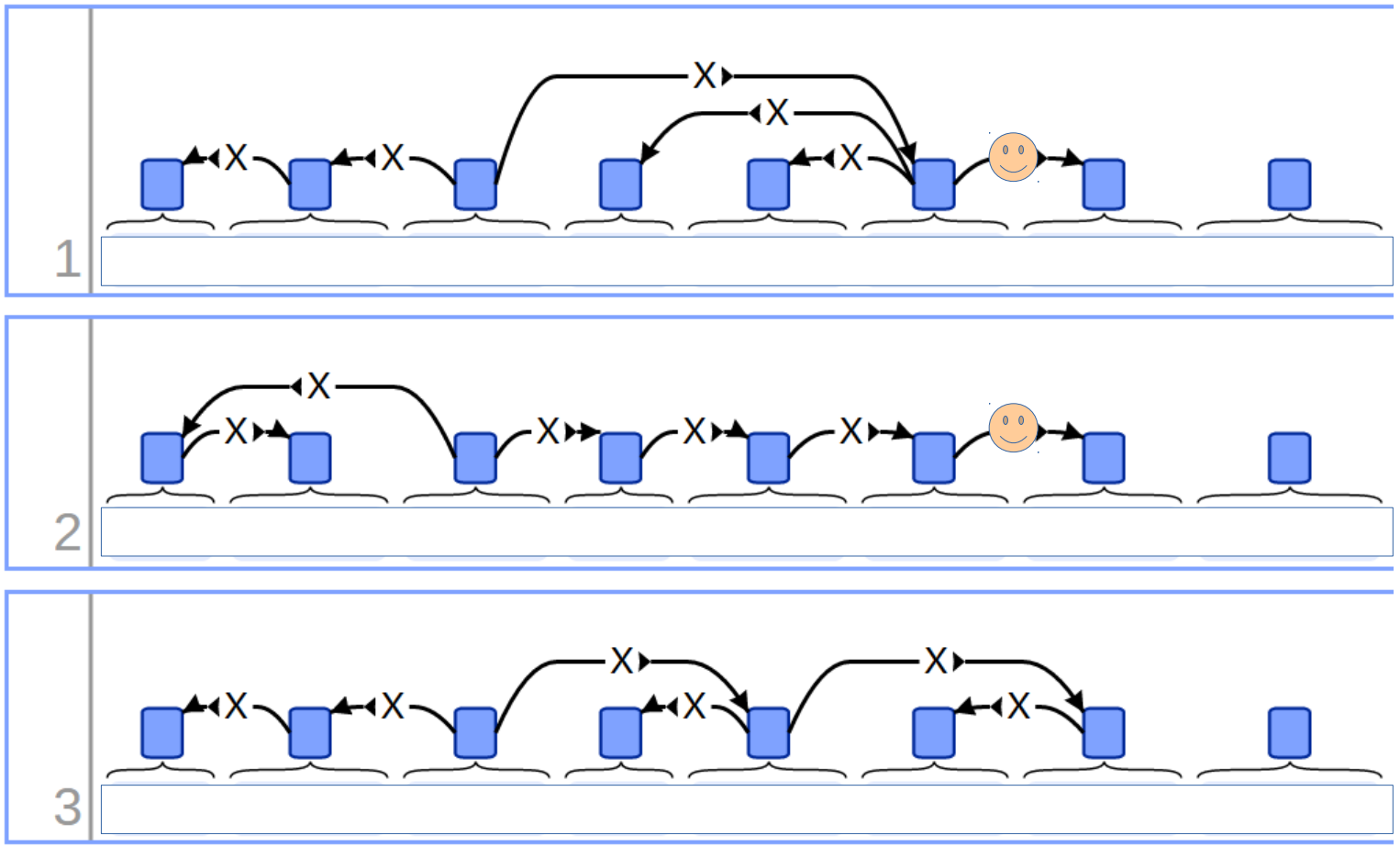
Dependency Treebanks

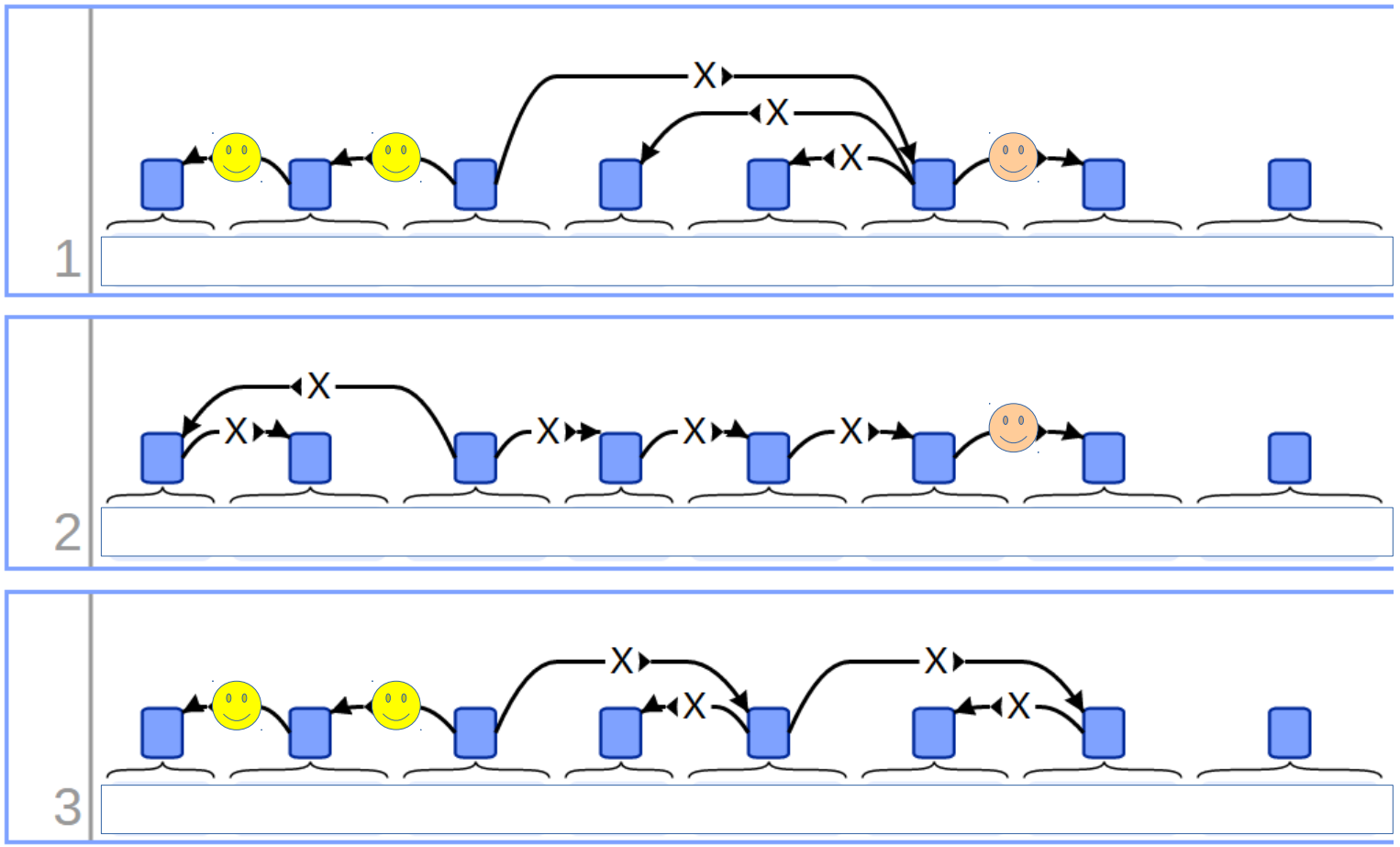


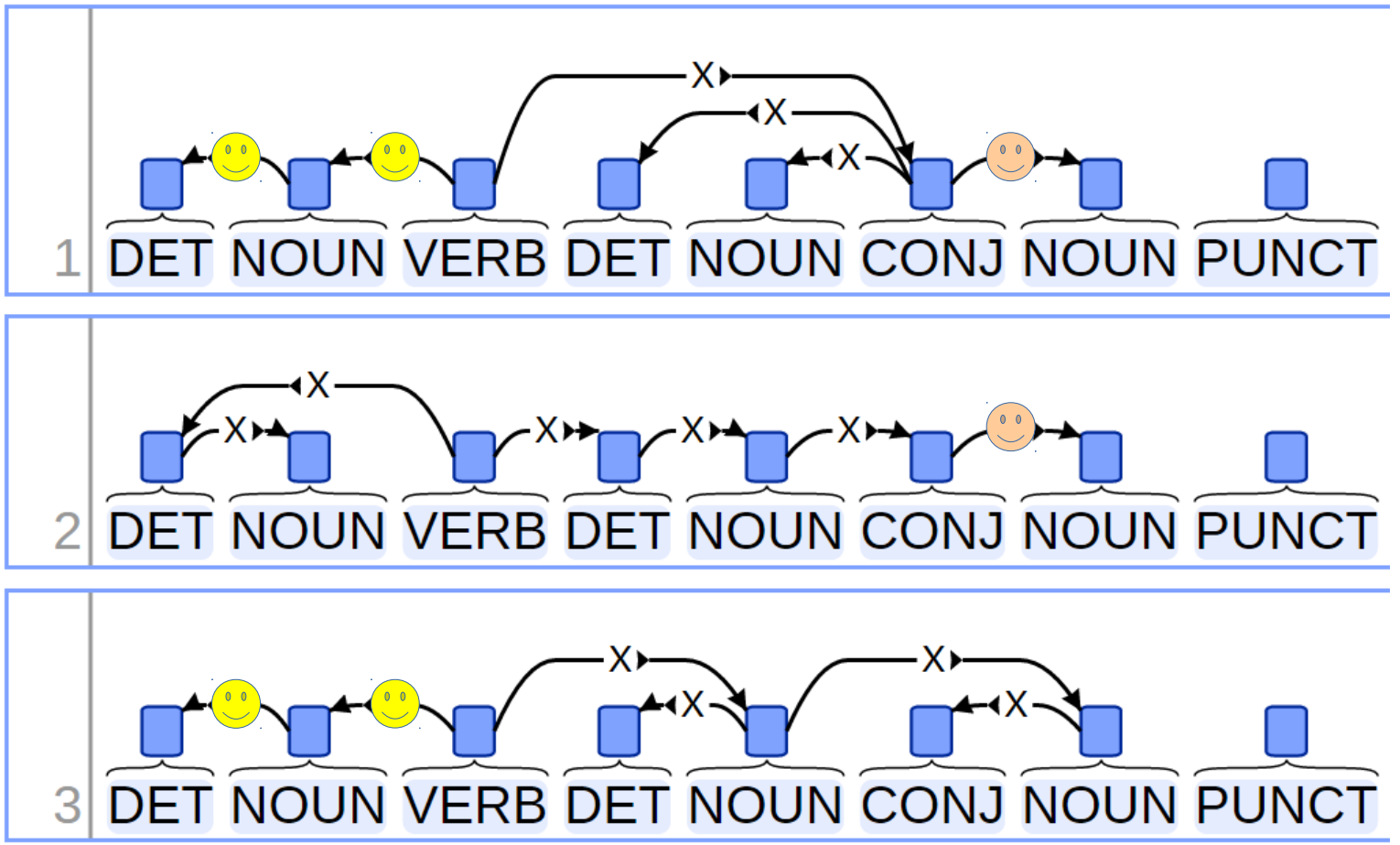
Dependency Treebanks

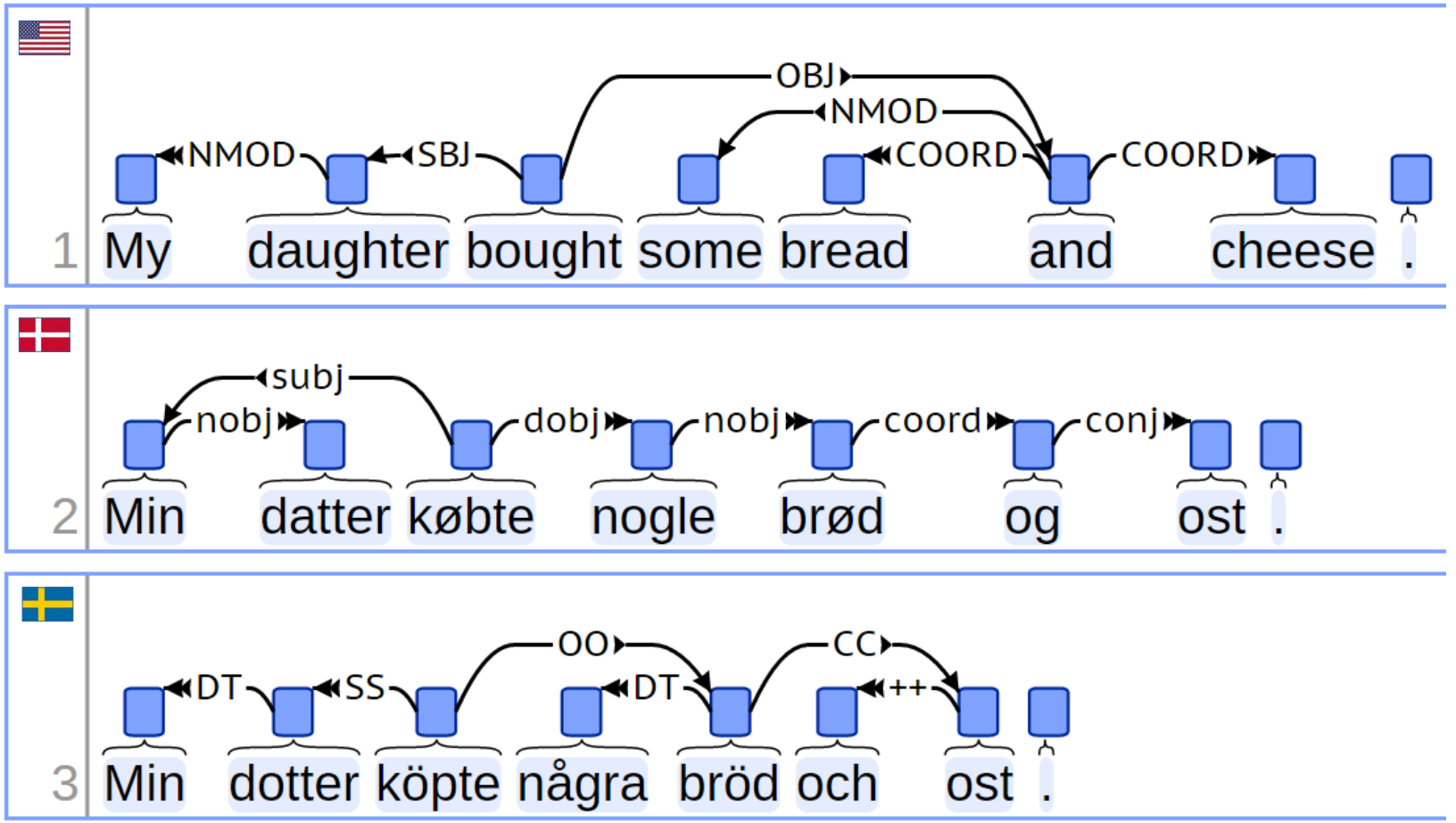



1
X
X
X
X
X
X
2
X
X
X
X
X
X
3
X
X
X
X
X
X


1
X
X
X
X
X
2
X
X
X
X
X
3
X
X
X
X
X


X
X
X
1
X
X
X
X
X
2
X
X
X
X
3


1
DET NOUN VERB DET NOUN CONJ NOUN PUNCT
X
X
X
2
DET NOUN VERB DET NOUN CONJ NOUN PUNCT
X
X
X
X
X
X
3
DET NOUN VERB DET NOUN CONJ NOUN PUNCT
X
X
X
X

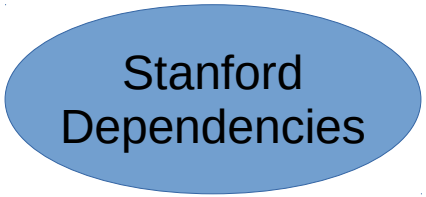

1 My daughter bought some bread and cheese .
NMOD SBJ OBJ NMOD COORD COORD
2 Min datter købte nogle brød og ost .
subj nobj dobj nobj coord conj
3 Min dotter köpte några bröd och ost .
DT SS OO DT CC ++

Universal Dependencies

http://universaldependencies.github.io/docs/

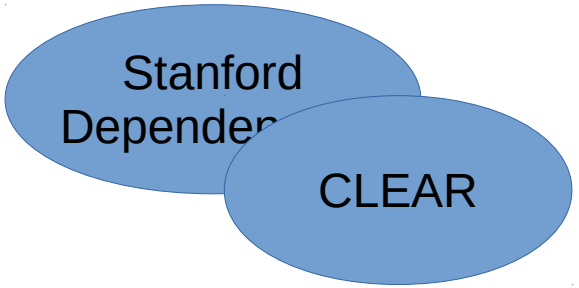
Universal Dependencies

http://universaldependencies.github.io/docs/



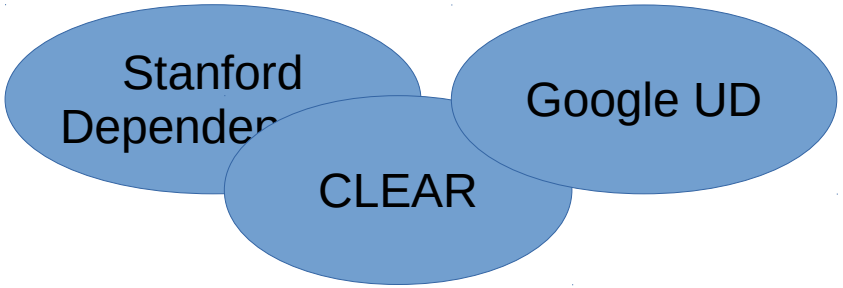
Universal Dependencies

http://universaldependencies.github.io/docs/



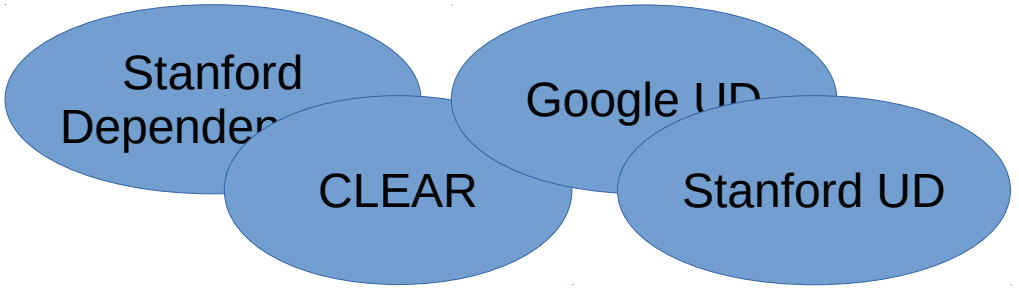
Universal Dependencies

http://universaldependencies.github.io/docs/



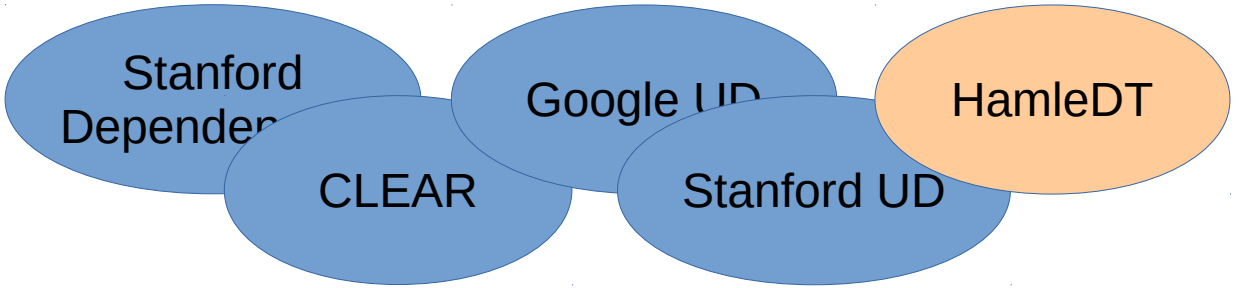
Universal Dependencies

http://universaldependencies.github.io/docs/



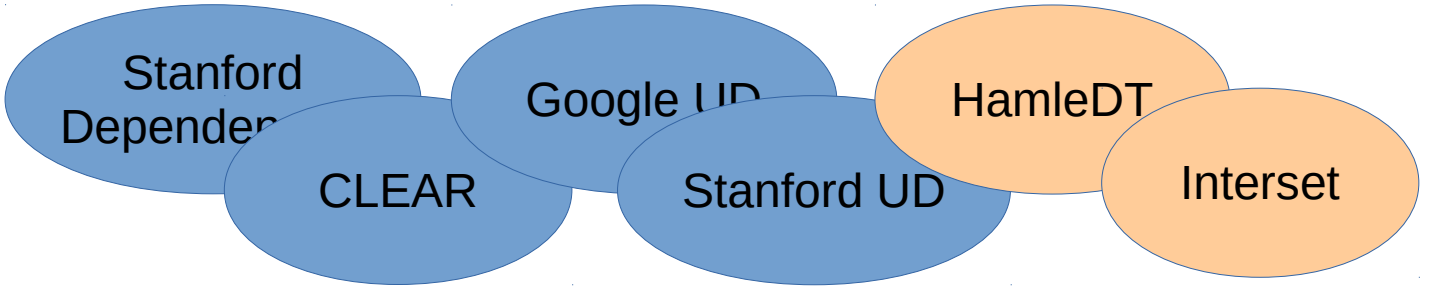
Universal Dependencies

http://universaldependencies.github.io/docs/



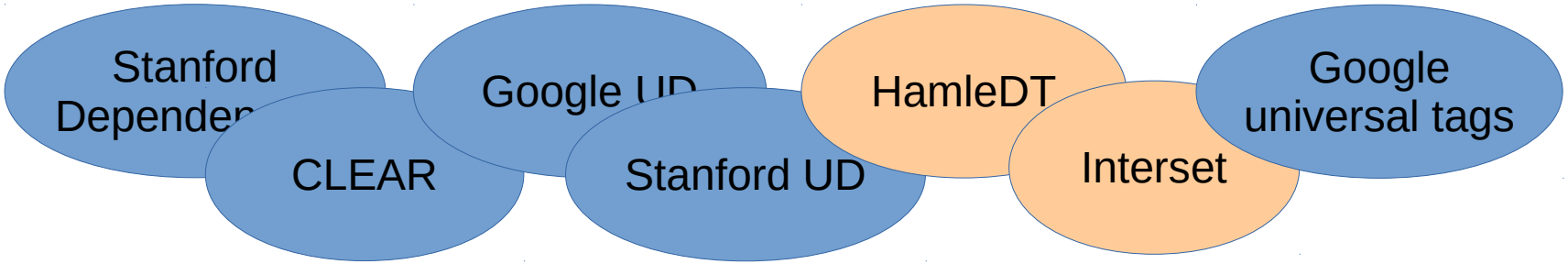
Universal Dependencies

http://universaldependencies.github.io/docs/



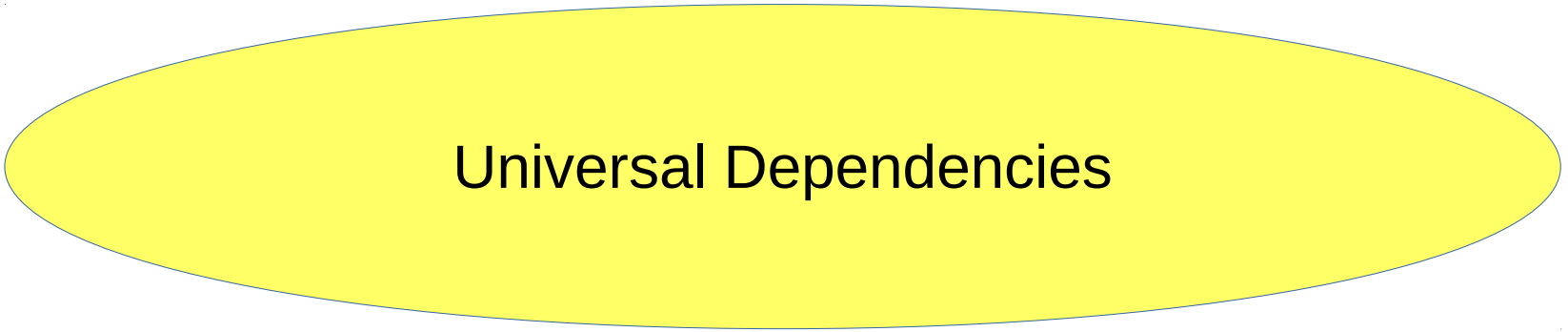
Universal Dependencies

http://universaldependencies.github.io/docs/



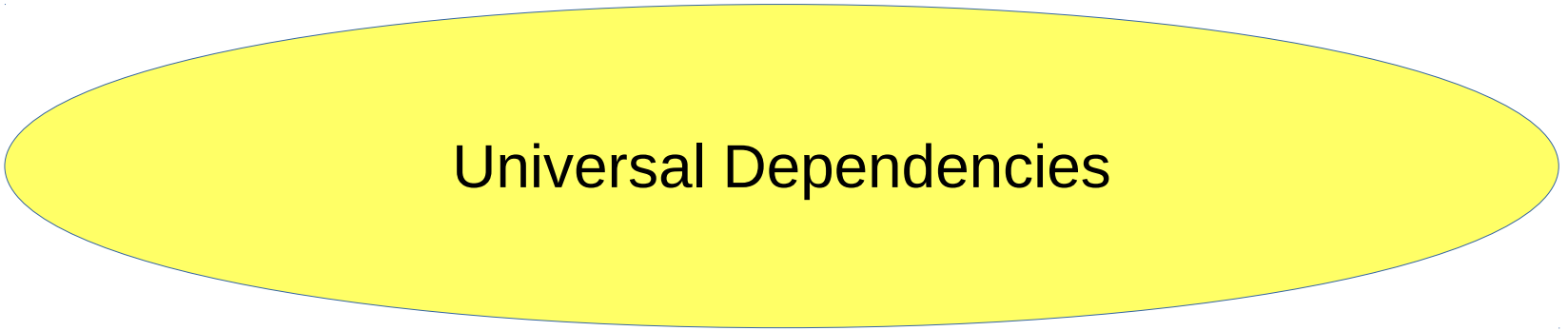
Universal Dependencies

http://universaldependencies.github.io/docs/



Universal Dependencies

http://universaldependencies.github.io/docs/



- Milestones:
  - 2014-04: EACL Göteborg, kick-off meeting
  - 2014-10: UD guidelines version 1
  - 2015-01: released treebanks of 10 languages (UD 1.0)
  - 2015-05: released treebanks of 18 languages (UD 1.1)
  - 2015-11: next release

Goals and Requirements

- Cross-linguistically consistent grammatical annotation

Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP

Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP
- Based on common usage and existing de facto standards

Goals and Requirements

- Cross-linguistically consistent grammatical annotation
- Support multilingual research and development in NLP
- Based on common usage and existing de facto standards
- Caveats:
  - Not a new linguistic theory –
    but linguistically informed and relevant
  - Not an ideal parsing representation –
    but useful for comparative evaluation
  - Not the ultimate annotation scheme –
    but a lightweight lingua franca

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
  - Don’t annotate things that are not there
  - Balance: is it still the same thing?
  - Allow language-specific extensions

Morphology

Některé	dívky	si	nicméně	pochvalovaly	zmrzlinu	.

Morphology

Některé	dívky	si	nicméně	pochvalovaly	zmrzlinu	.
některý	dívka	se	nicméně	pochvalovat	zmrzlina	.

- Lemma representing the semantic content of the word

Morphology

Některé	dívky	si	nicméně	pochvalovaly	zmrzlinu	.
některý	dívka	se	nicméně	pochvalovat	zmrzlina	.
DET	NOUN	PRON	CONJ	VERB	NOUN	PUNCT

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

Morphology

Některé	dívky	si	nicméně	pochvalovaly	zmrzlinu	.
některý	dívka	se	nicméně	pochvalovat	zmrzlina	.
DET	NOUN	PRON	CONJ	VERB	NOUN	PUNCT
PronType=Ind Gender=Fem Number=Plur Case=Nom	Gender=Fem Number=Plur Case=Nom	PronType=Prs Reflex=Yes Case=Dat		VerbForm=Part Tense=Past Voice=Act Aspect=Imp Gender=Fem Number=Plur	Gender=Fem Number=Sing Case=Acc	

- 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	CONJ	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

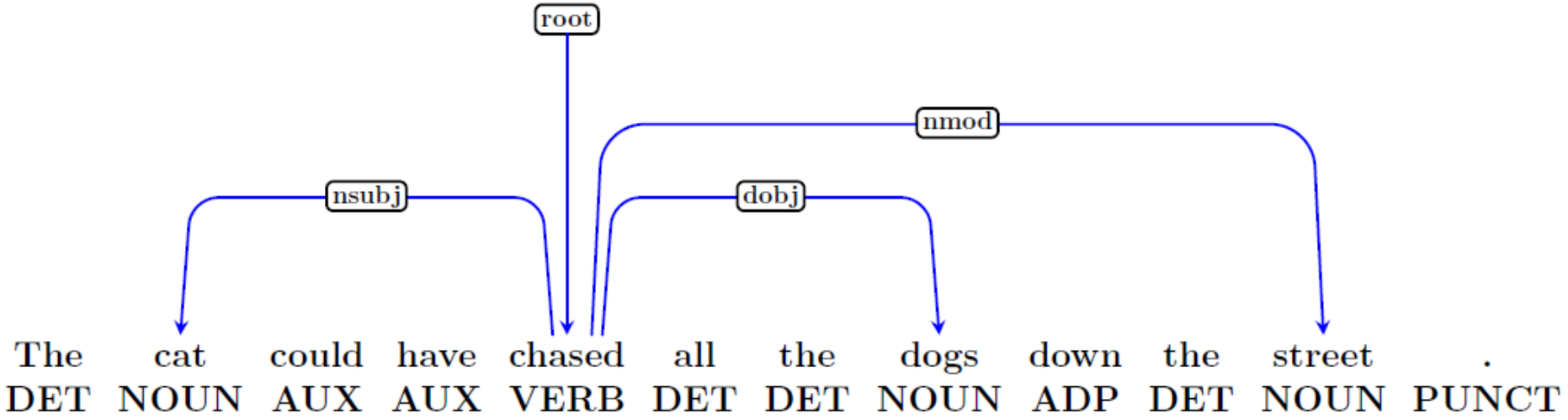
Lexical	Inflectional / Nominal	Inflectional / Verbal
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	Number	Tense
Reflex	Case	Aspect
	Definite	Voice
	Degree	Person
		Negative

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

Syntax

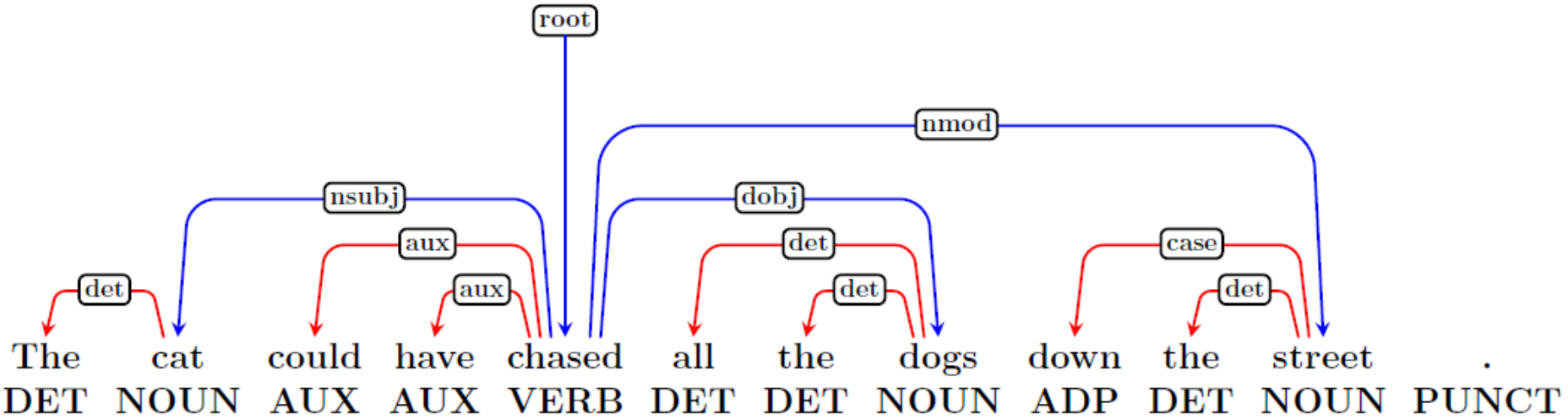
The	cat	could	have	chased	all	the	dogs	down	the	street	.
DET	NOUN	AUX	AUX	VERB	DET	DET	NOUN	ADP	DET	NOUN	PUNCT

Syntax



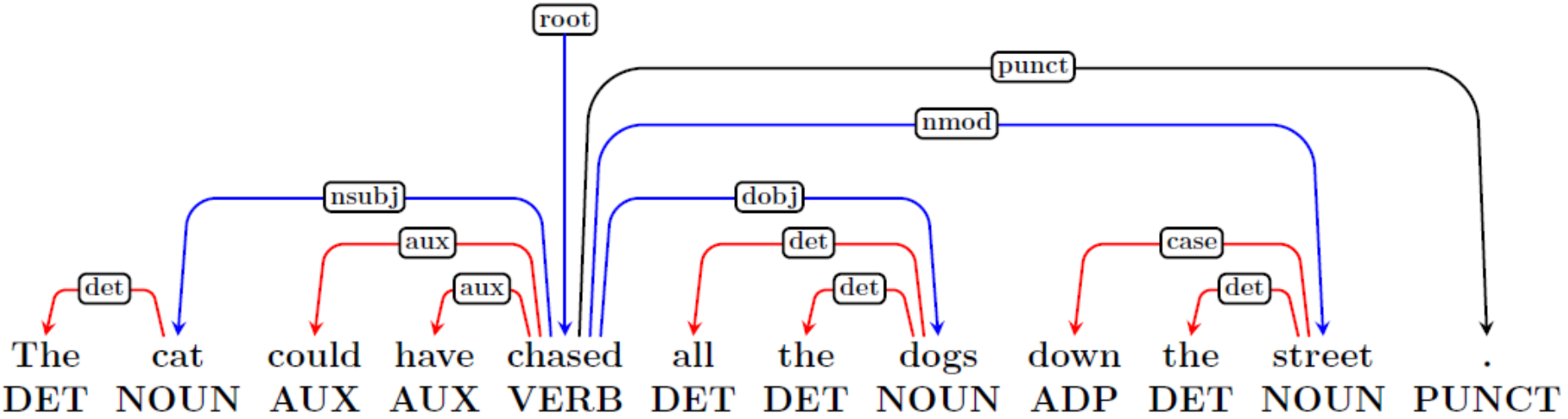
- Content words are related by dependency relations

Syntax

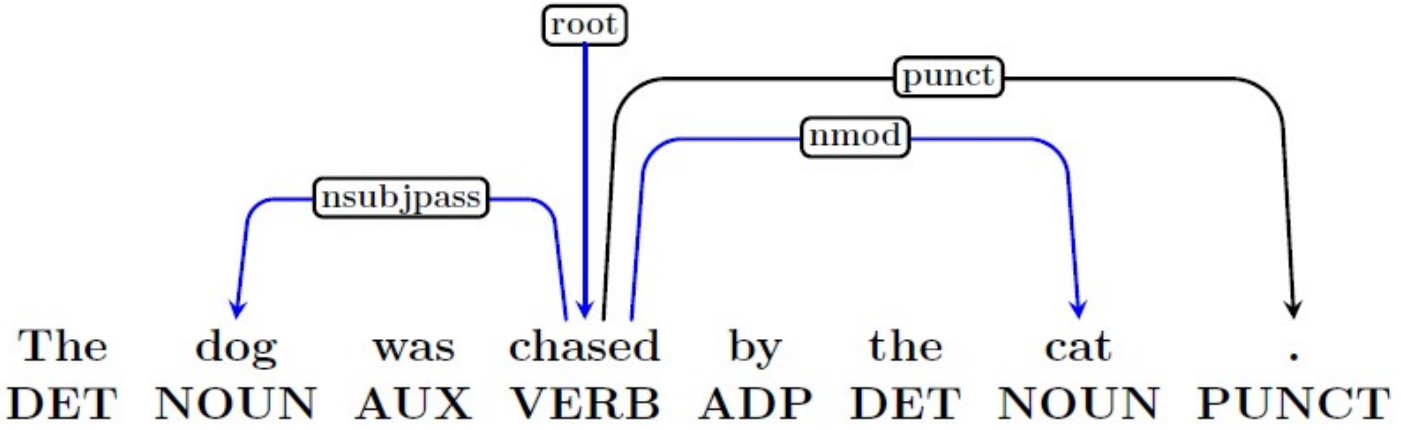


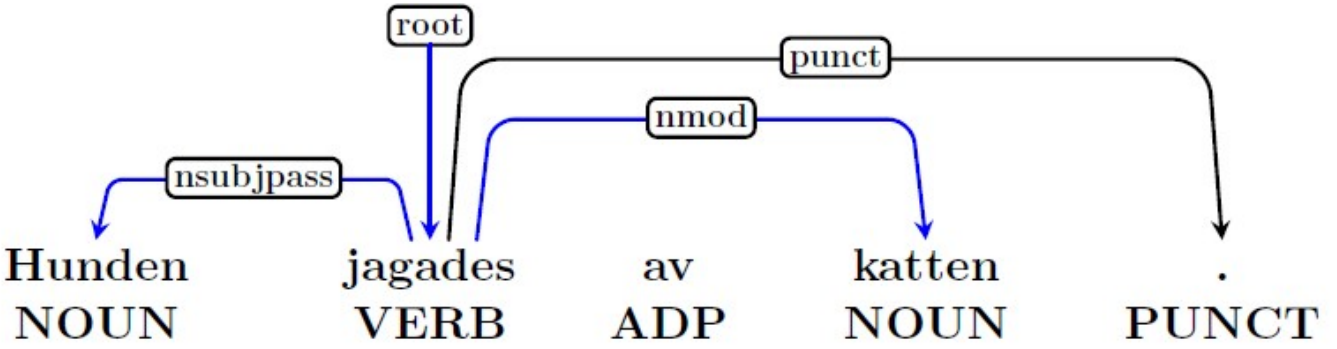
- Content words are related by dependency relations
- Function words attach to closest content word

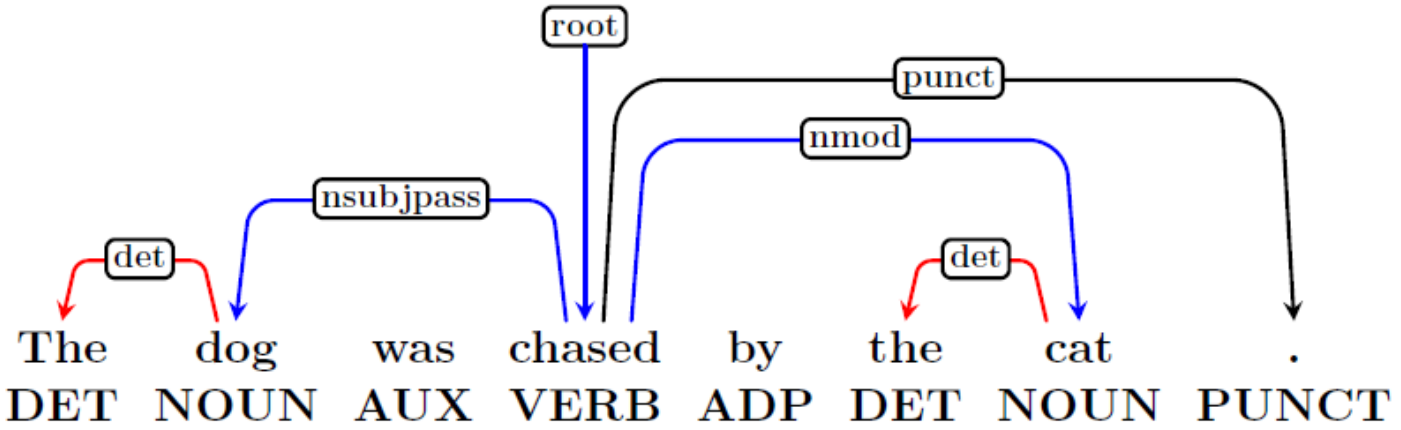
Syntax

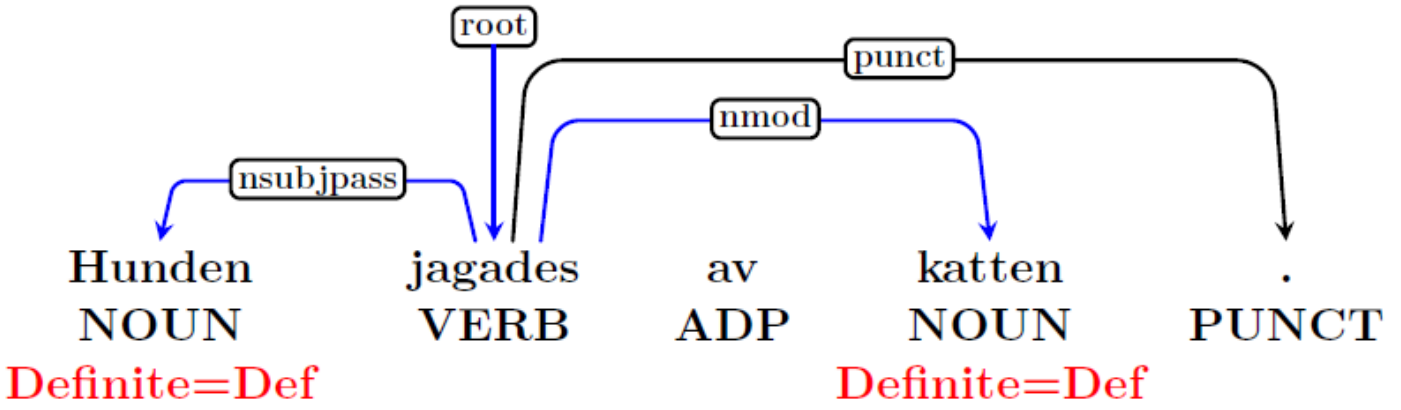


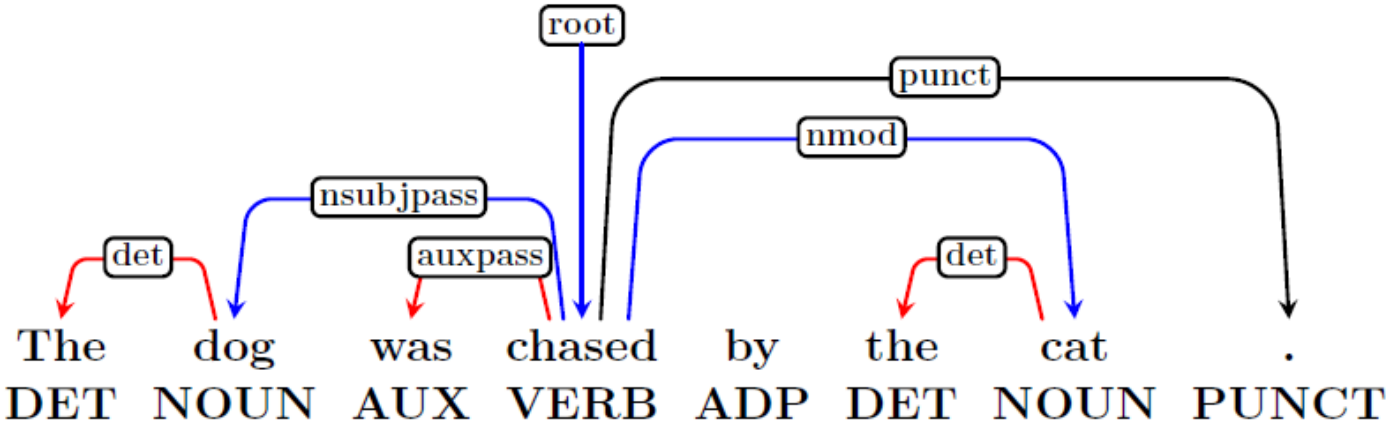
- Content words are related by dependency relations
- Function words attach to closest content word
- Punctuation attach to head of phrase or clause

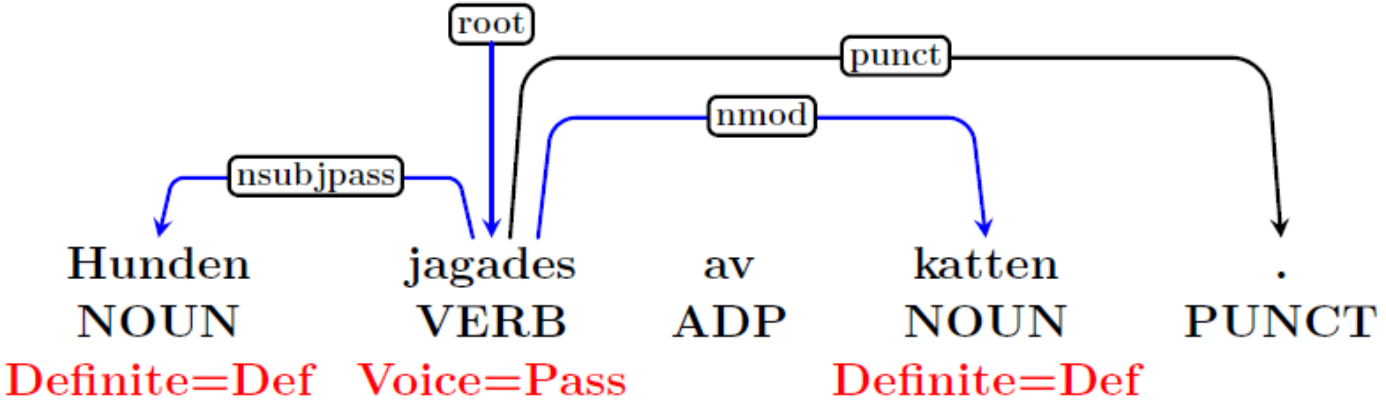

root
punct
nmod
nsubjpass
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT

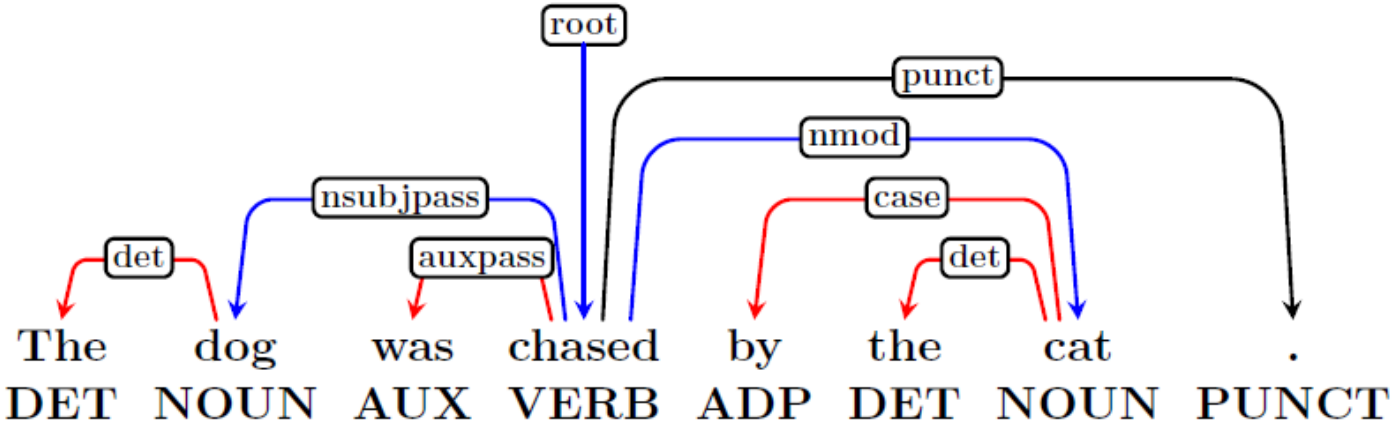

root
punct
nmod
nsubjpass
Hunden jagades av katten .
NOUN VERB ADP NOUN PUNCT

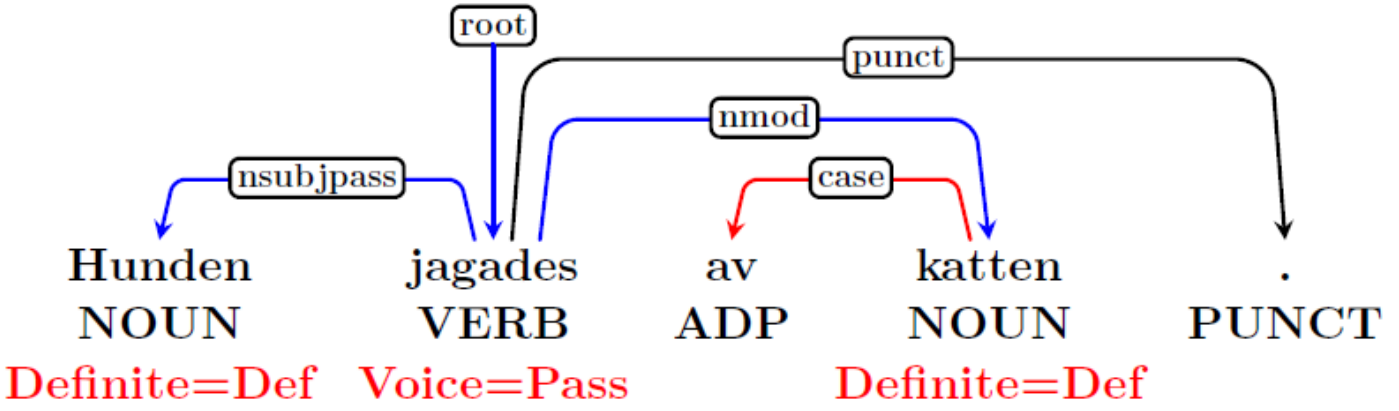

root
punct
nmod
nsubjpass
det
det
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT


root
punct
nmod
nsubjpass
Hunden jagades av katten .
NOUN VERB ADP NOUN PUNCT
Definite=Def
Definite=Def


root
punct
nmod
nsubjpass
det
auxpass
det
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT


root
punct
nmod
nsubjpass
Hunden jagades av katten .
NOUN VERB ADP NOUN PUNCT
Definite=Def Voice=Pass Definite=Def


root
punct
nmod
nsubjpass
case
det
auxpass
det
The dog was chased by the cat .
DET NOUN AUX VERB ADP DET NOUN PUNCT

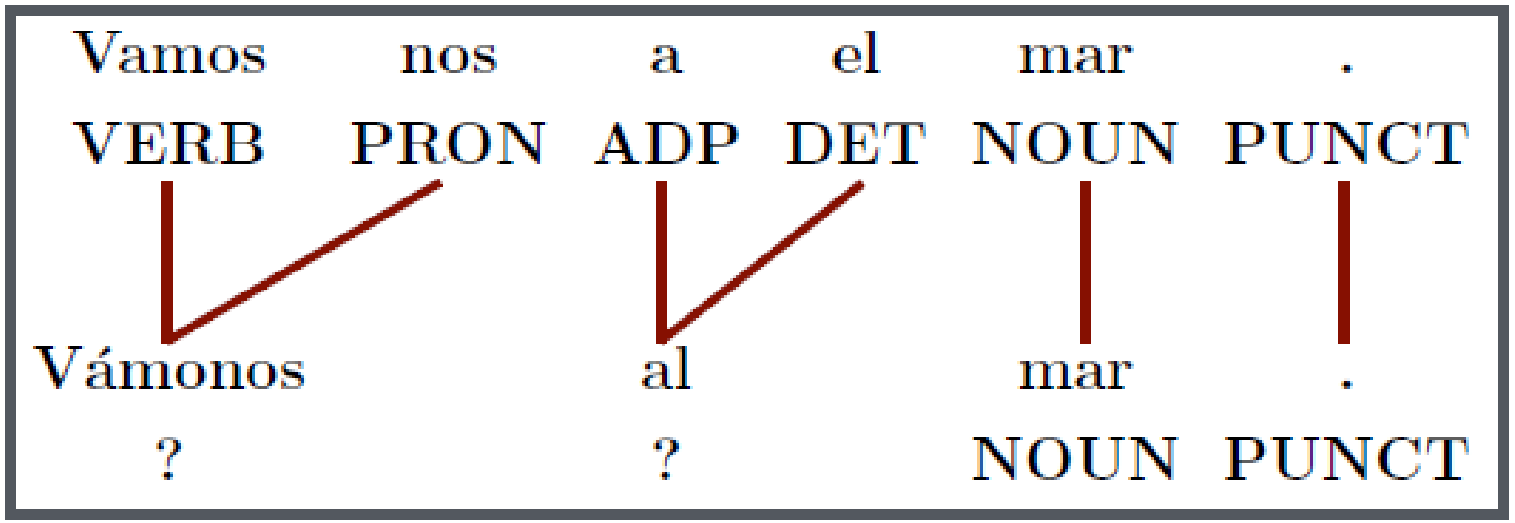

root
punct
nmod
nsubjpass
case
Hunden jagades av katten .
NOUN VERB ADP NOUN PUNCT
Definite=Def Voice=Pass Definite=Def

Dependency Relations

- Taxonomy of 40 universal grammatical relations, broadly attested in language typology (de Marneffe et al., 2014)
  - Language-specific subtypes may be added

Word Segmentation

- Fusions
  - *al = a + el*
  - *naň = na + něj*
- Clitics
  - *vámonos = vamos + nos*
  - *изменяться = изменять + ся*



Where are we now?

- Universal Dependencies, Version 1
  - Guidelines released October 2014
  - Treebank release May 2014 (v. 1.1):
    Basque, Bulgarian, Croatian, Czech, Danish, English, Finnish, French, German, Greek, Hebrew, Hungarian, Indonesian, Irish, Italian, Persian, Spanish, Swedish
- Future plans
  - New releases every six months (May, November)
  - Revision of guidelines as needed
- November 2015
  - Improve consistency of existing data
  - New languages
    Ancient Greek, Arabic, Dutch, Estonian, Latin, Polish, Portuguese, Slovenian, Tamil; Gothic? Hindi? Kazakh? Norwegian? Old Church Slavonic? Romanian? Urdu?

Ancient Greek	309K	LF					
Arabic	282K	F	-				
Basque	60K	LF				✔	
Bulgarian	125K	LF			-	✔	
Catalan	-				-	-	-
Croatian	87K	LF		-	-	✔	
Czech	1,503K	LF			✔	✔	
Danish	100K	LF		-	-	✔	
Dutch	200K	LF		-			
English	254K	LF				✔	
Estonian	9K	LF		-	-		-
Finnish	181K	LFD			✔	✔	
Finnish-FTB	161K	LF		-		✔	
French	388K				-	✔	
German	293K			-	-	✔	
Greek	59K	LF			-	✔	
Hebrew	115K	F		-	-	✔	
Hindi	-			-	-	-	-
Hungarian	26K	LF			-	✔	
Indonesian	121K			-	-	✔	
Irish	23K	L			-	✔	
Italian	258K	LF			✔	✔	
Japanese	-						
Kazakh	-				-		-
Korean	-			-	-	-	-
Latin	53K	LF		-			
Latin-ITT	259K	LF		-			
Norwegian	-			-			-
Persian	151K	F			-	✔	
Polish	83K	LF		-			
Portuguese	212K	LF		-			
Romanian	-			-	-		-
Slovenian	140K	LF					
Spanish	424K	LF			✔	✔	
Swedish	96K	LF			✔	✔	
Tamil	9K	LF		-			
Turkish	-				-	-	-

Existing Slavic Treebanks

Language	Code	Treebank	Sent	Tok
Bulgarian	[bg]	BulTreeBank	13,221	196K
Church Slavonic	[cu]	PROIEL	7,818	72K
Croatian	[hr]	SETimes.HR	3,736	84K
Czech	[cs]	PDT	87,913	1504K
Polish	[pl]	IPI PAN	8,227	84K
Russian	[ru]	SynTagRus	63,000	900K
Slovak	[sk]	SNK	63,238	994K
Slovene	[sl]	Ssj200k	11,411	236K

Issues of Slavic Languages in UD

- Pronouns vs. determiners, numerals and quantifiers
- Attachment of cardinal numbers
- Verbs, participles, adjectives
- Auxiliary verbs and modal verbs
- Direct object, indirect object and nmod
- Reflexive pronouns, reflexive passive
- Comparative constructions

Pronouns and Determiners

- English + Romance languages: DET = article or pronominal adjective *(this, which, every)*

Pronouns and Determiners

- English + Romance languages: DET = article or pronominal adjective *(this, which, every)*
- We don’t have this category! (Traditionally → PRON.)

Pronouns and Determiners

- English + Romance languages: DET = article or pronominal adjective *(this, which, every)*
- We don’t have this category! (Traditionally → PRON.)
- But we have the words (except for articles).

Pronouns and Determiners

- English + Romance languages: DET = article or pronominal adjective *(this, which, every)*
- We don’t have this category! (Traditionally → PRON.)
- But we have the words (except for articles).
- Currently functional borderline (but ellipsis?)
  This.DET *car is expensive.*
  This.PRON *is expensive.*

Pronouns Only

- Personal pronouns (including reflexives, but not possessives)
- Interrogative *who, what*
- Indefinite and negative derivatives
- Relative [cs] *jenž*
  - cs: *já, ty, on, my, vy, oni, se, kdo, co, někdo, něco, nikdo, nic*
  - sk: *já, ty, on, my, vy, oni, sa, kto, čo, niekto, niečo, nikto, nič*
  - pl: *ja, ty, on, my, wy, oni, się, kto, co, ktoś, coś, nikt, nic*
  - ru: *я, мы, он, мы, вы, они, ся, кто, что, кто-нибудь, что-нибудь, никто, ничто*
  - sl: *jaz, ti, on, mi, vi, oni, se, kdo, kaj, nekdo, nekaj, nihče, nič*
  - hr: *ja, ti, on, mi, vi, oni, se, tko, što, neki, nešto, nitko, ništa*
  - bg: *аз, ти, ние, вие, се, кой, кое, някой, нещо, никой, нищо*
  - cu: *азъ, тъı, мъı, въı, и, сѧ, къто, чьто*

Possessives: Determiners

- If they occur without a noun … ellipsis

Můj otec je starší. Tvůj má ale více zkušeností.

- cs: *můj, tvůj, jeho, její, náš, váš, jejich, svůj*
- sk: *môj, tvoj, jeho, jej, náš, váš, ich, svoj*
- …

Both Possible

- Demonstratives
  - *ten, to, tento, tamten, jaký, který, čí, nějaký, některý, něčí, každý, všechen, žádný, …*
- Adjectival interrogatives/relatives, indefinites, negatives
  - *jaký, který, čí, nějaký, některý, něčí, každý, žádný*
  - *všechen, všichni, všechno*
- Relative pronouns cannot be explained by ellipsis!
  - *Muž, kterého *muže jsem vám představil.*
  - *The man, which *man I introduced to you.*

Numerals and Quantifiers

- NUM: *jeden, dva, tři, čtyři, pět, šest, …, sto*

  (one, two, three, four, five, six, …, hundred)

Numerals and Quantifiers

- NUM: *jeden, dva, tři, čtyři, pět, šest, …, sto*
- NUM/NOUN: *tisíc, milión, miliarda*
- NOUN: *polovina, třetina, čtvrtina, setina*

(thousand, million, billion)
(half, one-third, one-fourth, one-hundredth)

Numerals and Quantifiers

- NUM: *jeden, dva, tři, čtyři, pět, šest, …, sto*
- NUM/NOUN: *tisíc, milión, miliarda*
- NOUN: *polovina, třetina, čtvrtina, setina*
- NUM: *dvé, tré, čtvero, patero, šestero; jedny, dvoje, troje, čtvery, patery*

  (one set of, two sets of, …)

Numerals and Quantifiers

- ADJ: *první, druhý, třetí, čtvrtý, …, stý, tisící; dvojí, trojí, čtverý, paterý*

  (first, second, third, fourth, …, hundredth…)

Numerals and Quantifiers

- ADJ: *první, druhý, třetí, čtvrtý, …, stý, tisící; dvojí, trojí, čtverý, paterý*
- ADV: *jedenkrát, dvakrát, třikrát, čtyřikrát; poprvé, podruhé, potřetí, posté; kolikrát, pokolikáté*

  (once, twice, three times, four times)
  (for the first time, for the second time…)
  (how many times)

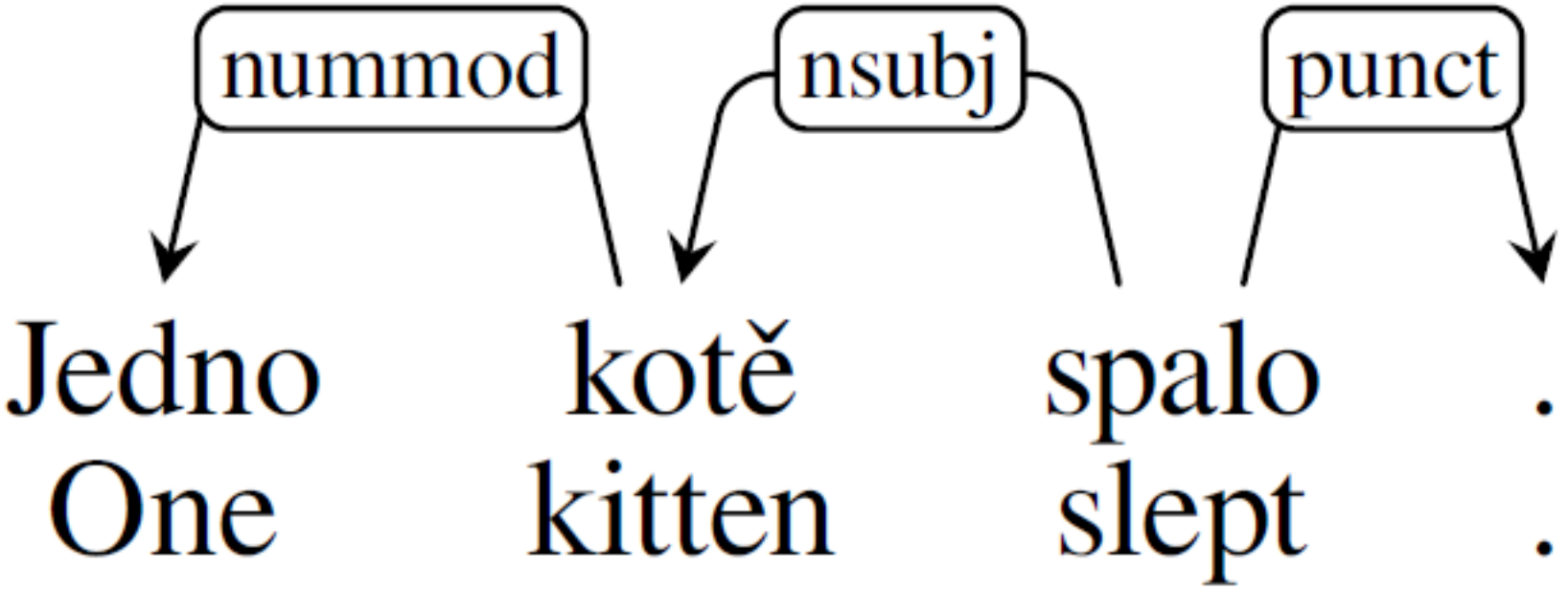
Numerals and Quantifiers

- ADJ: *první, druhý, třetí, čtvrtý, …, stý, tisící; dvojí, trojí, čtverý, paterý*
- ADV: *jedenkrát, dvakrát, třikrát, čtyři[illegible]; poprvé, podruhé, potřetí, posté; kolikrát, pokolikáté*
- DET: *kolik, tolik, několik, mnoho, málo, kolikátý, kolikerý*

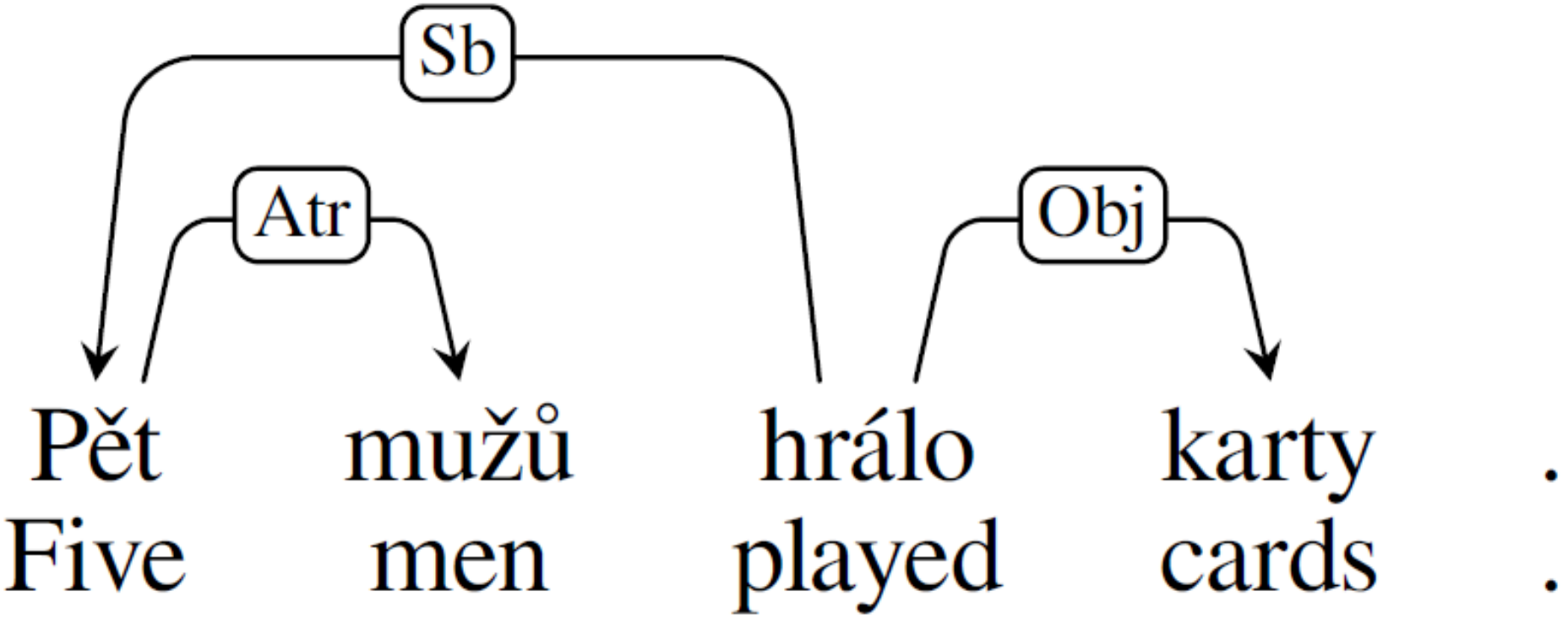
but: *více, méně, …*

(how many, so many, some, few, …)

Quantified Noun Phrases



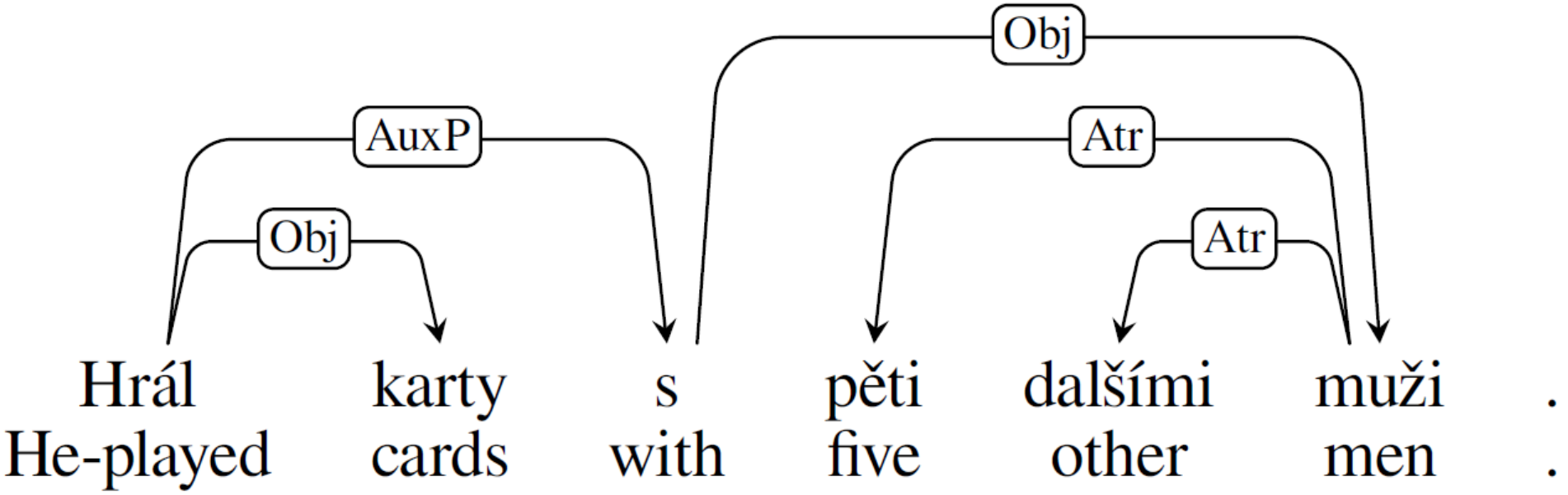
“Five” in PDT



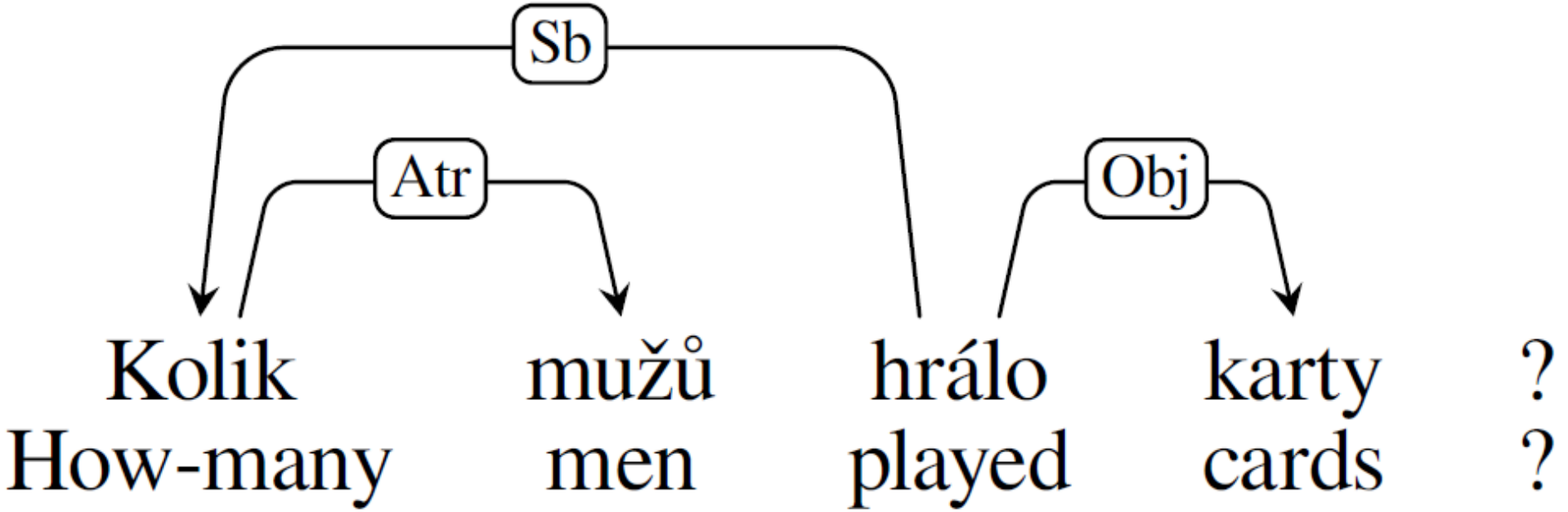
Cases: Numeral and Noun

Phrase Case	Example	Numeral Case	Noun Case
★ Nom	pět mužů	Nom	Gen
☆ Gen	pěti mužů	Gen	Gen
Dat	pěti mužům	Dat	Dat
★ Acc	pět mužů	Acc	Gen
★ Voc	pět mužů	Voc	Gen
Loc	pěti mužích	Loc	Loc
Ins	pěti muži	Ins	Ins

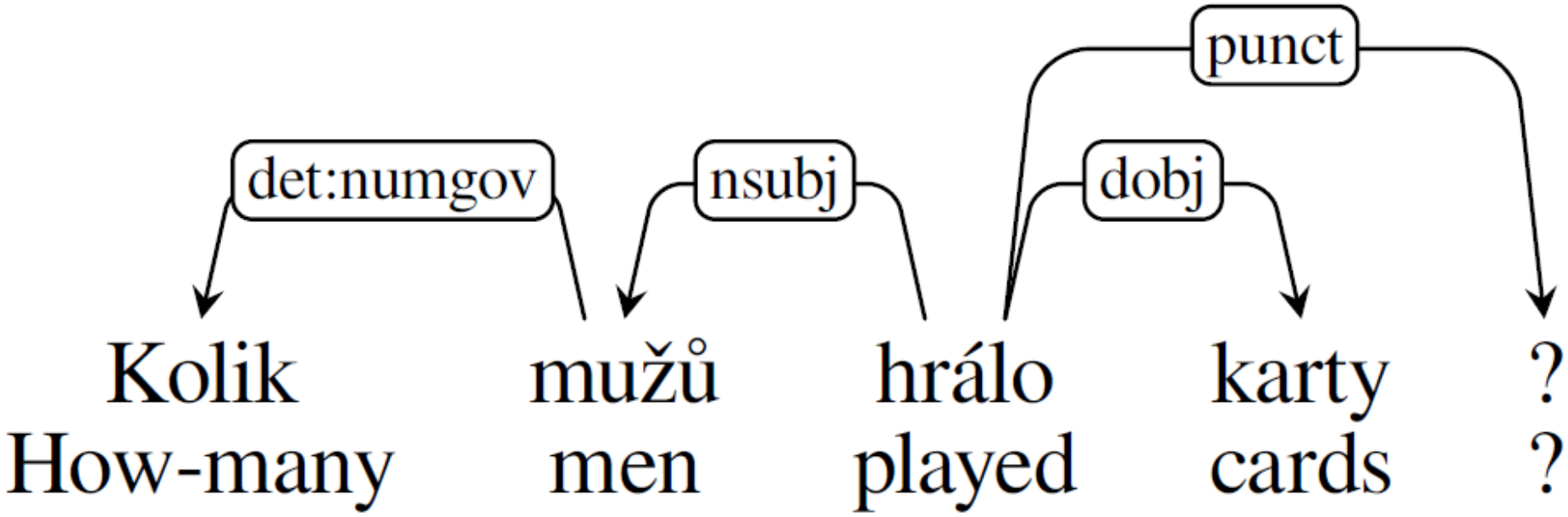
“Five” in Instrumental in PDT



Pronominal Quantifiers in PDT



Pronominal Quantifiers in UD



Language-Specific Labels

	Numeric	**Pronominal**
Noun governs	nummod	det:nummod
Numeral governs	nummod:gov	det:numgov

dobj / iobj

- Not as easy as accusative vs. dative.
- Default: dobj
- Heuristics for iobj
  - *Cením si vaší pomoci.* (Gen)
  - *Čelíme velkým problémům.* (Dat)
  - *Nedisponuje takovým rozpočtem.* (Ins)
  - *Učí mou dceru fyziku.* (2 × Acc)

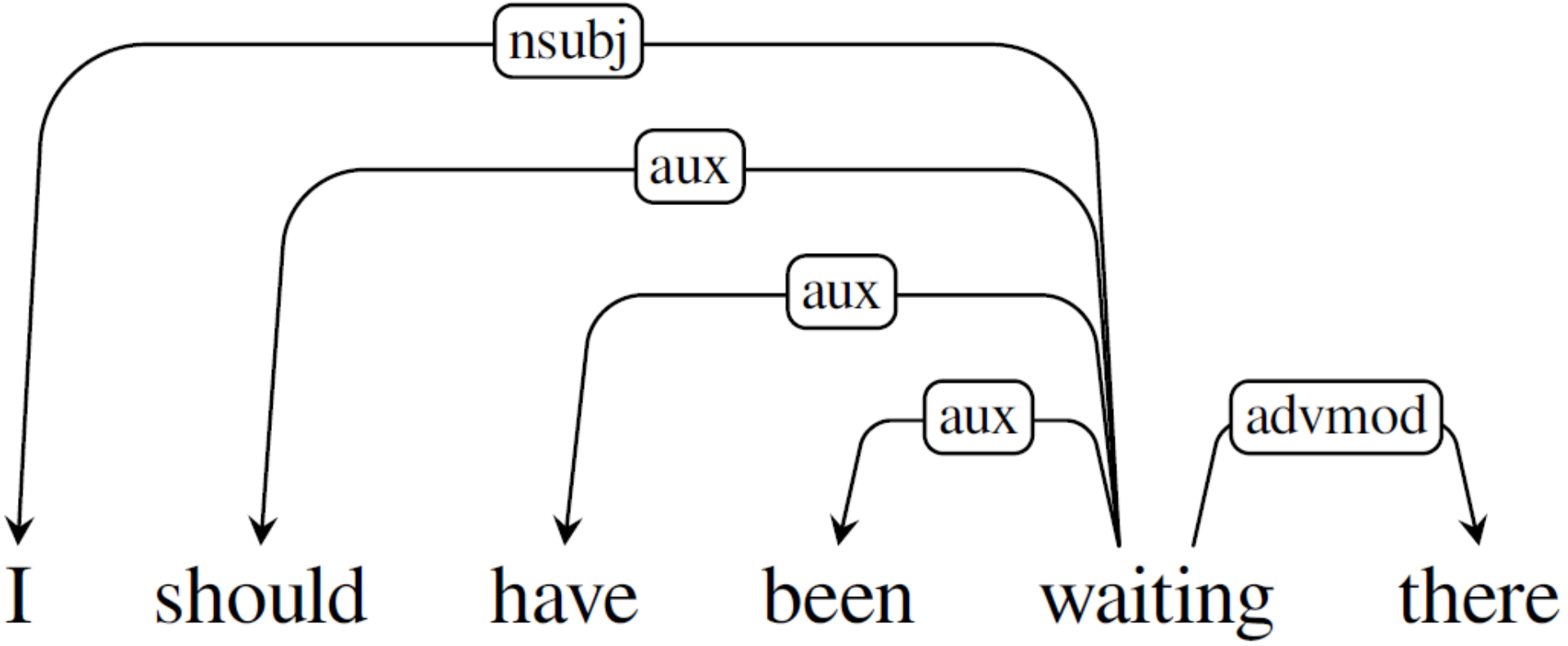
dobj / iobj / nmod

- **Core arguments:** what exactly is it?
- English:
  - *He gave John the book.* (iobj)
  - *He gave the book to John.* (nmod)
- Spanish:
  - *Dio el libro a John.* (iobj)
- Czech:
  - Every Obj is translated to dobj, regardless the case and the presence of preposition

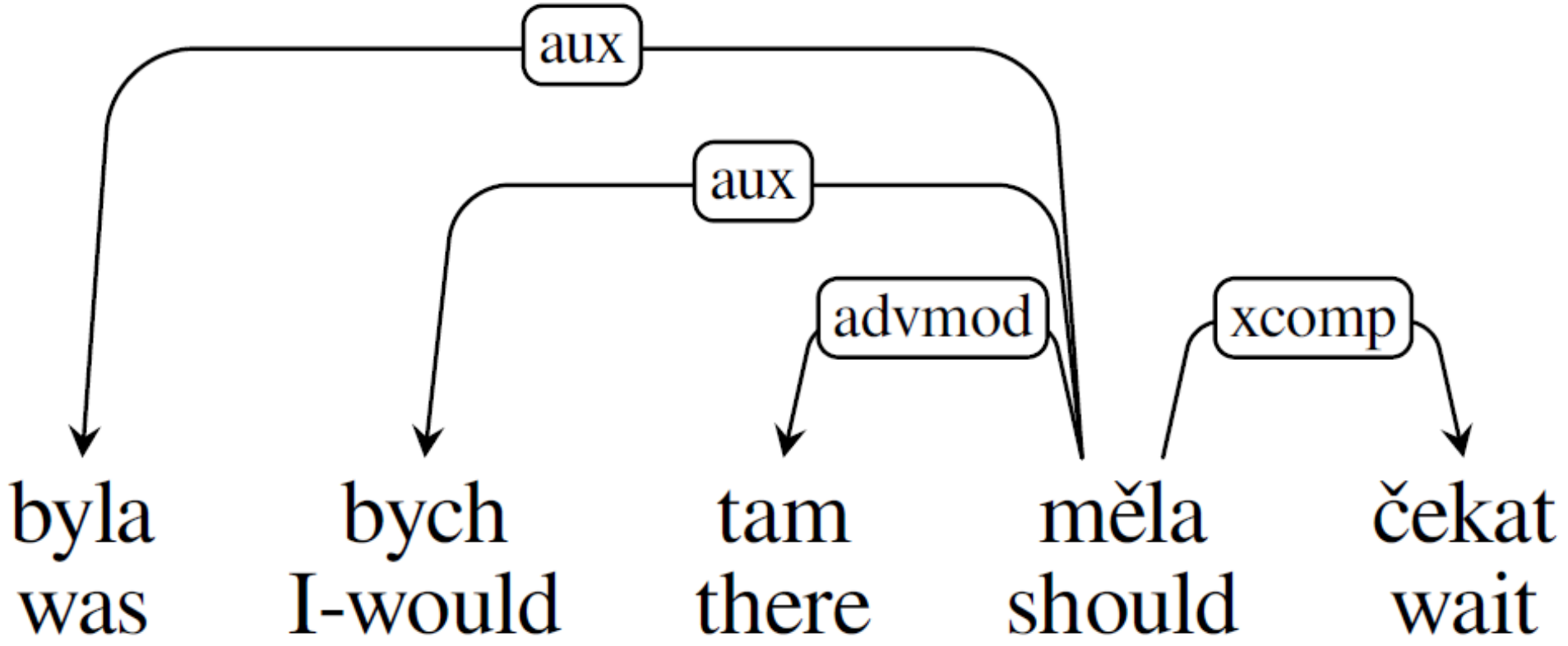
Reflexive Pronouns

- Direct or indirect object (dobj, iobj):
  Řízl se do prstu / Řízl ho do prstu.
  - Including reciprocal usage:
    Políbili se. / They kissed each other.
- Inherently reflexive verbs: *smát se, bát se*
  - `compound:reflex` (analogy to English compound:prt in *give up, come on, …*) NOW CHANGED:
    `expl`
- Reflexive passive:
  To se snadněji řekne než udělá.
  - `auxpass:reflex`

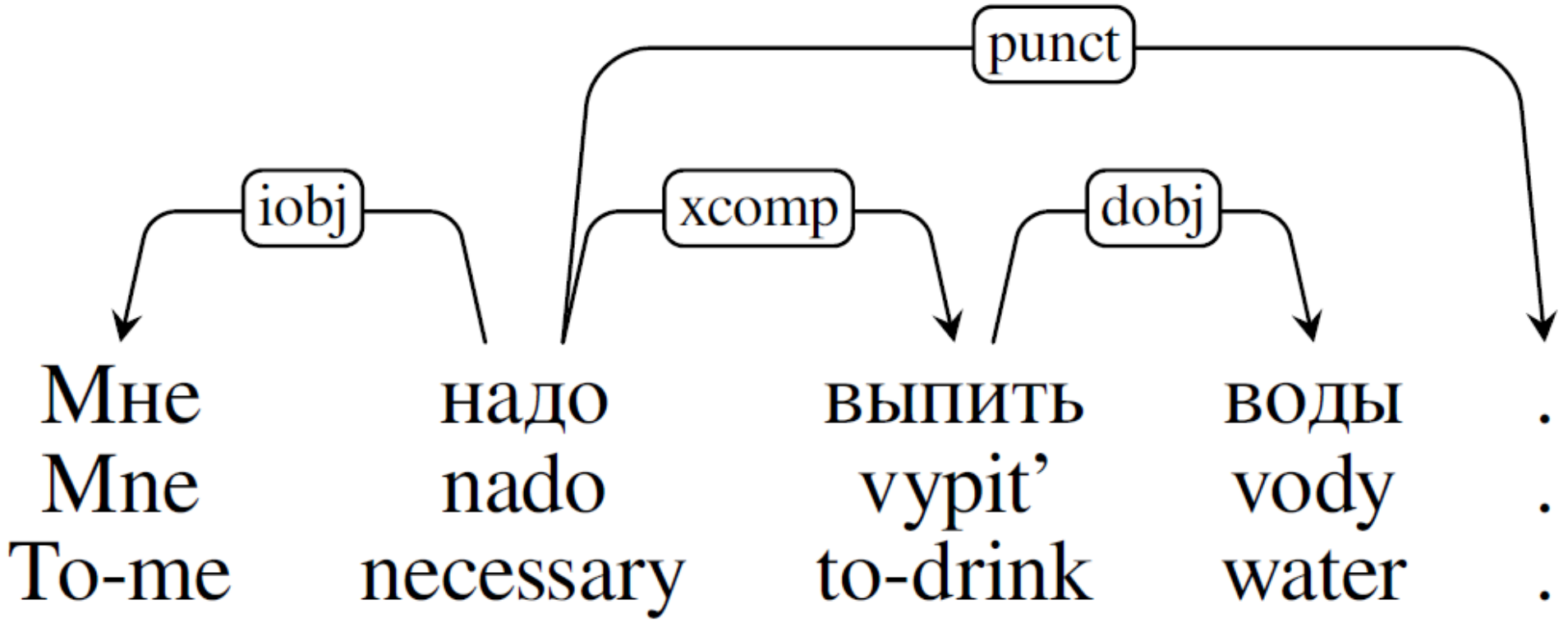
Modal Auxiliary in English



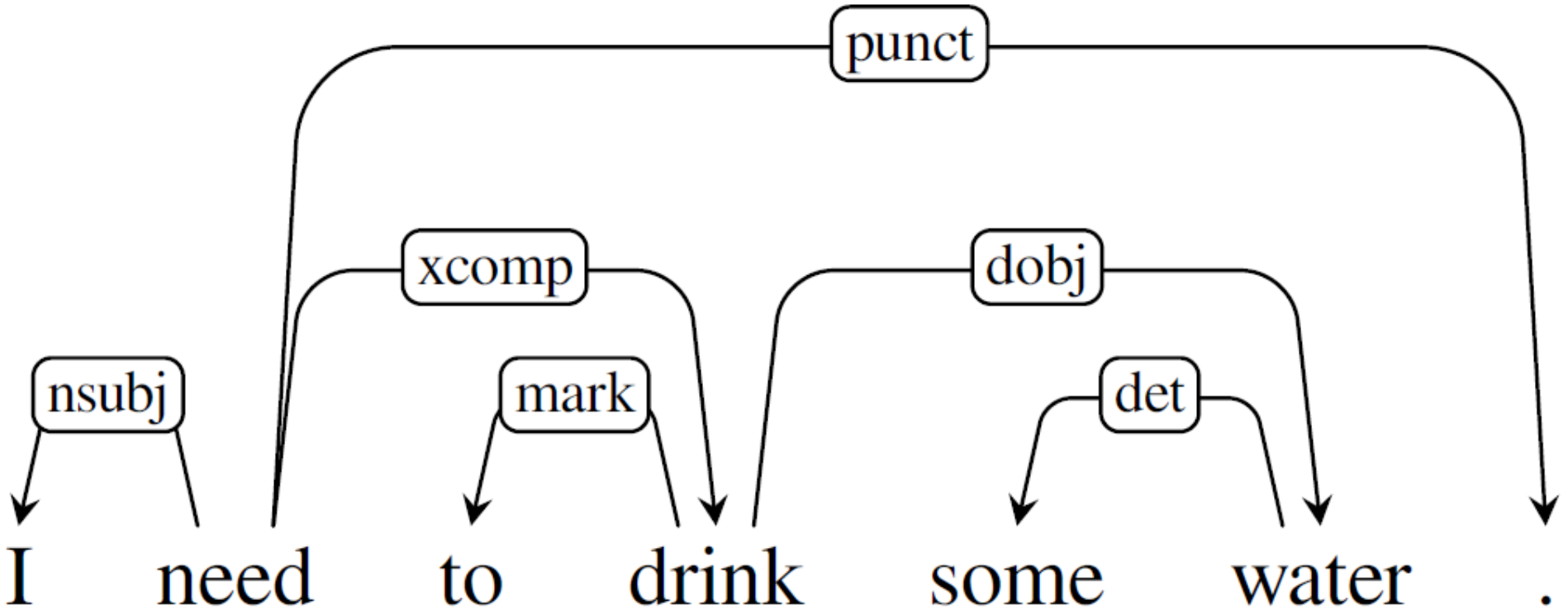
Modal Verb in Czech



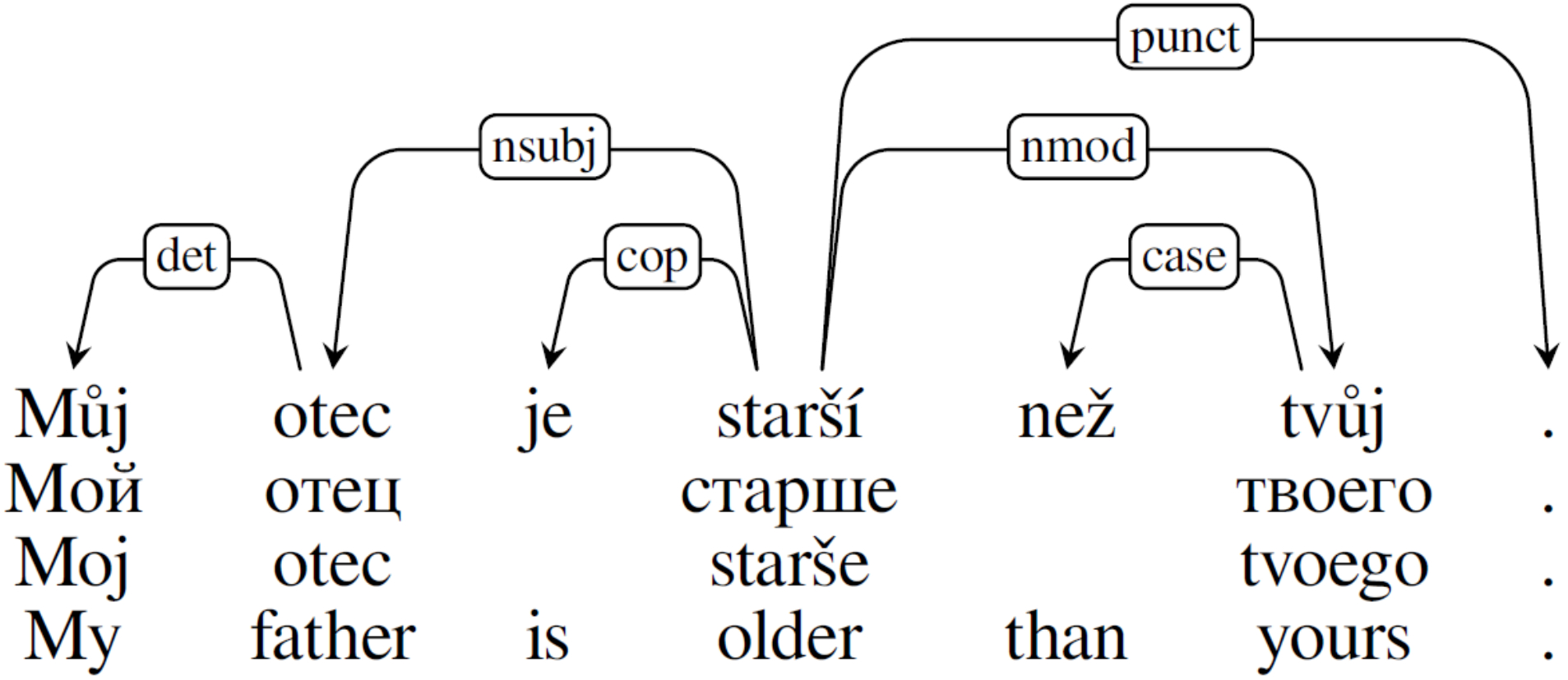
Modal Adverb in Russian



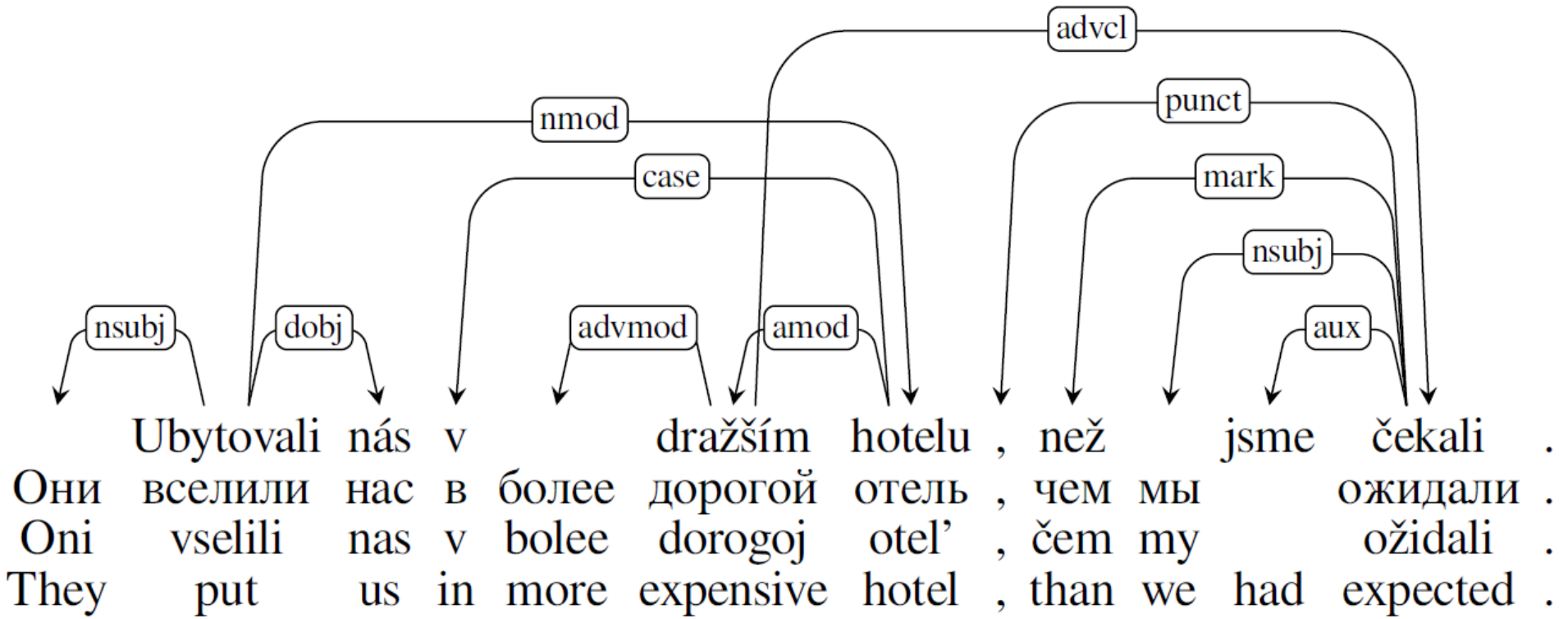
Modal / Control Verb in English



Comparative Constructions



Comparative Constructions



Summary

- Universal Dependencies =
  cross-linguistically applicable annotation of

  POS + morpho features + dependency relations

- Parallelism among Slavic languages
  - Determiners, Quantifiers, d/i Objects, Reflexives, Modal verbs/auxiliaries, Comparatives

Děkuji!
Otázky?

Благодарѭ!
Въпроси?

Ďakujem!
Otázky?

Благодаря!
Въпроси?

Thank you!
Questions?

Спасибо!
Вопросы?

Dziękuję!
Pytania?

Hvala!
Vprašanja?

Hvala!
Pitanja?