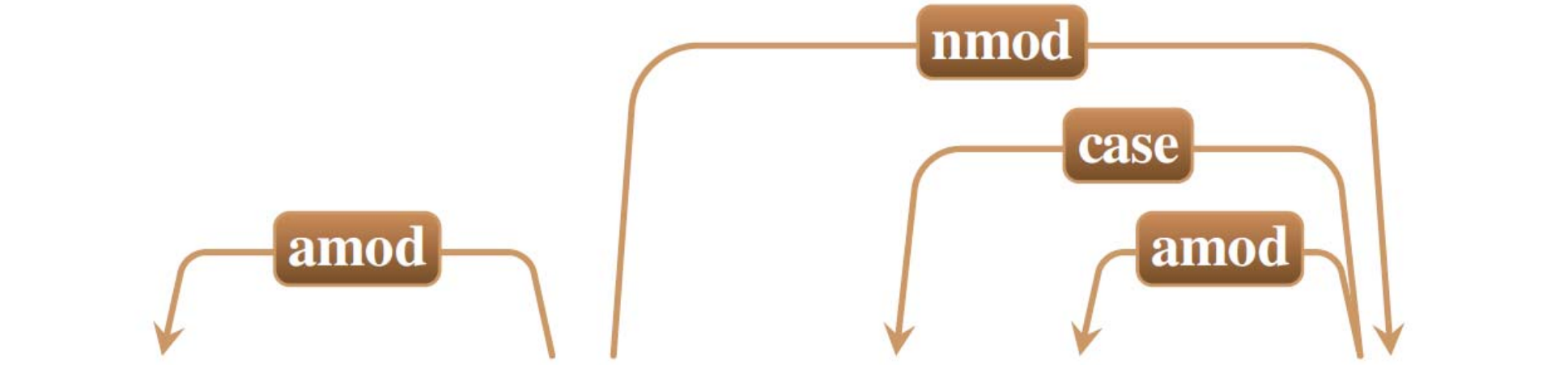
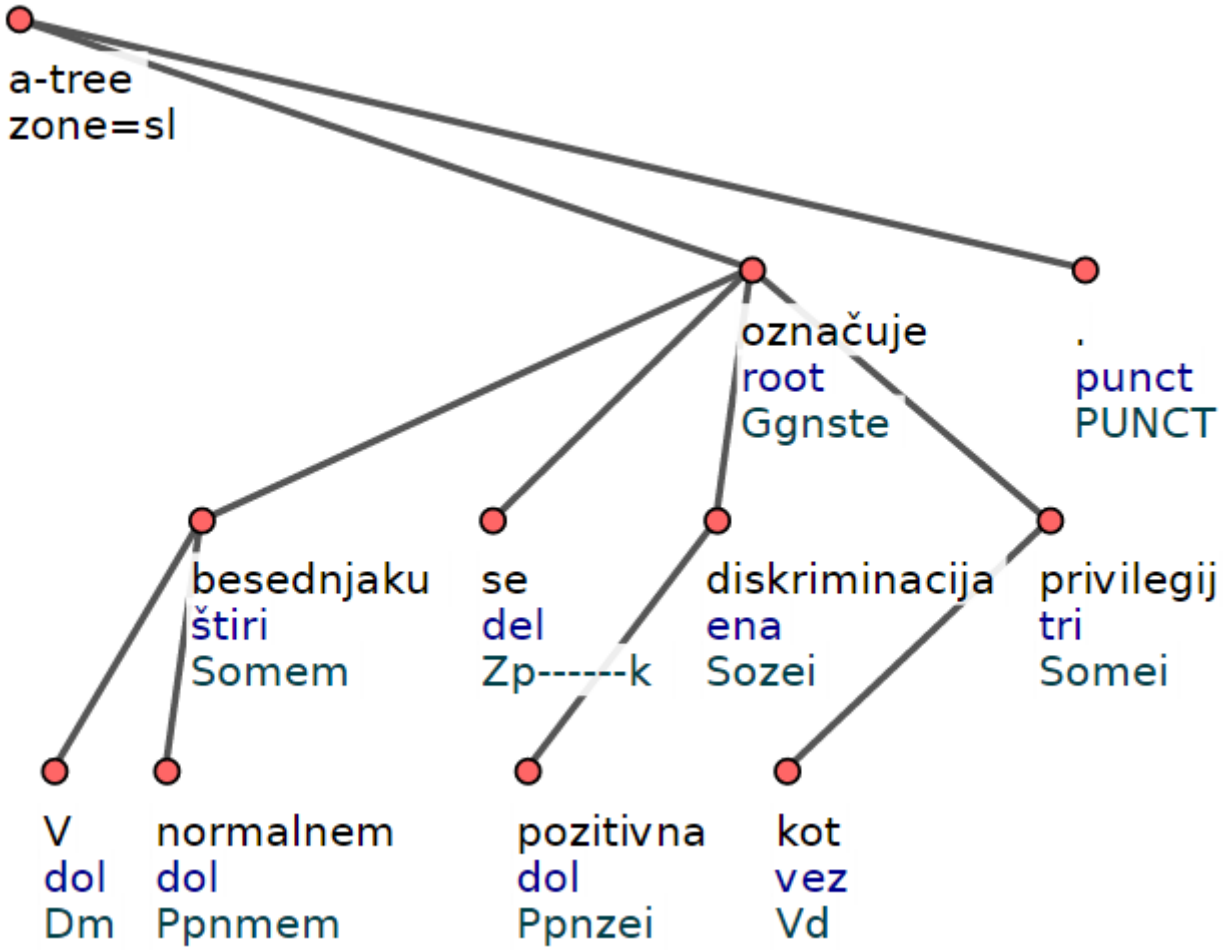
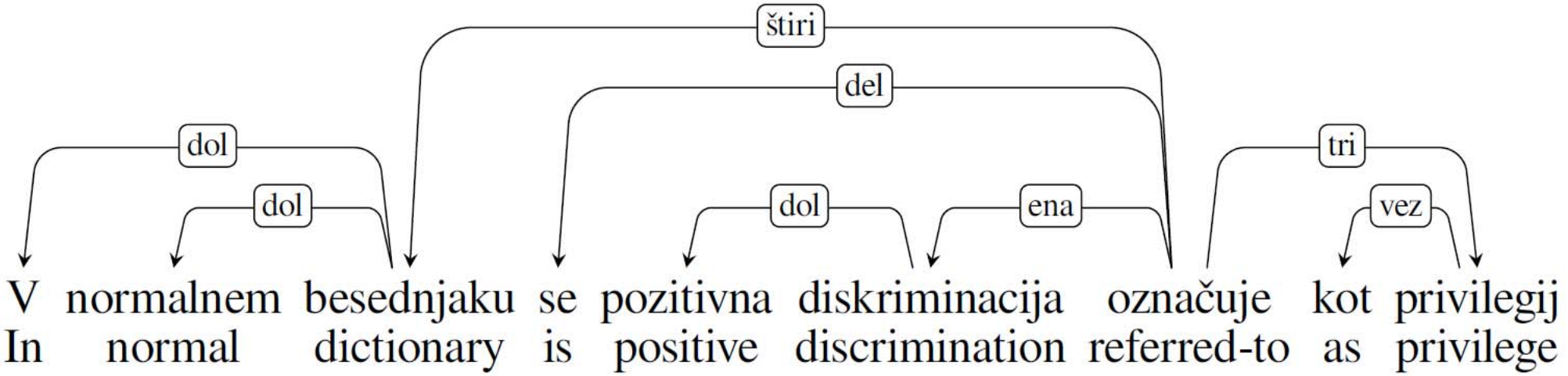


Dan Zeman
Charles University, Prague

Dependency Treebanks

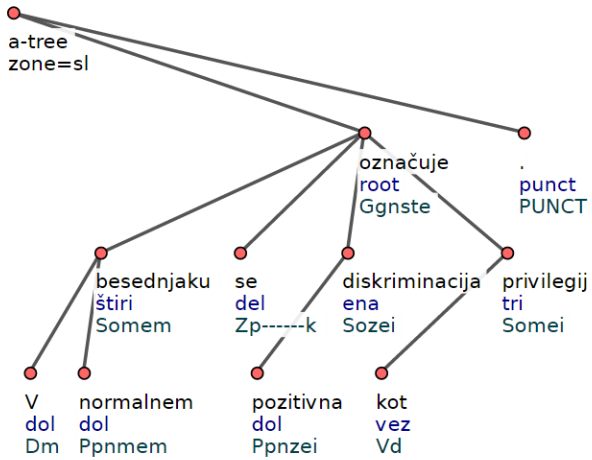


Dependency Treebanks

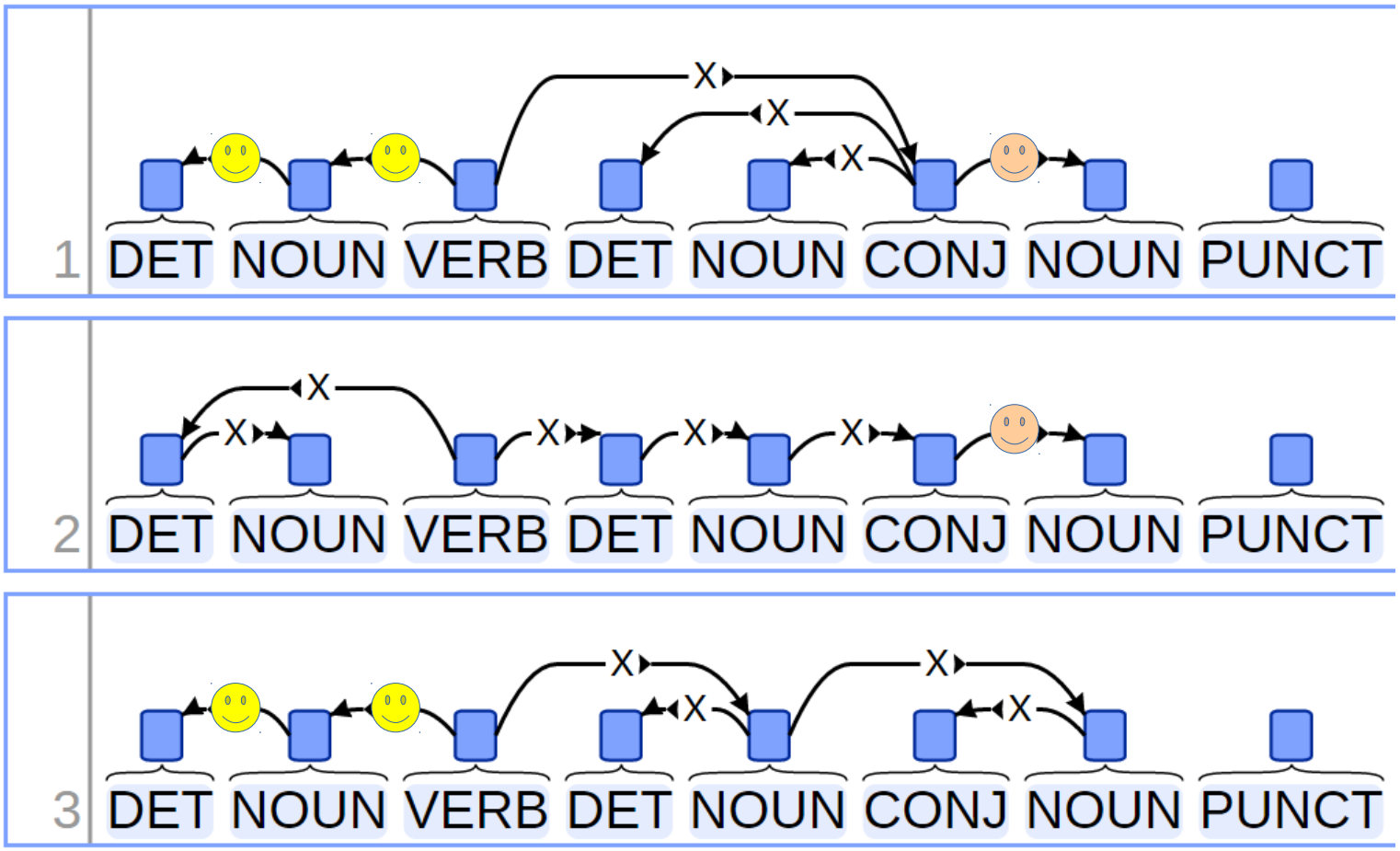


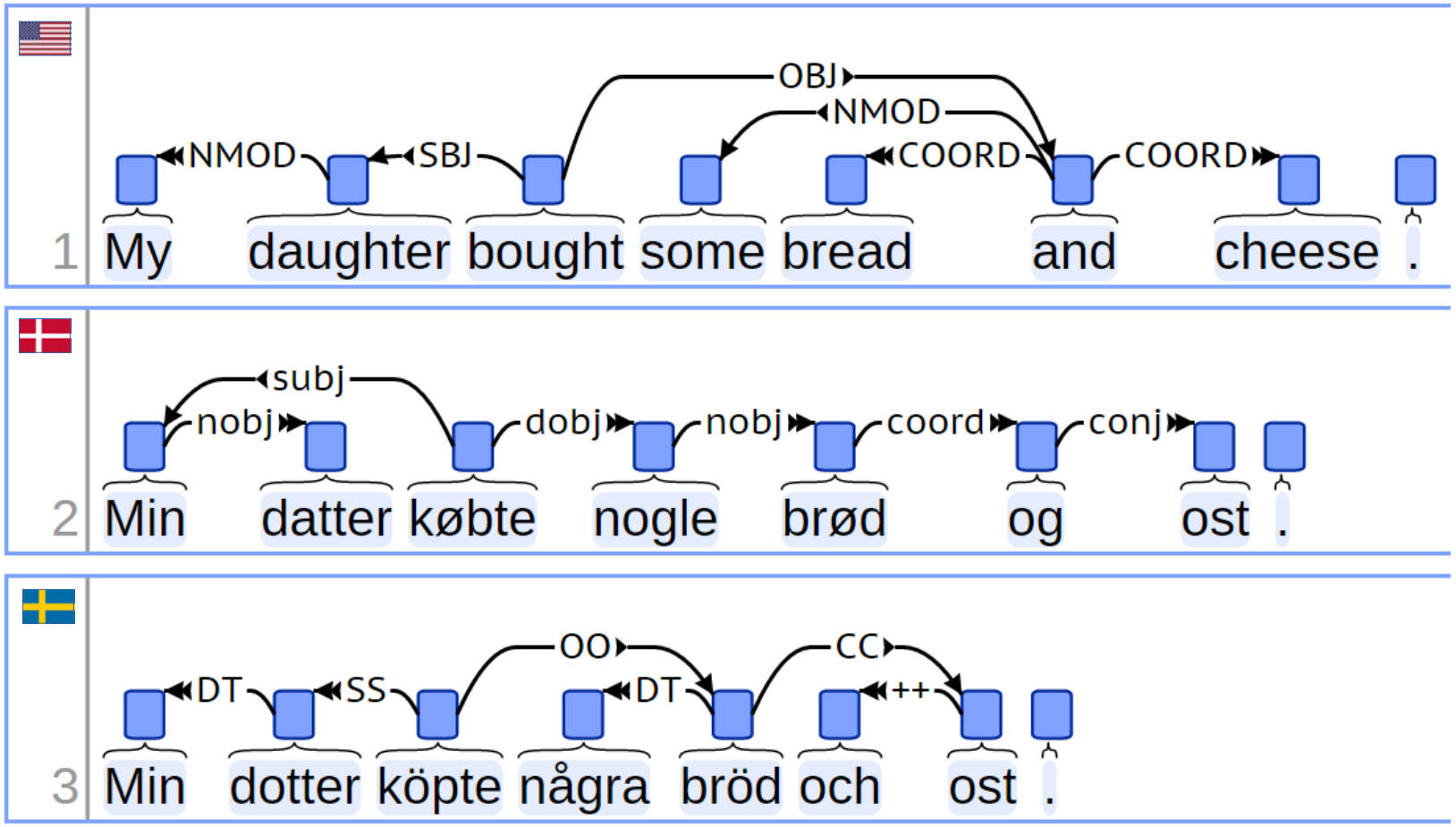
Why?

- Linguistic research
  - Corpus query



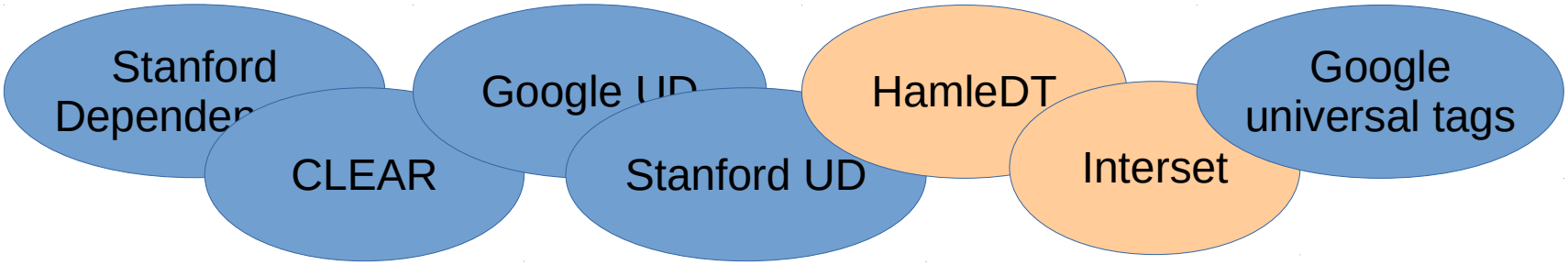
- Training tools (parsers) for NLP
  - Downstream applications


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


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

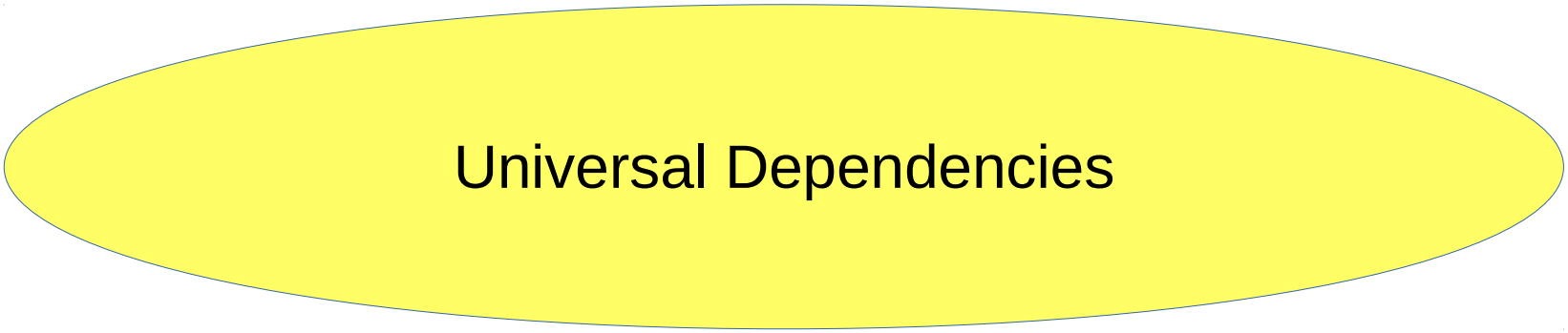
Universal Dependencies

http://universaldependencies.org/



Universal Dependencies

http://universaldependencies.org/



Universal Dependencies

http://universaldependencies.org/

- Milestones:
  - 2014-04: EACL Göteborg, kick-off meeting
  - 2014-10: UD guidelines version 1
  - 2015-01: released 10 treebanks of **10** languages (UD 1.0)
  - 2015-05: released 19 treebanks of **18** languages (UD 1.1)
  - 2015-11: released 37 treebanks of **33** languages (UD 1.2)
  - 2016-05: released 54 treebanks of **40** languages (UD 1.3)
  - 2016-11: UD release 1.4, ~7 new languages
  - 2016 fall: UD guidelines version 2

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

Not
"Universal"
in the strictly
typological
sense!

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

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

Part-of-Speech Tags

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

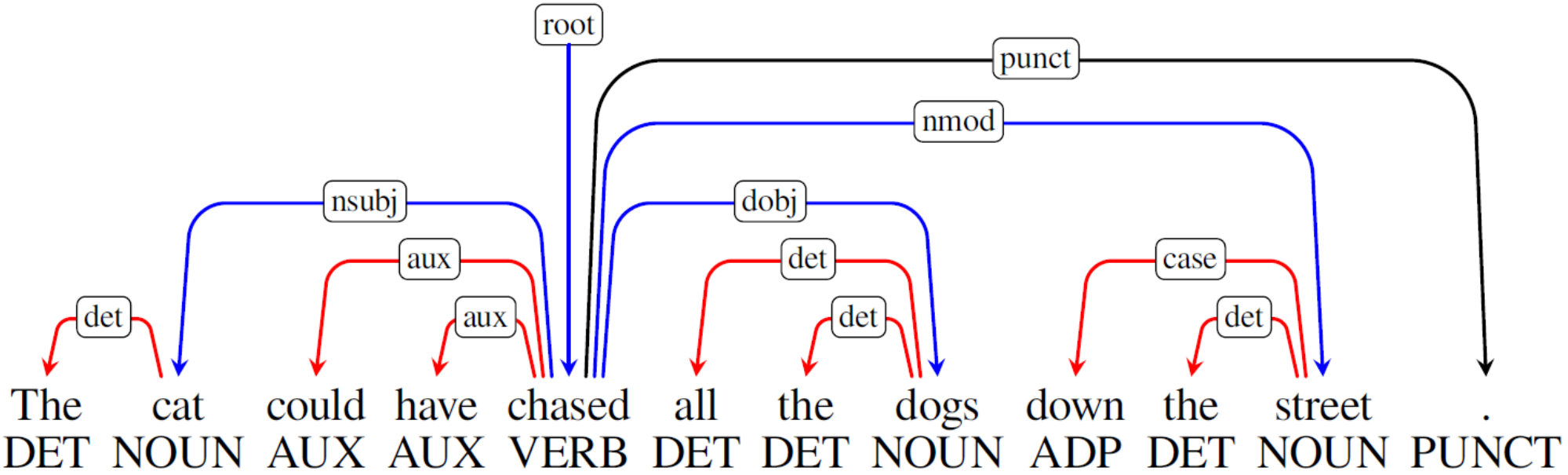
- Taxonomy of 17 universal part-of-speech tags, based on the Google Universal Tagset (Petrov et al., 2012)
- All languages use the same inventory, but not all tags have to be used by all languages

Features

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

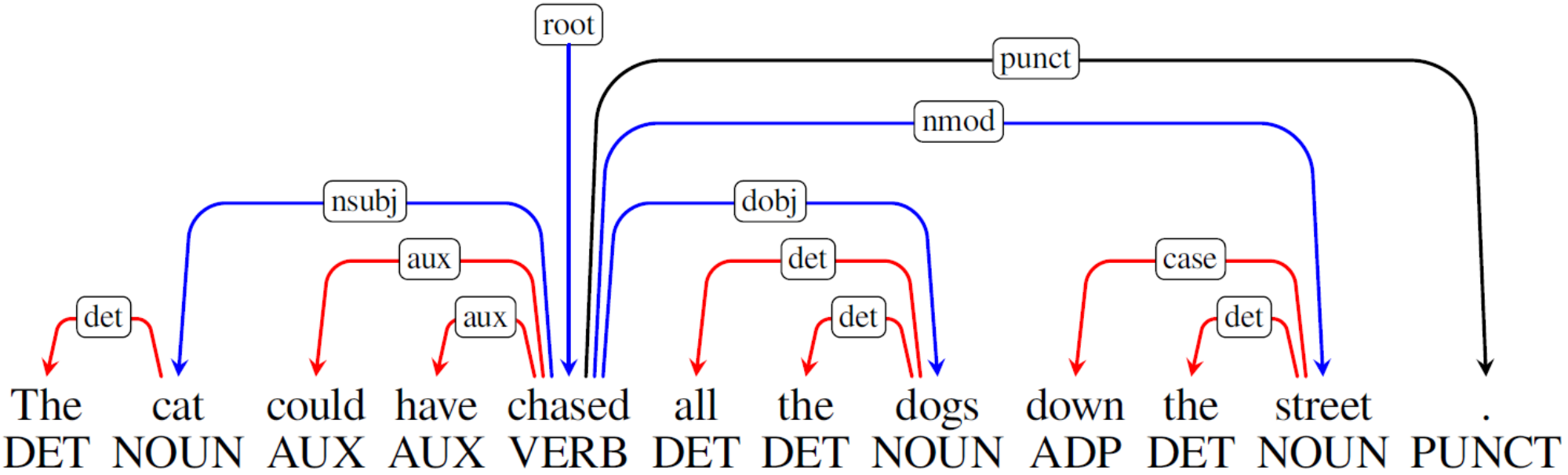
- Standardized inventory of morphological features, based on Interset (Zeman, 2008)
- Languages select relevant features and can add language-specific features or values with documentation

Syntax



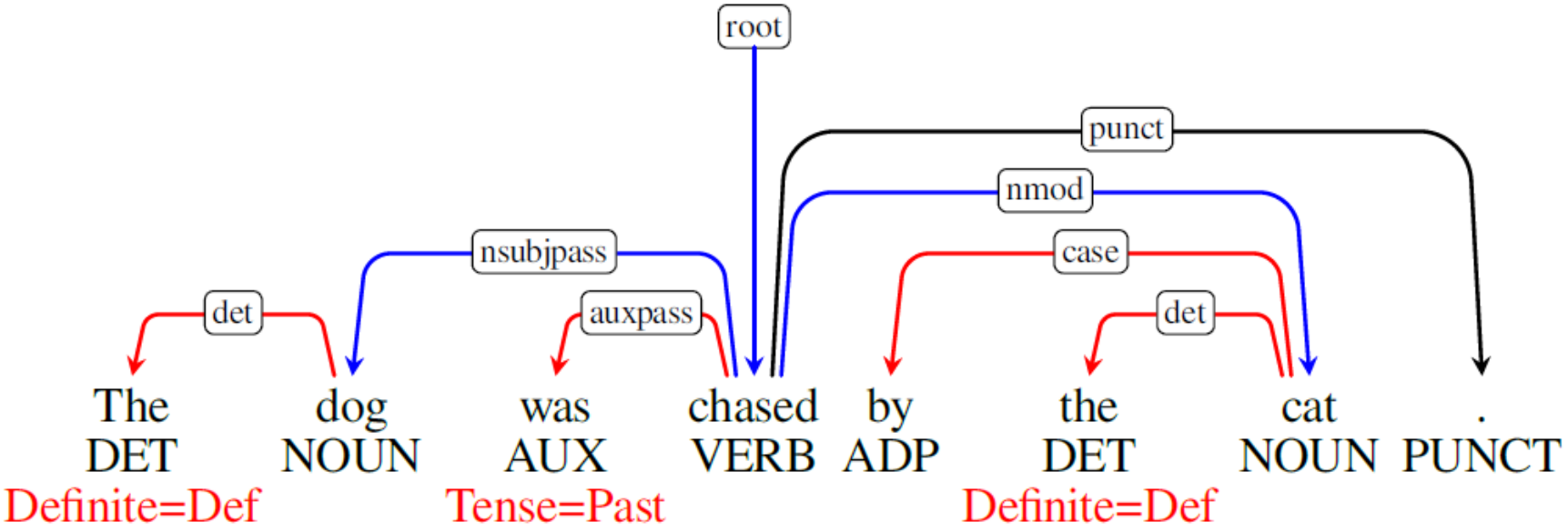
- Content words are related by dependency relations
- Function words attach to closest content word
- Punctuation attach to head of phrase or clause

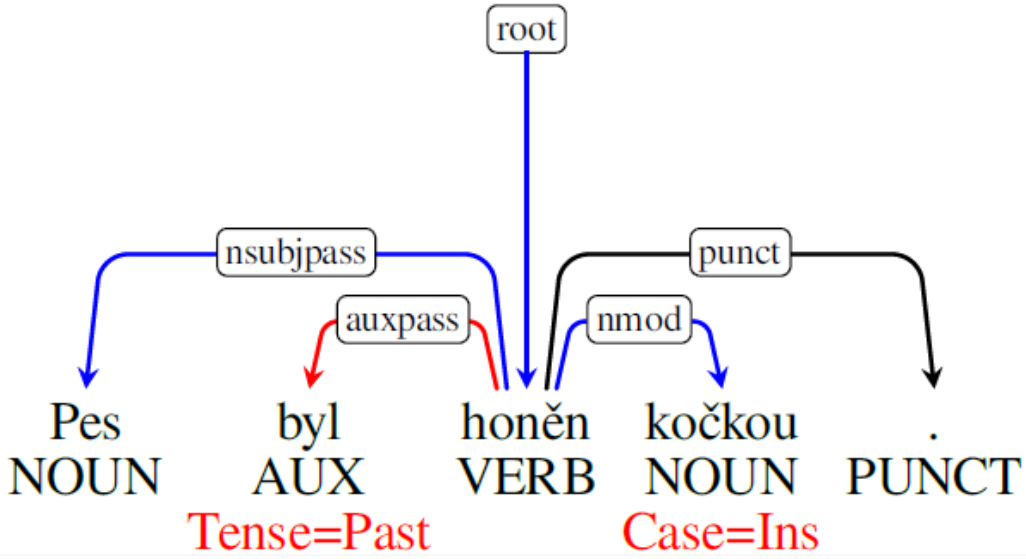
Syntax



- Content words are related by dependen lations
- Function words attach to closest cont
- Punctuation attach to head of phrase

Not "dependency" in the strictly syntactic sense!


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

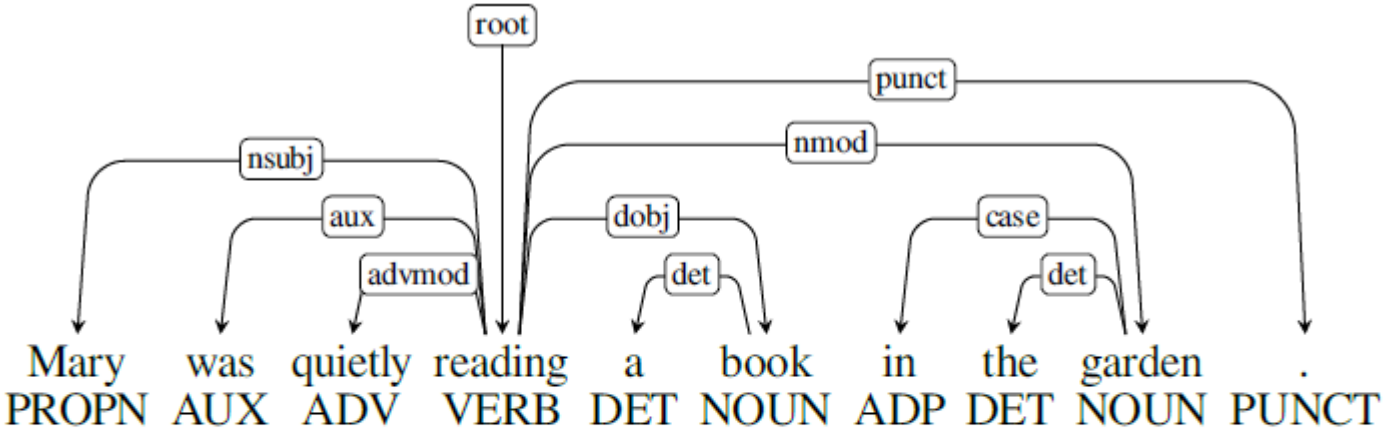

root
nsubjpass
punct
auxpass
nmod
Pes NOUN
byl AUX Tense=Past
honěn VERB
kočkou NOUN Case=Ins
. PUNCT

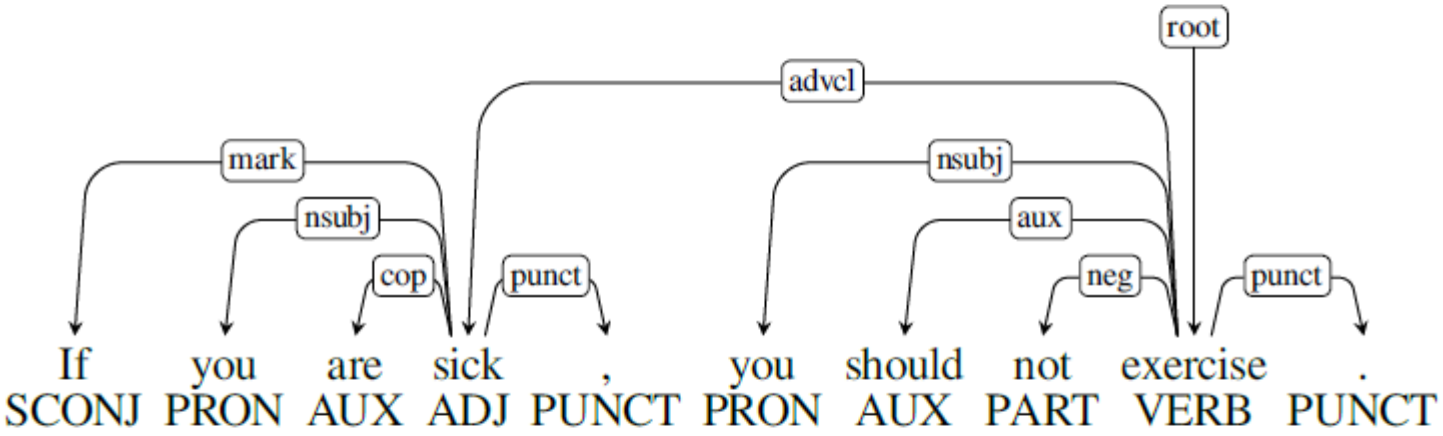
Dependency Relations

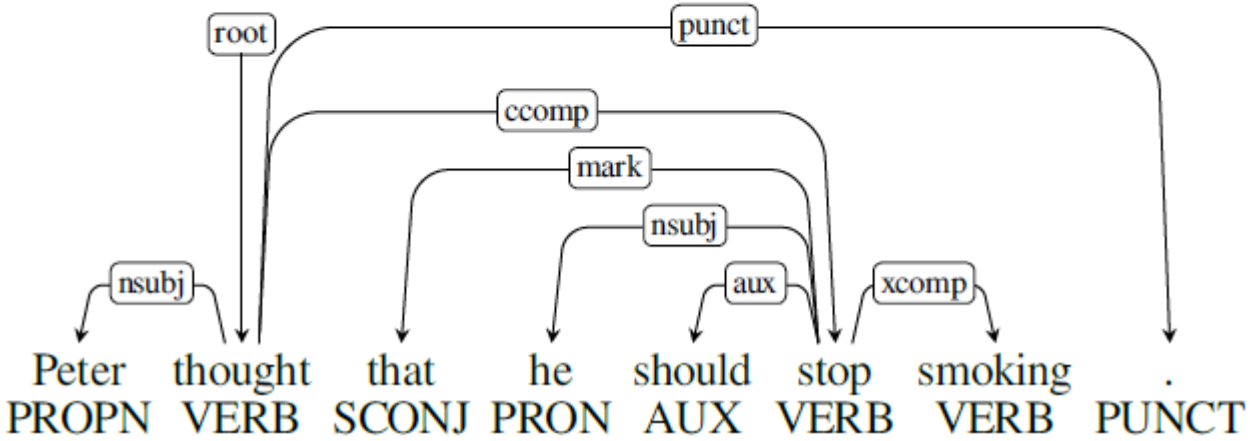
- Taxonomy of 40 universal grammatical relations, broadly attested in language typology (de Marneffe et al., 2014)
  - Language-specific subtypes may be added
- Organizing principles
  - Three types of structures: nominals, clauses, modifiers
  - Core arguments vs. other dependents (not arguments vs. adjuncts)

Dependents of Clausal Predicates

	Nominal	Clausal	Other
Core	nsubj nsubjpass dobj iobj	csubj csubjpass ccomp xcomp	
Non-Core	nmod vocative discourse expl	advcl	advmod neg aux auxpass cop mark punct

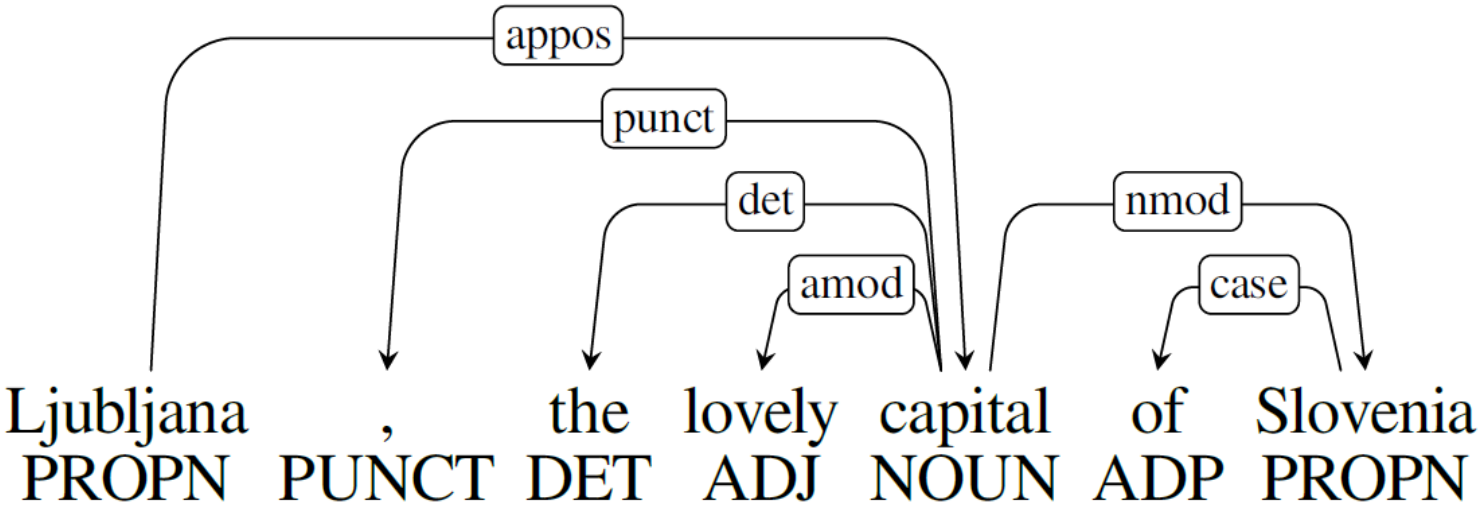

root
punct
nsubj
nmod
aux
dobj
case
advmod
det
det
Mary was quietly reading a book in the garden .
PROPN AUX ADV VERB DET NOUN ADP DET NOUN PUNCT


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

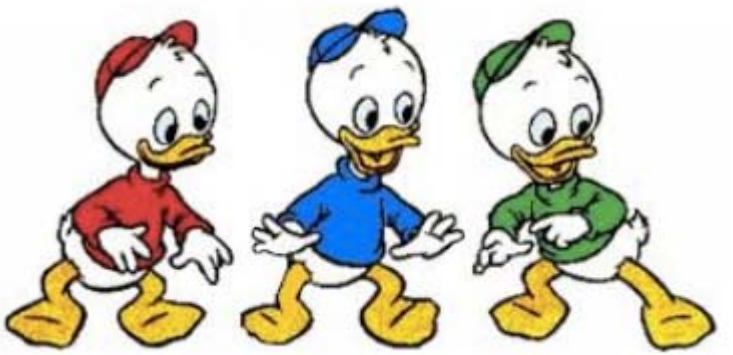

root
punct
ccomp
mark
nsubj
nsubj
aux
xcomp
Peter thought that he should stop smoking .
PROPN VERB SCONJ PRON AUX VERB VERB PUNCT

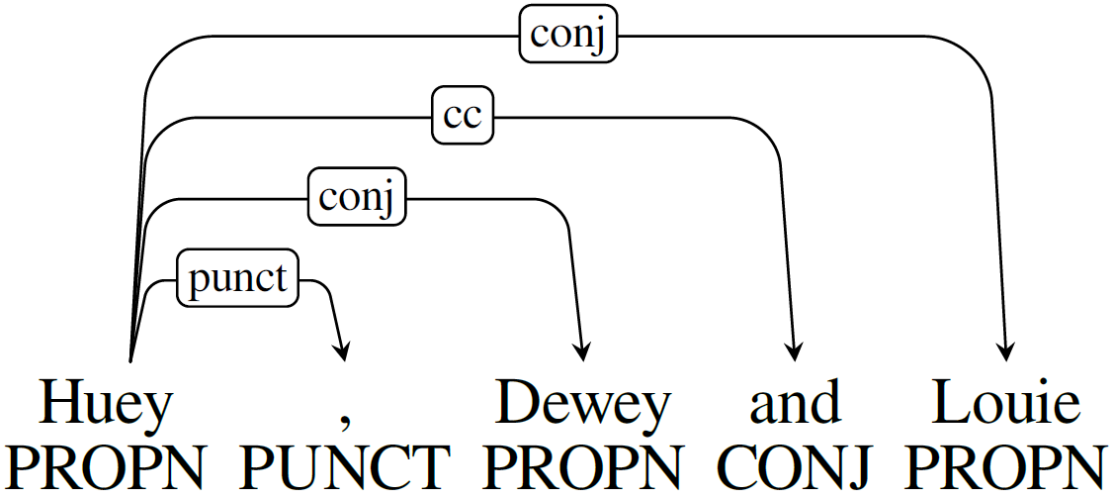
Dependents of Nominals

Nominal	Clausal	Other
nmod appos nummod	acl	amod det neg case



“Stanford-style” Coordination





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

Multiword Expressions

Relation	Examples
mwe	*in spite of, as well as, ad hoc*
name	*Roger Bacon, New York*
compound	*phone book, four thousand, dress up*
goeswith	*notwith standing, with out*

- UD annotation does not permit “words with spaces”
  - Multiword expressions are analyzed using special relations
  - The **mwe**, **name** and **goeswith** relations are always head-initial
  - The **compound** relation reflects the internal structure

Other Relations

Relation	Explanation
parataxis	Loosely linked clauses of same rank
list	Lists without syntactic structure
remnant	Orphans in ellipsis linked to parallel elements
reparandum	Disfluency linked to (speech) repair
foreign	Elements within opaque stretches of code switching
dep	Unspecified dependency
root	Syntactically independent element of clause/phrase

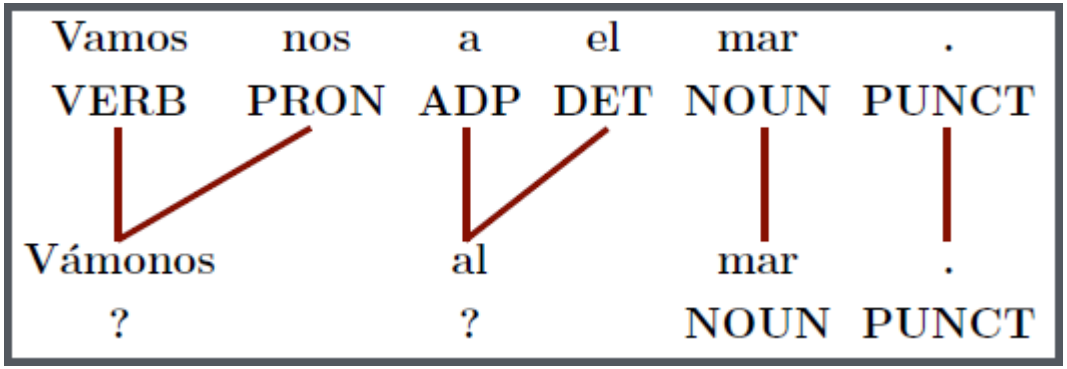
Language-Specific Relations

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

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

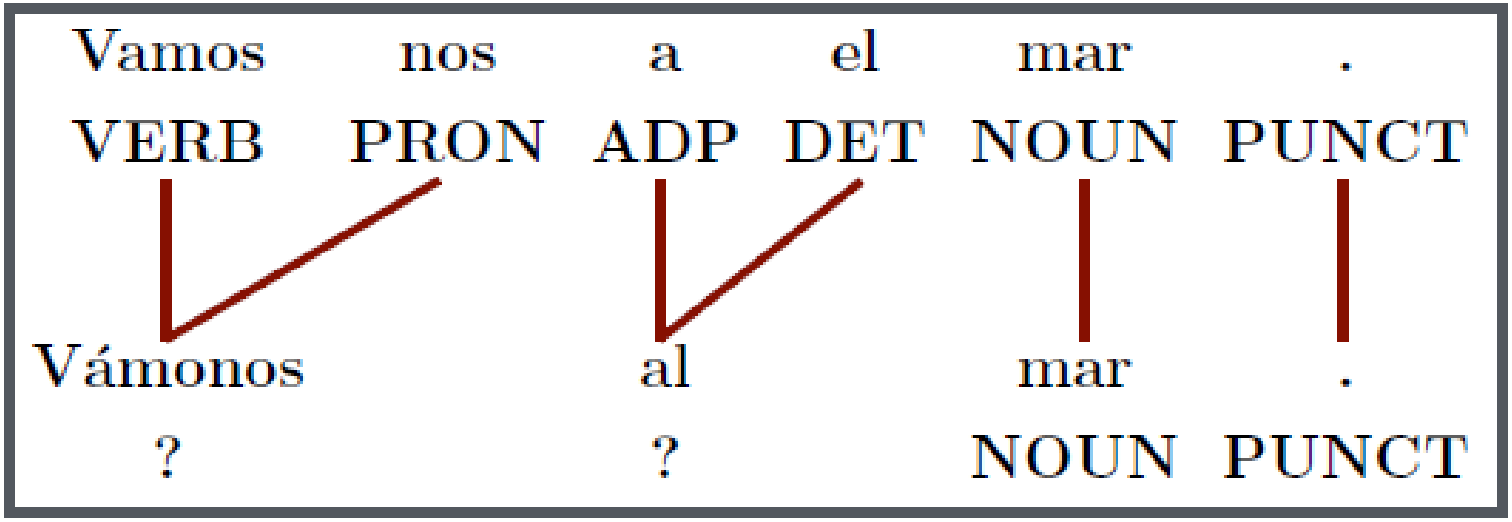
Word Segmentation

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



Word Segmentation

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



Where Are We Now?

- Two years of UD version 1
- 4 treebank releases (every 6 months)
- 54 (61) treebanks
- 40 (47) languages (over 50% world's population)
- Over 11M tokens; treebanks range from 1K to 1.5M
- Over 120 contributors
  - language group consistency SIGs
  - version 2 guidelines coming soon

47 Languages and Growing

Ancient Greek-PROIEL	206K	LF	-	⚙		Irish	23K	LF	🗋	⚙✓
Arabic	242K	LF	-	⚙		Italian	252K	LF	🗋	⚙✓
Basque	121K	LF	🗋	⚙		Japanese-KTC	267K	L	🗋	⚙
Bulgarian	156K	LF	🗋	⚙✓		Kazakh	4K	L	🗋	👤
Buryat	5K	L	-	👤		Korean	-		-	-
Catalan	530K	LF	🗋	⚙✓		Latin	47K	LF	-	⚙
Chinese	123K	F	🗋	⚙✓		Latin-ITTB	291K	LF	-	⚙
Coptic	4K	L	🗎	👤		Latin-PROIEL	165K	LF	-	⚙
Croatian	87K	LF	-	⚙✓		Latvian	20K	LF	-	⚙
Czech	1,503K	LF	🗎	⚙✓		Norwegian	311K	LF	🗋	⚙
Czech-CAC	493K	LF	🗎	⚙✓		Old Church Slavonic	57K	LF	-	⚙
Czech-CLTT	35K	LF	🗎	⚙✓		Persian	151K	F	🗋	⚙✓
Danish	100K	LF	🗋	⚙✓		Polish	83K	LF	-	⚙
Dutch	209K	LF	-	⚙		Portuguese	209K	LF	-	⚙
Dutch-LassySmall	98K	LF	-	⚙		Portuguese-BR	298K	F	-	⚙
English	254K	LF	🗎	👤		Romanian	145K	LF	🗋	⚙✓
English-ESL	97K	L	🗎	👤		Russian	99K	F	🗋	⚙✓
English-LinES	82K		🗋	⚙✓		Russian-SynTagRus	1,032K	LF	🗋	⚙✓
Estonian	234K	LF	-	⚙✓		Sanskrit	1K	LF	-	⚙
Faroese	119K	F	-	⚙		Slovenian	140K	LF	🗋	⚙
Finnish	181K	LFD	🗎	⚙✓		Slovenian-SST	29K	LF	🗋	👤
Finnish-FTB	159K	LF	-	⚙✓		Spanish	423K	LF	🗋	⚙✓
French	390K	LF	🗎	⚙✓		Spanish-AnCora	547K	LF	🗋	⚙✓
Galician	138K	L	🗋	⚙✓		Swedish	96K	LF	🗎	⚙✓
German	293K	LF	-	⚙		Swedish-LinES	79K		🗋	⚙✓
Gothic	56K	LF	-	⚙		Tamil	8K	LF	-	⚙
Greek	59K	LF	🗋	⚙		Turkish	56K	LF	🗋	⚙
Hebrew	115K	F	-	⚙		Ukrainian	-		-	⚙✓
Hindi	351K	LF	-	⚙		Urdu	-		-	⚙
Hungarian	42K	LF	🗎	👤		Uyghur	45K	F	-	⚙
Indonesian	121K		-	⚙		Vietnamese	43K	L	-	⚙

Where Are We Going?

- UD guidelines version 2 coming soon
- Consistency checking

Common vocabulary is great …

… because we finally
understand each other …

… almost

Childs of you be vary acute!

Consistency Checking

- Automatic tests catch only a fraction
- Focus groups on
  - Romance, Germanic, Slavic, Uralic, Turkic languages

Existing Slavic Treebanks

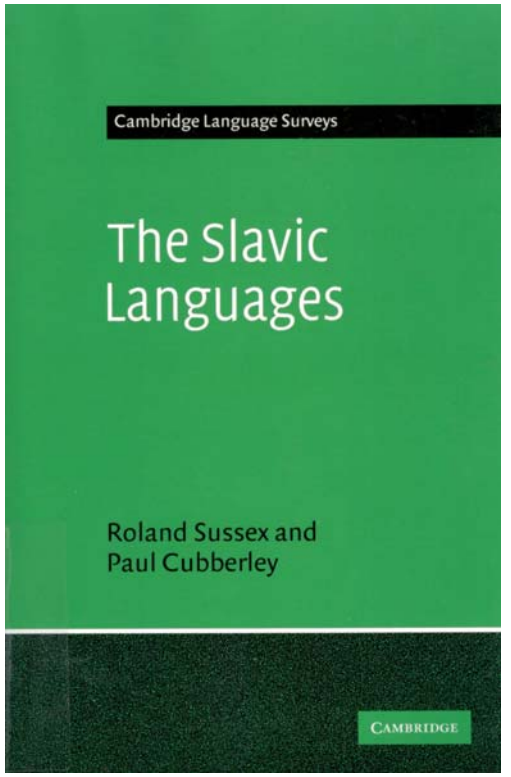
Language	Code	Treebank	Sent	Tok
Bulgarian	[bg]	BulTreeBank	13,221	196K
Church Slavonic	[cu]	PROIEL	7,818	72K
Croatian	[hr]	SETimes.HR	3,736	84K
Croatian	[hr]	HOBS	4,626	117K
Czech	[cs]	PDT	87,913	1504K
Czech	[cs]	CAC	24,709	494K
Polish	[pl]	IPI PAN	8,227	84K
Russian	[ru]	SynTagRus	59,130	1033K
Russian	[ru]	Google	5,030	99K
Slovak	[sk]	SNK	63,238	994K
Slovenian	[sl]	SSJ500K	27,829	500K
Slovenian	[sl]	SST	3,188	29K

Issues of Slavic Languages in UD

- Pronouns vs. determiners, numerals and quantifiers
- Attachment of cardinal numbers
- Verbs, participles, adjectives
- Core arguments
- Reflexive pronouns (clitics)
- Auxiliary verbs and modal verbs
- Comparative constructions

Pronouns and Determiners

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



Pronouns and Determiners

- English + Romance languages: DET = article or pronominal adjective *(this, which, every)*
- We don’t have this category! (Traditionally → PRON.)
- We have the words (except for articles).
- Currently functional borderline (but ellipsis?)
  This.DET *car is expensive.*
  This.PRON *is expensive.*
- Less strict in UD v2.

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: *ja, 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, ты, 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.
My father is older. But yours is more experienced.

- sl: *moj, tvoj, njegov, njen, najin, vajin, njun, naš, vaš, njihov, svoj*
- bg: *мой, твой, негов, неин, наш, ваш, техен, свой*
- cs: *můj, tvůj, jeho, její, náš, váš, jejich, svůj*
- sk: *môj, tvoj, jeho, jej, náš, váš, ich, svoj*
- cu: *мои, твои, нашь, вашь, свои / его, еѭ, ею, ихъ*

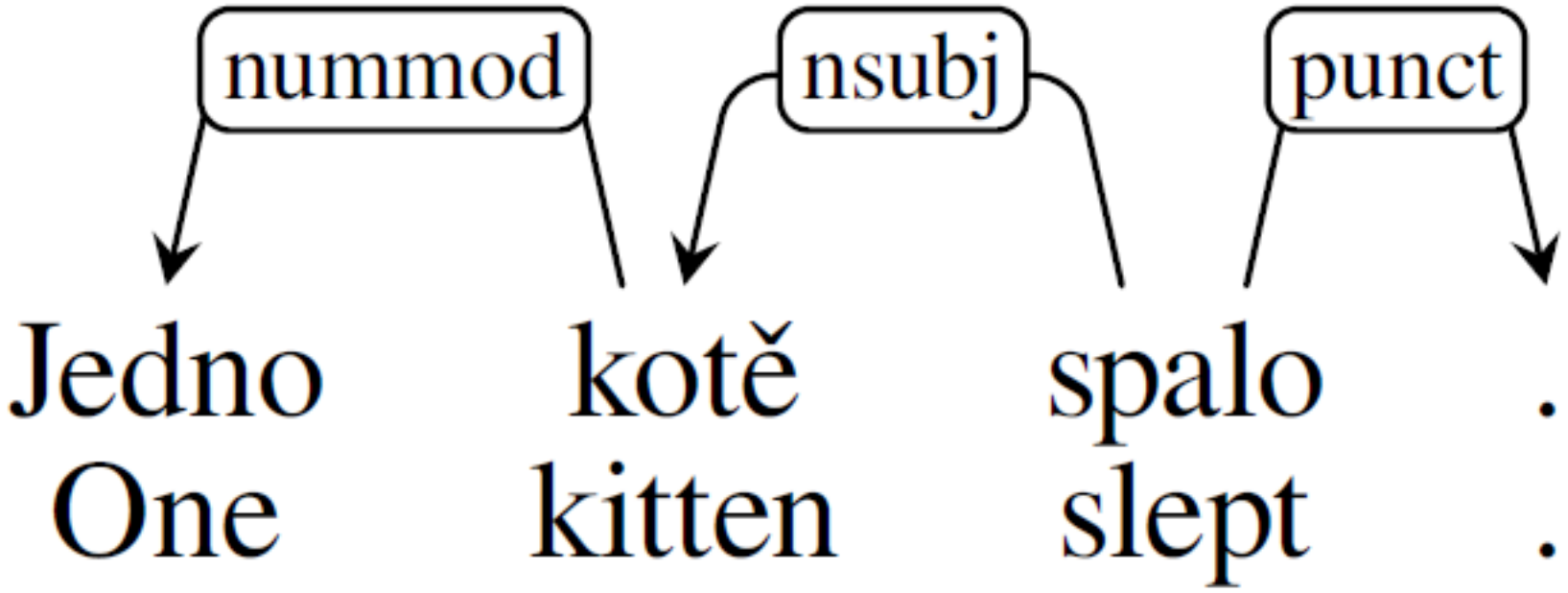
Both Possible?

- Demonstratives
  - cs: *ten, to, tento, tenhle, tamten, …*
  - sl: *ta, to, tisti, oni, takšen, …*
- 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.*

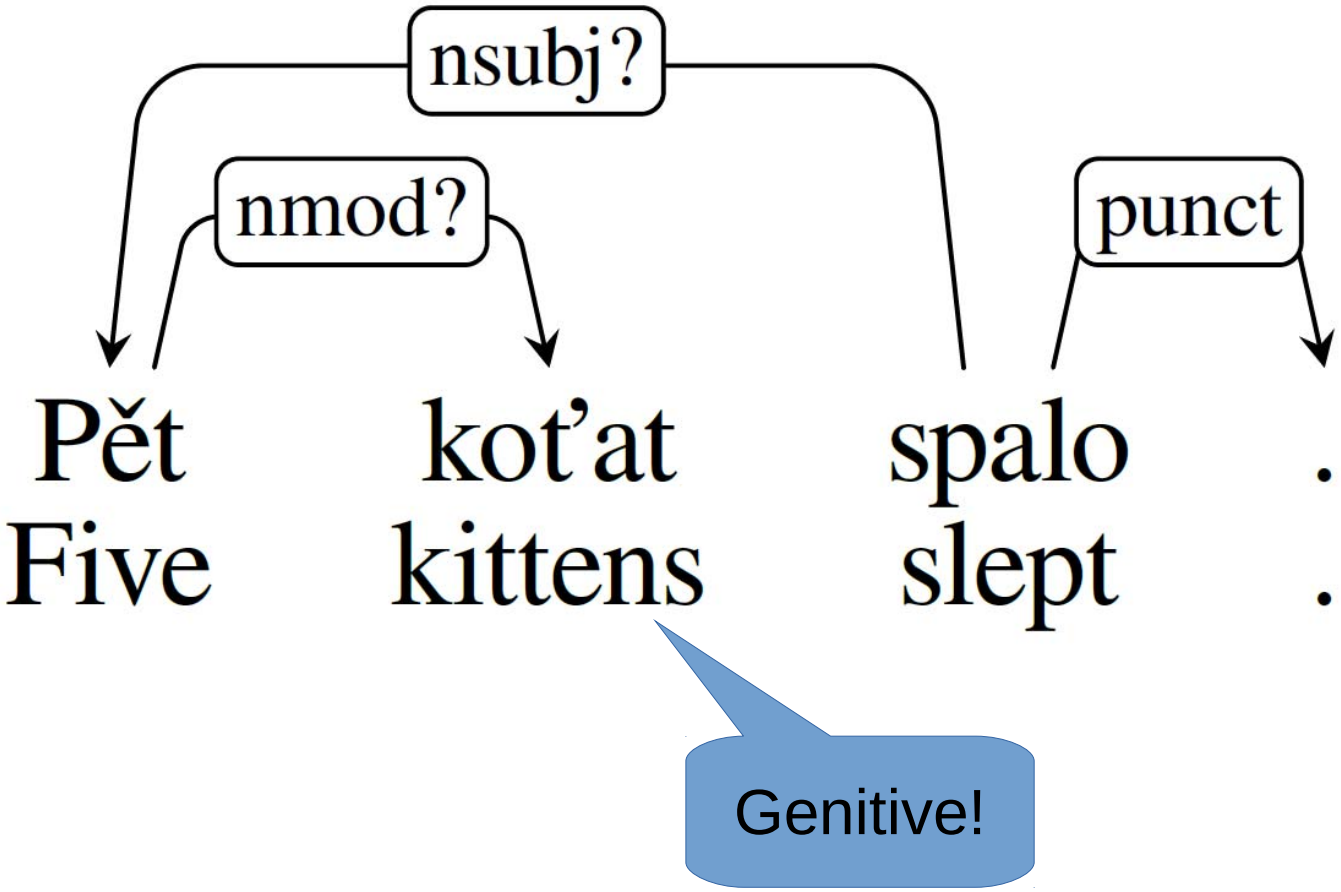
Issues of Slavic Languages in UD

- Pronouns vs. determiners, numerals and quantifiers
- Attachment of cardinal numbers
- Verbs, participles, adjectives
- Core arguments
- Reflexive pronouns (clitics)
- Auxiliary verbs and modal verbs
- Comparative constructions

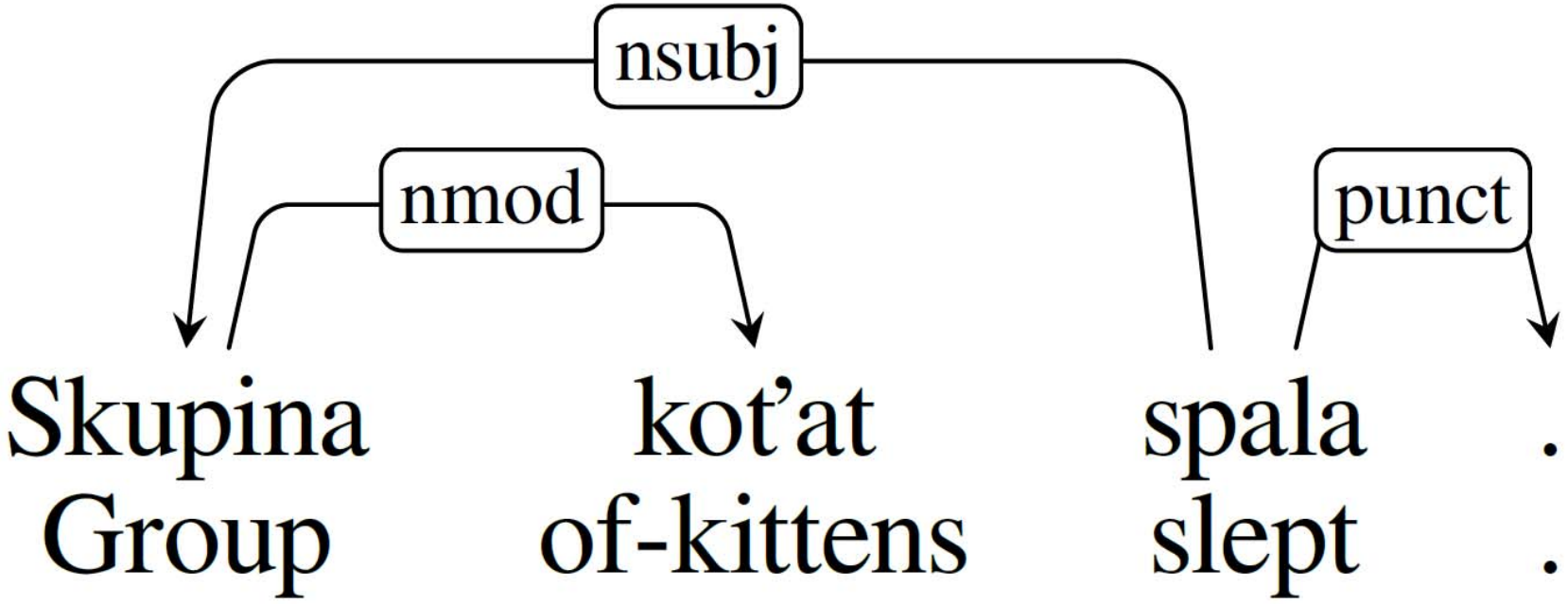
Quantified Noun Phrase



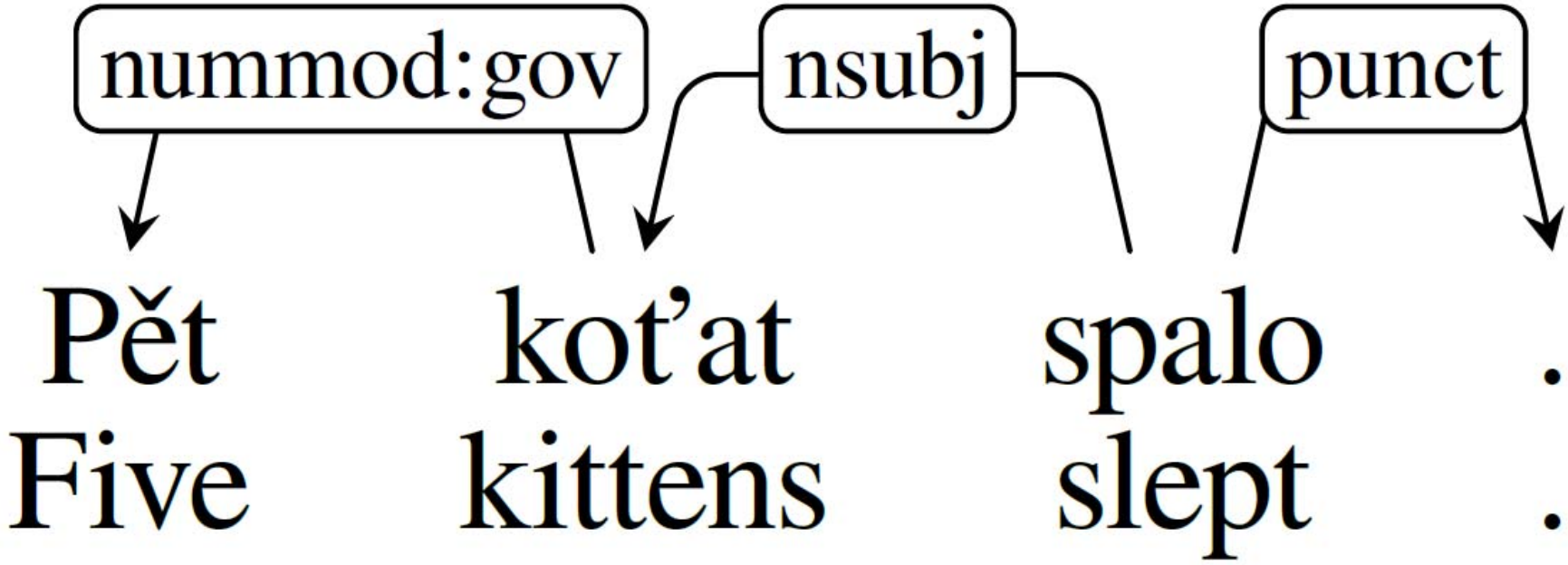
Quantified Noun Phrase



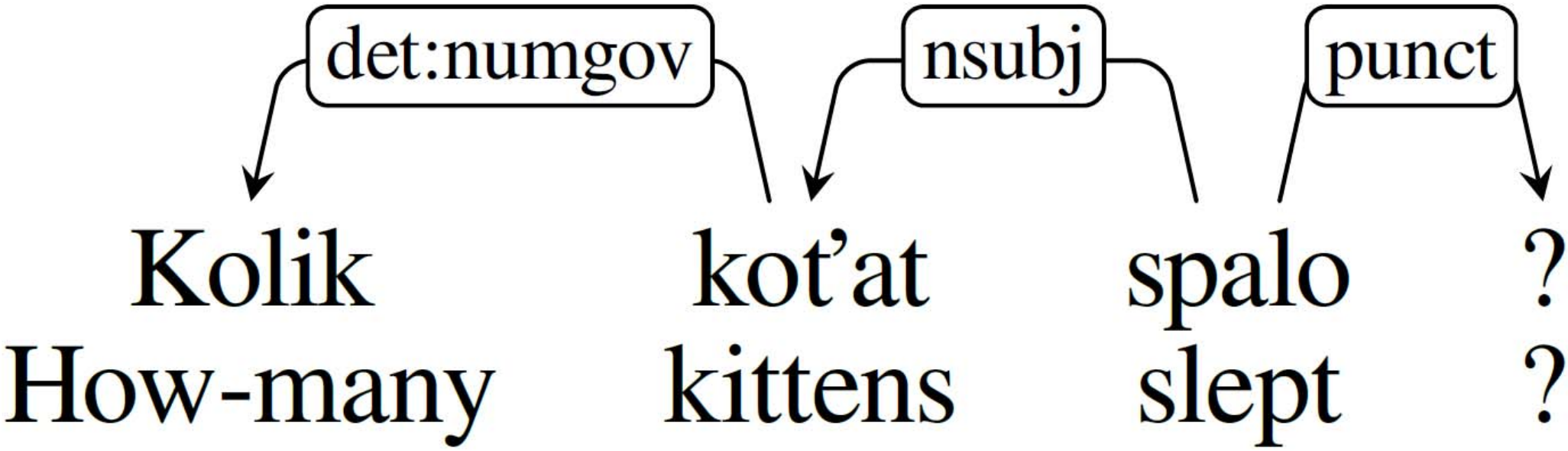
Quantified Noun Phrase



Quantified Noun Phrase



Pronominal Quantifiers



Language-Specific Labels

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

Issues of Slavic Languages in UD

- Pronouns vs. determiners, numerals and quantifiers
- Attachment of cardinal numbers
- Verbs, participles, adjectives
- Core arguments
- Reflexive pronouns (clitics)
- Auxiliary verbs and modal verbs
- Comparative constructions

Verb Forms

- Conflicting terminologies in traditional grammars
- Participle … verb or adjective?
- Converb … verb or adverb?

- Tags and features apply to individual words!

Verb Forms

- POS tags and features apply to individual words!
- *A ko so se leta 1942 vračali, …*
  - past tense
- *… da ne bi v Atene prišli …*
  - conditional mood
- *… v prihodnje ne bodo vozili zgolj les …*
  - future tense

Verb Forms

- POS tags and features apply to individual words!
- *A ko so se leta 1942 vračali, …*
  - t tense
- *… da ne bi v Atene prišli …*
  - nal mood
- *… v prihodnje ne bodo vozili zgolj les …*
  - future tens

Present

Past???

Participle

Conditional

Participle

Future

Participle

Verb Forms

- *vračali, prišli, vozili*

- [cs] “active participle” / “past tense”
- [ru] “past tense” / “finite!”
  - Active participle is something else: *нарушивший*
- [bg] “participle + past (aorist) / imperfect” (two subtypes)
- [cu] “participle + resultative aspect” (lang-spec)
- “l-participle”
  - But that would be a language-specific verb form.

Issues of Slavic Languages in UD

- Pronouns vs. determiners, numerals and quantifiers
- Attachment of cardinal numbers
- Verbs, participles, adjectives
- Core arguments
- Reflexive pronouns (clitics)
- Auxiliary verbs and modal verbs
- Comparative constructions

Core Arguments

- Easier cross-linguistically than argument-adjunct?
- Subject of intransitive verb
- Agent of transitive verb
- Patient (direct object) of transitive verb

- Indirect object? Dative only?

Core vs. Oblique Dependents

- **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

dobj / iobj

- Not as easy as accusative vs. dative.
- Default: dobj
- Heuristics for iobj
  - *Cením si vaší pomoci.* (Gen)
    I appreciate your help.
  - *Čelíme velkým problémům.* (Dat)
    We are facing big problems.
  - *Nedisponuje takovým rozpočtem.* (Ins)
    He does not have such budget.
  - *Učí mou dceru fyziku.* (2 × Acc)
    He teaches my daughter physics.

All Slavic Treebanks Have Non-Accusative “Direct” Objects

- *podrobit se testu; odpovídají smlouvě; jednat s někým*
- *mówi o niej; używa wielkich słów*
- *от которых зависит; относится к программам*
- *potrebuje informacij; slediti evropskim smernicam; ukvarjal se bom orožjem*
- *odriče se imuniteta; priključiti se naporima*
- *се характеризира с развитие; моля за внимание*

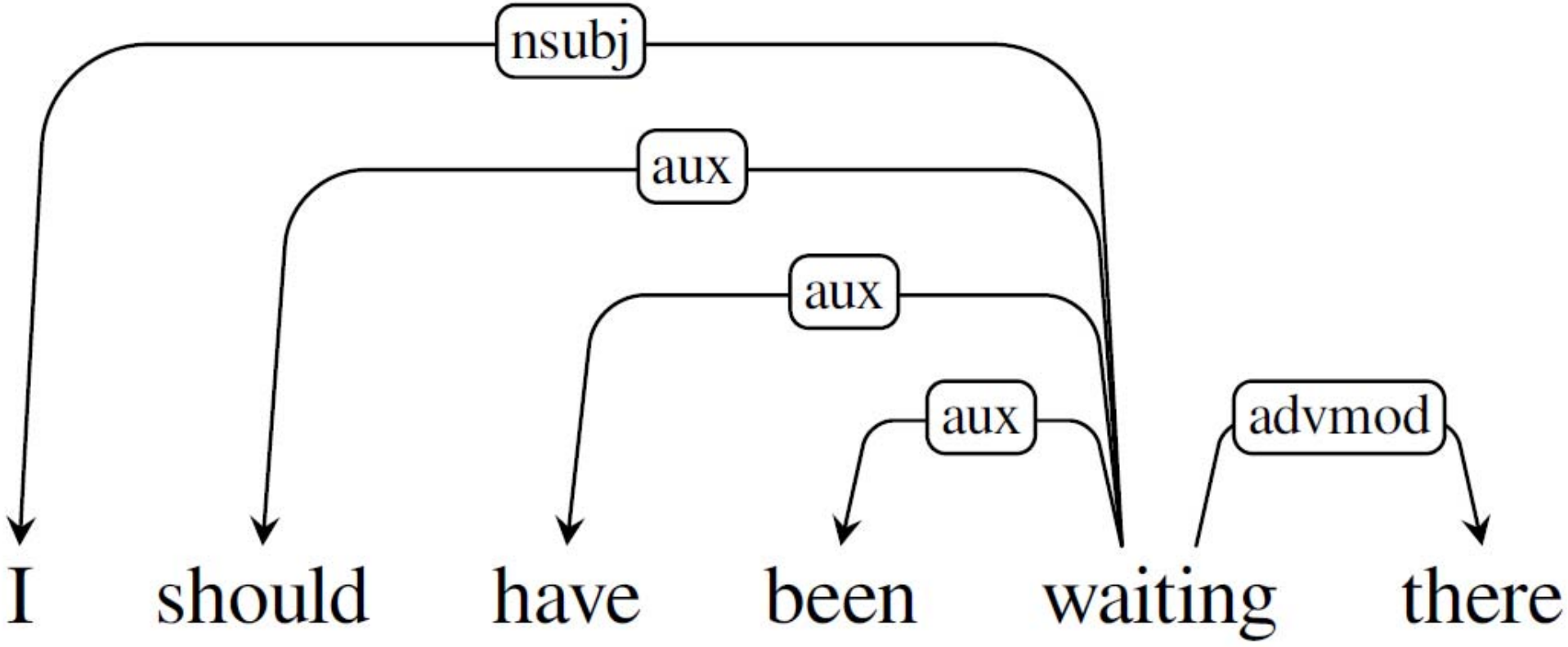
Reflexive Pronouns

- 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 / laugh, fear*
  - `expl:pv` (pronominal verb; previously `compound`)
- Reflexive passive:
  To se snadněji řekne než udělá. / That is easier said than done.
  - `expl:pass` (previously `auxpass:reflex`)
- Impersonal construction (~ passive?):
  Zde se mluví německy. / German is spoken here.
  - `expl:impers`

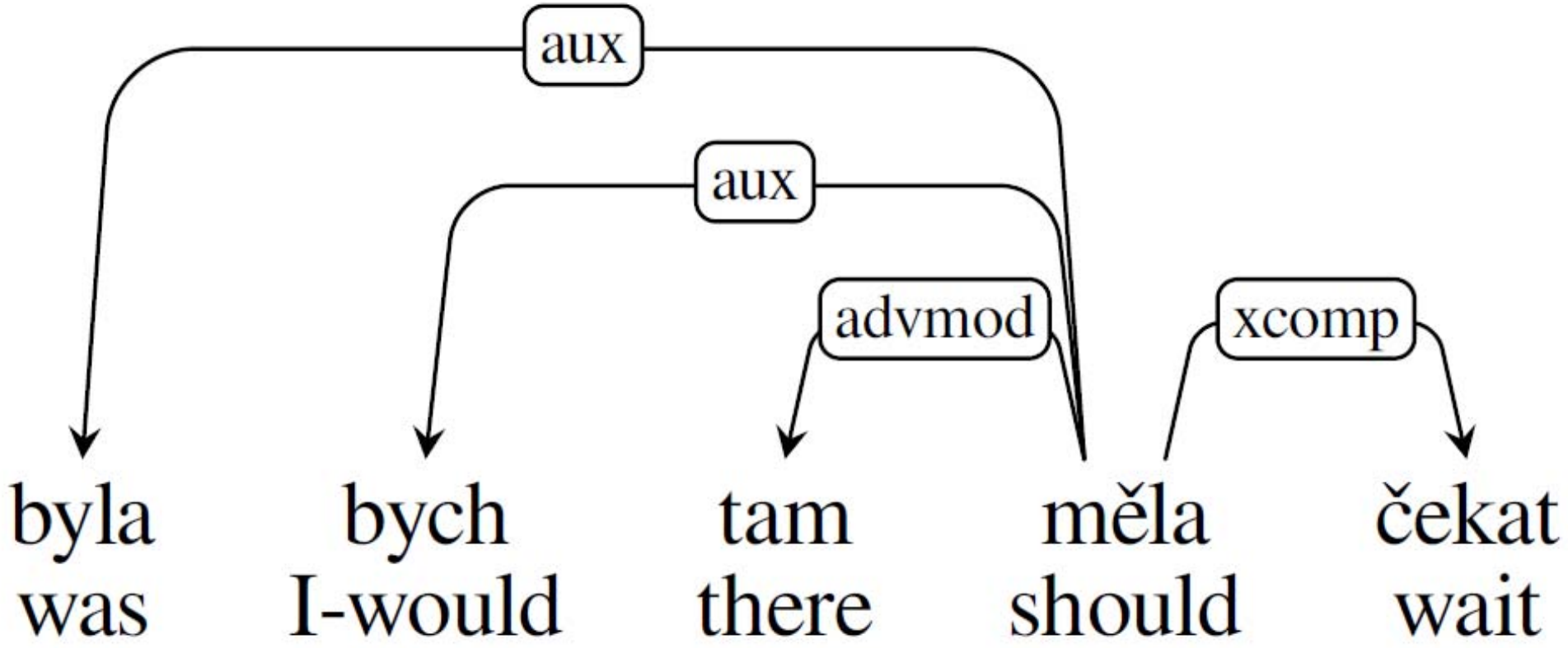
Issues of Slavic Languages in UD

- Pronouns vs. determiners, numerals and quantifiers
- Attachment of cardinal numbers
- Verbs, participles, adjectives
- Core arguments
- Reflexive pronouns (clitics)
- Auxiliary verbs and modal verbs
- Comparative constructions

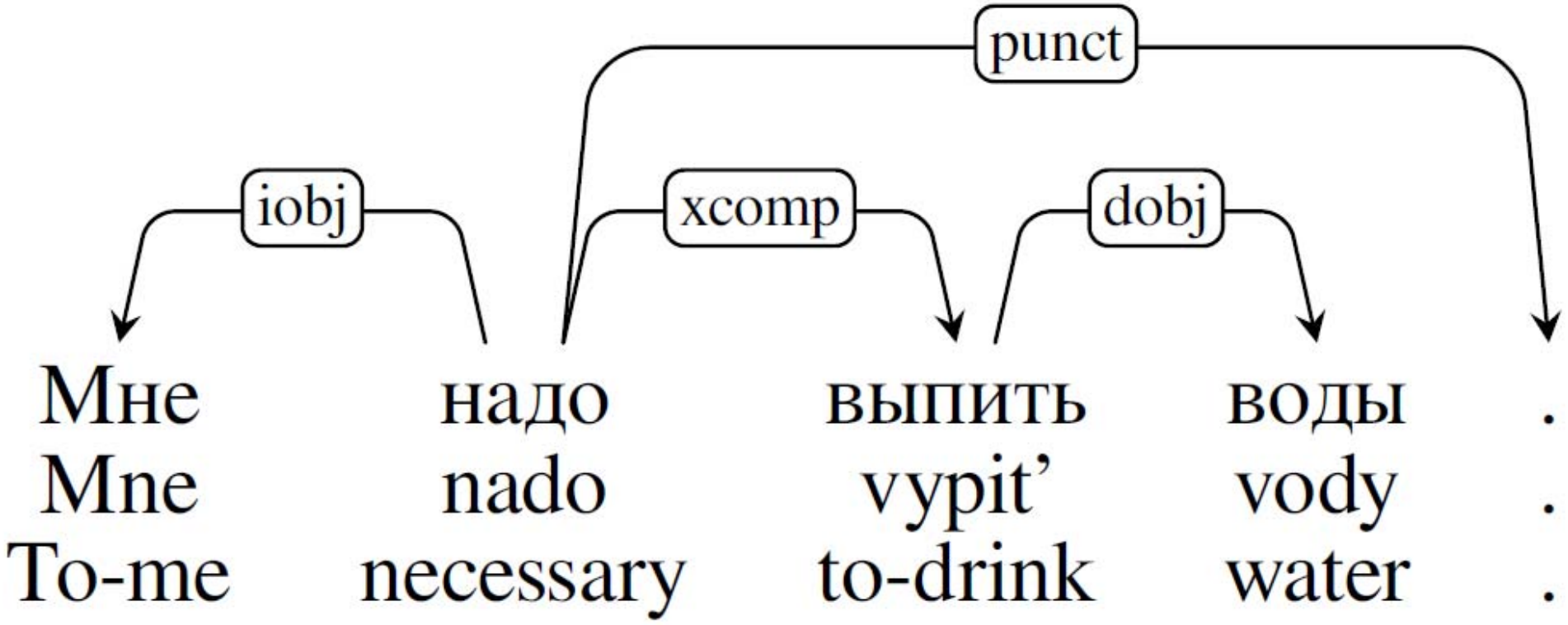
Modal Auxiliary in English



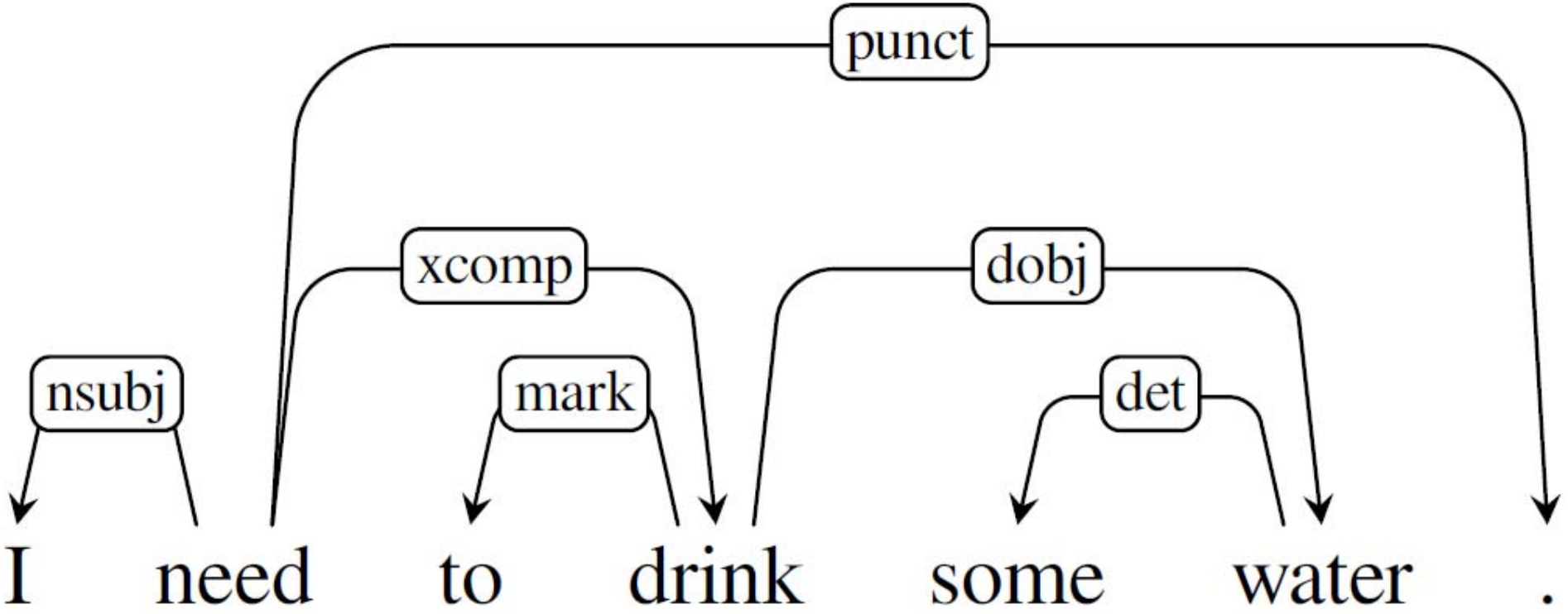
Modal Verb in Czech



Modal Adverb in Russian



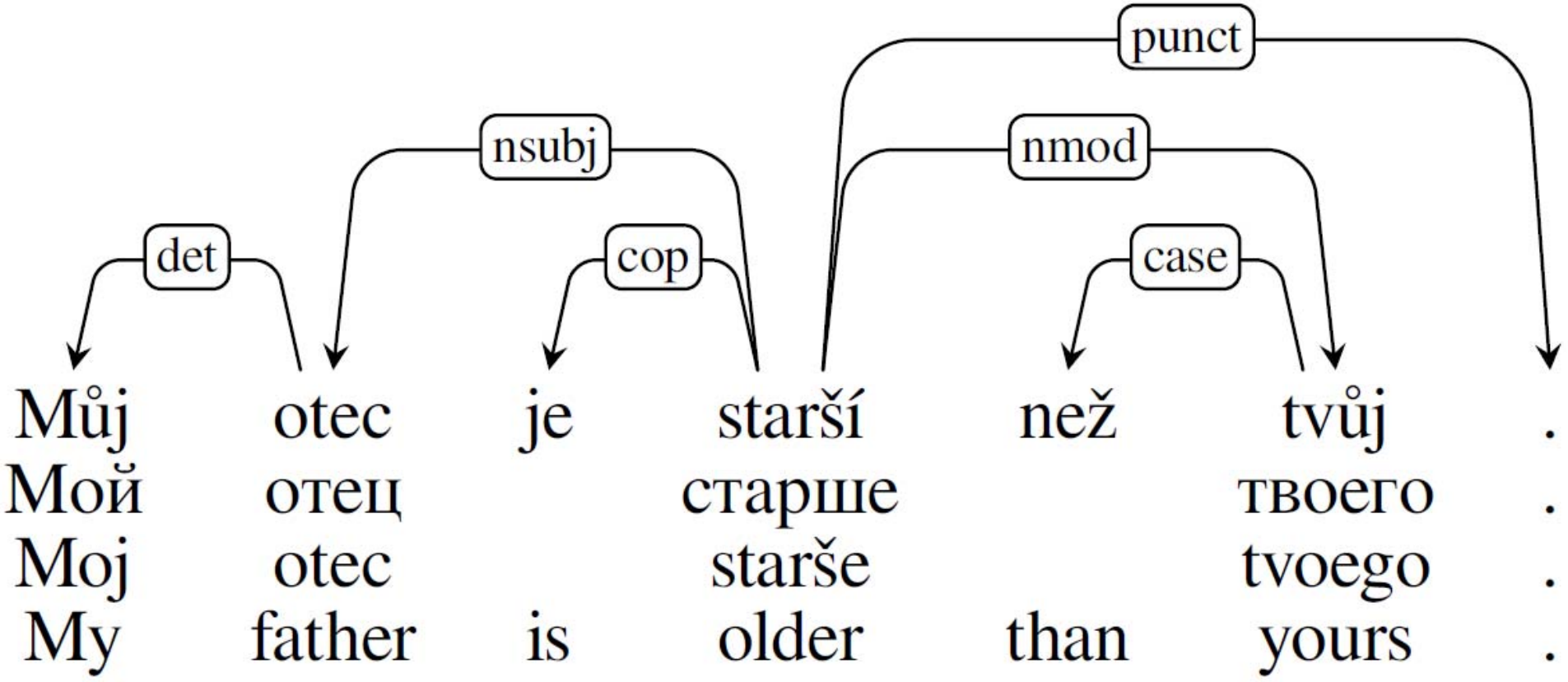
Modal / Control Verb in English



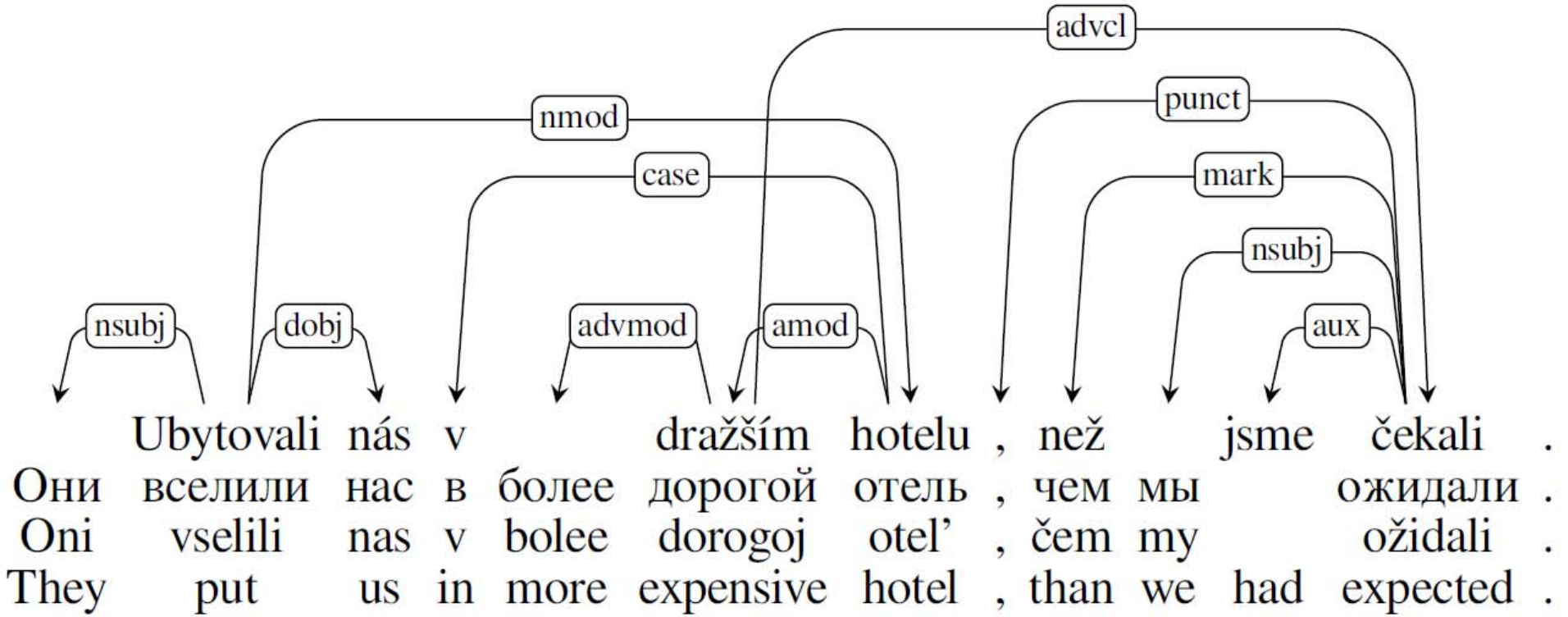
Issues of Slavic Languages in UD

- Pronouns vs. determiners, numerals and quantifiers
- Attachment of cardinal numbers
- Verbs, participles, adjectives
- Core arguments
- Reflexive pronouns (clitics)
- Auxiliary verbs and modal verbs
- Comparative constructions

Comparative Constructions



Comparative Constructions



Wrapping Up

Wrapping Up

- UD has had a great start
- Still a long way to go.
  Consistency matters!
- Get involved. It’s fun!

Děkuji!
Otázky?

Благодарѭ!
Въпроси?

Ďakujem!
Otázky?

Благодаря!
Въпроси?

Thank you!
Questions?

Спасибо!
Вопросы?

Dziękuję!
Pytania?

Hvala!
Vprašanja?

Hvala!
Pitanja?