UPPSALA
UNIVERSITET

UNIVERSITÉ
DE GENÈVE

Universal Dependencies as a Resource for Linguistic Research

Joakim Nivre

UD in a Nutshell

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

UD for Linguistic Research

Theory

- Can we make linguistic sense of UD representations?
- Are dependencies syntactic or semantic?

Data

- What kind of linguistic data is available in UD treebanks?
- How diverse are the data sets?

Studies

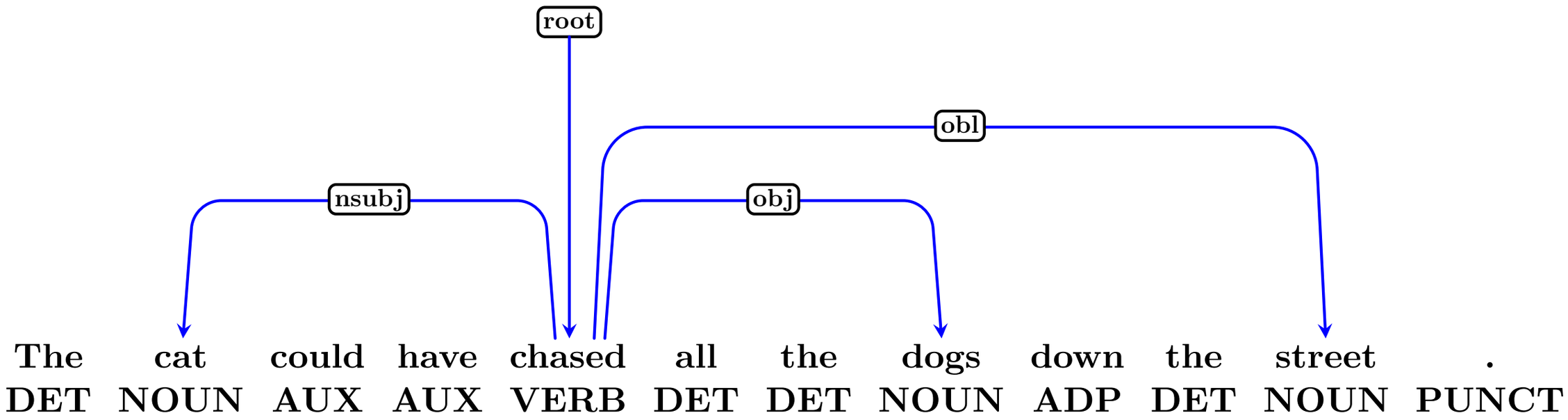
- Two case studies using UD resources

Theory

Syntax

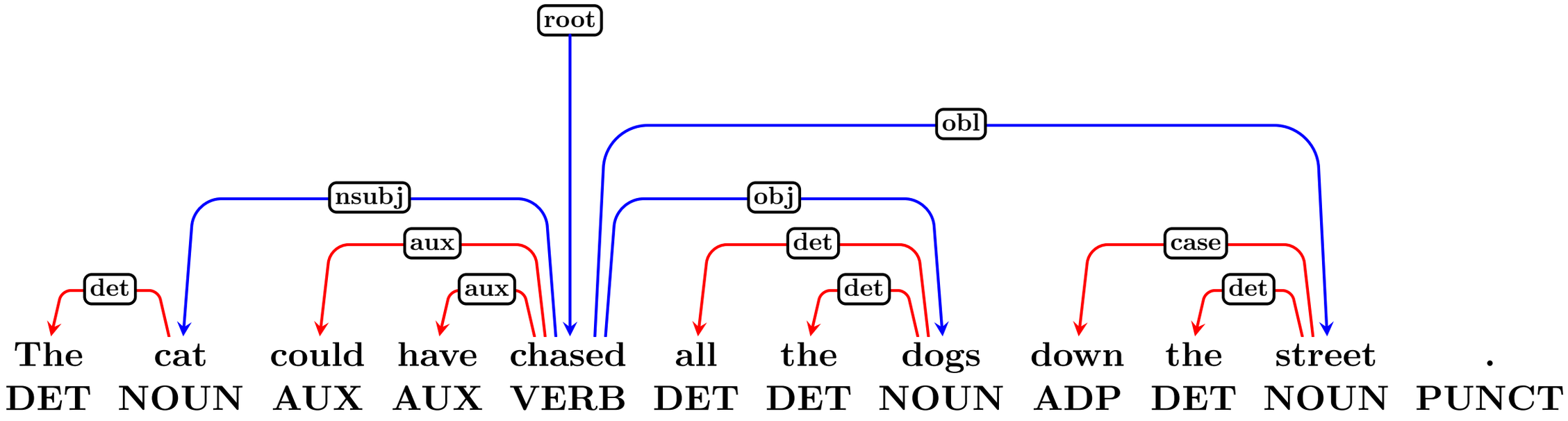
The	cat	could	have	chased	all	the	dogs	down	the	street	.
DET	NOUN	AUX	AUX	VERB	DET	DET	NOUN	ADP	DET	NOUN	PUNCT

Syntax



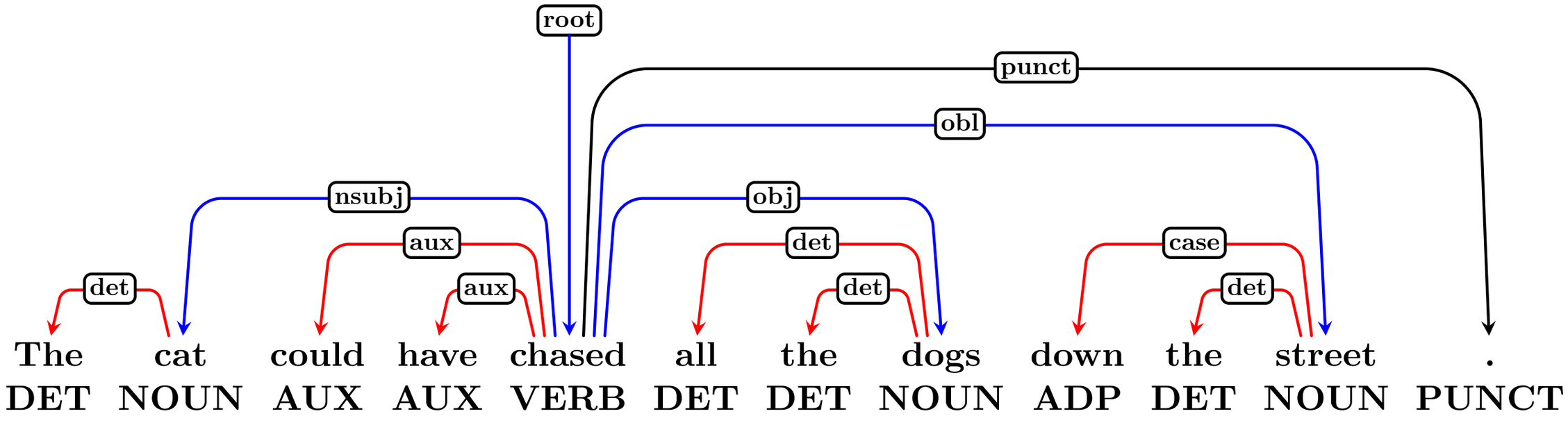
- Content words are related by dependency relations

Syntax



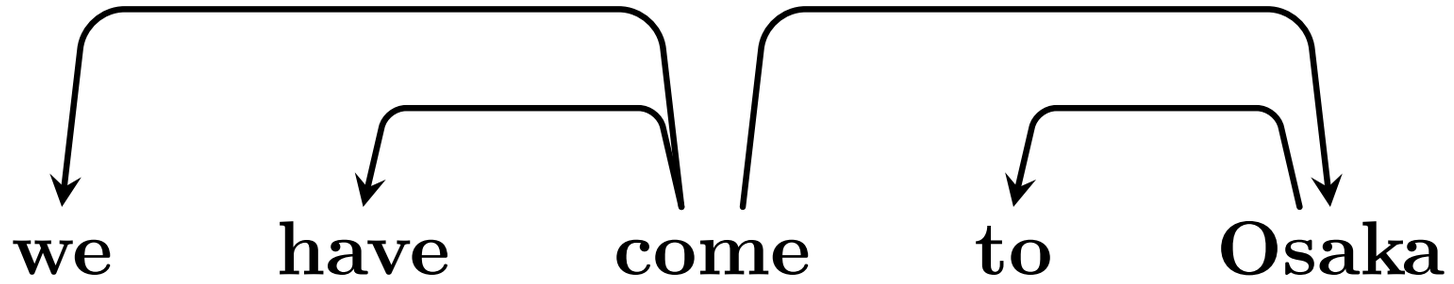
- Content words are related by dependency relations
- Function words attach to the content word they modify

Syntax

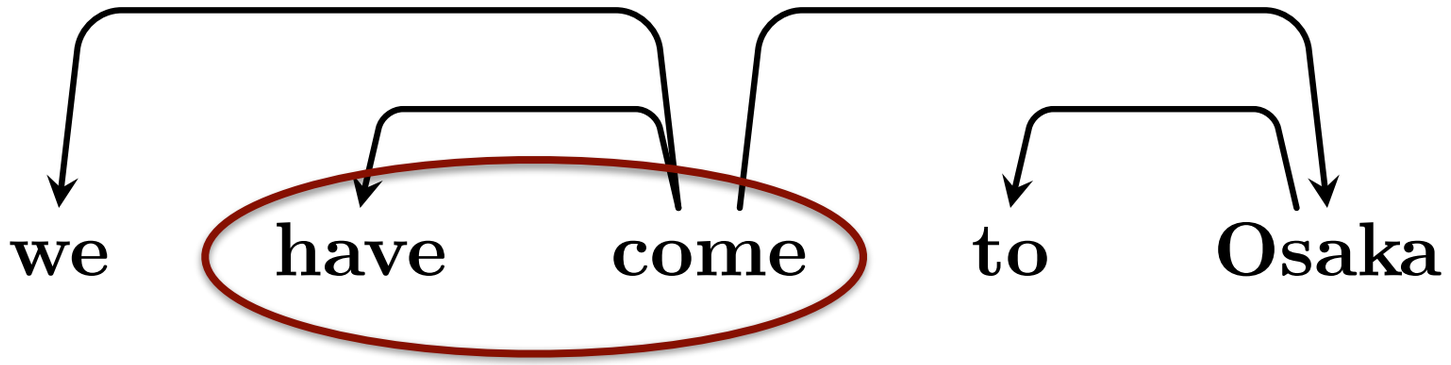


- Content words are related by dependency relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

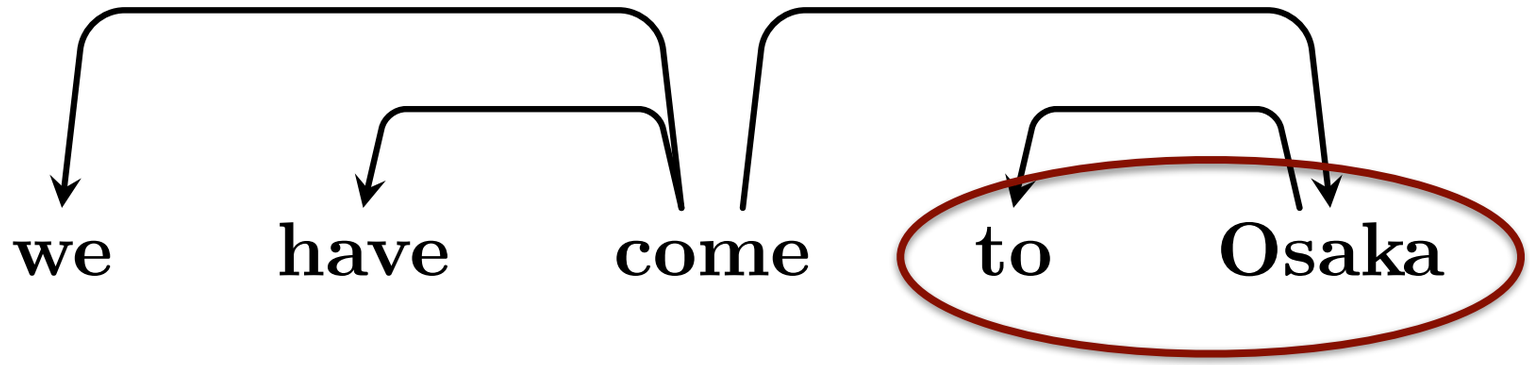
“Content-Head Dependencies”



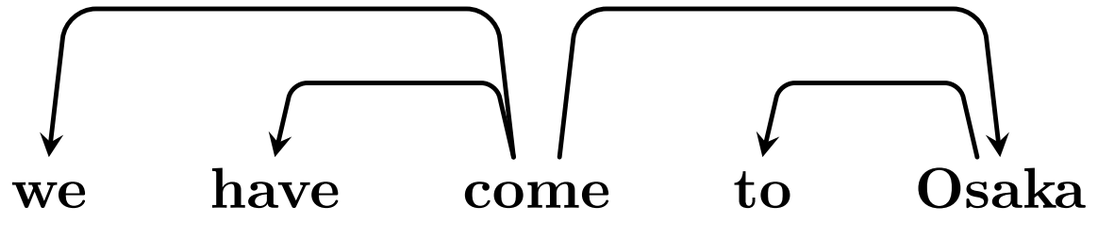
“Content-Head Dependencies”



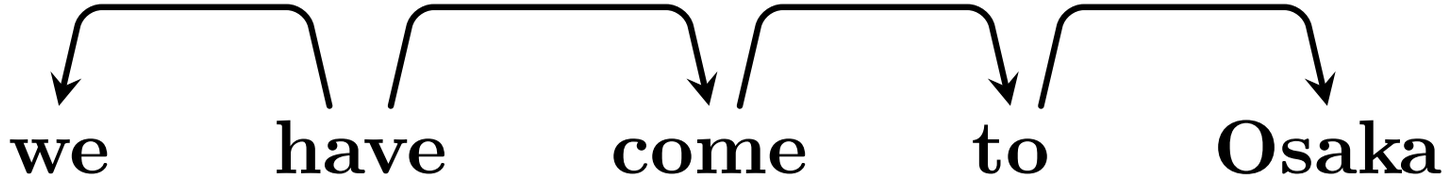
“Content-Head Dependencies”



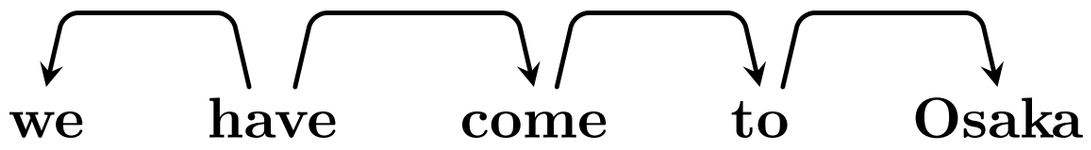
"Content-Head Dependencies"



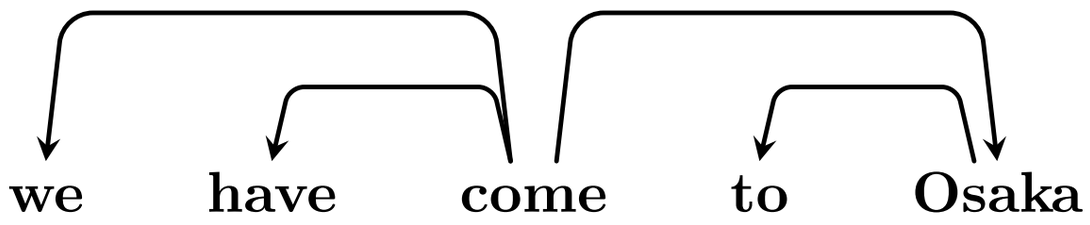
"Function-Head Dependencies"



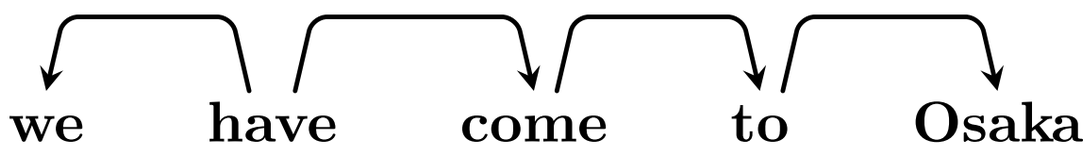
“Function-Head Dependencies”



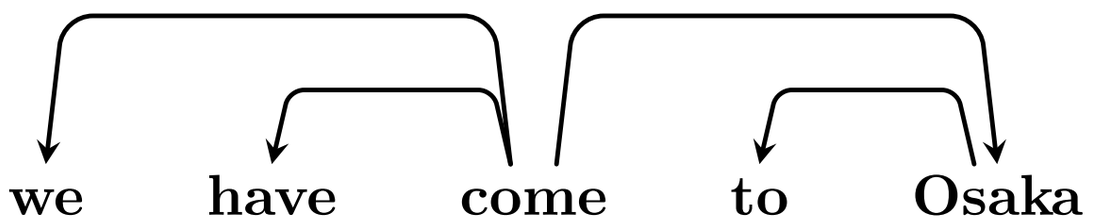
“Content-Head Dependencies”



“Function-Head Dependencies”



“Content-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)

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

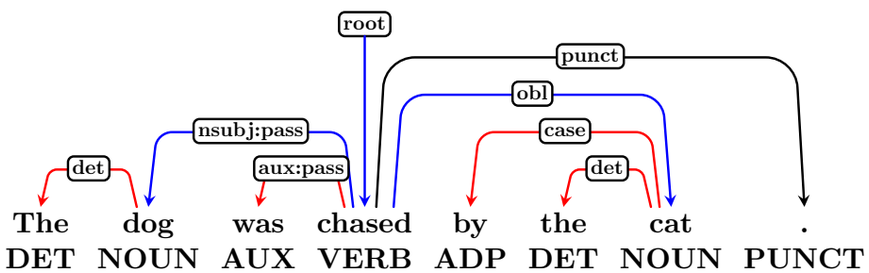
UD Syntax

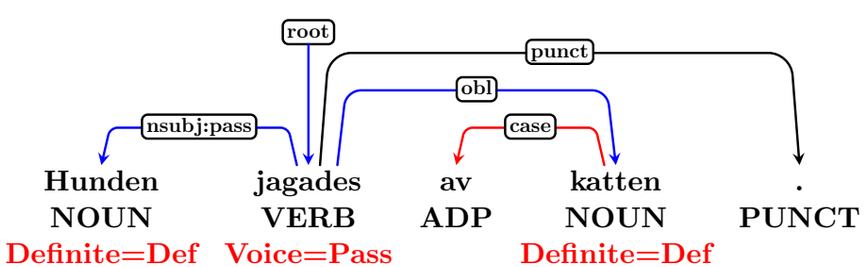
UD representations are mono-stratal – single tree

- 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





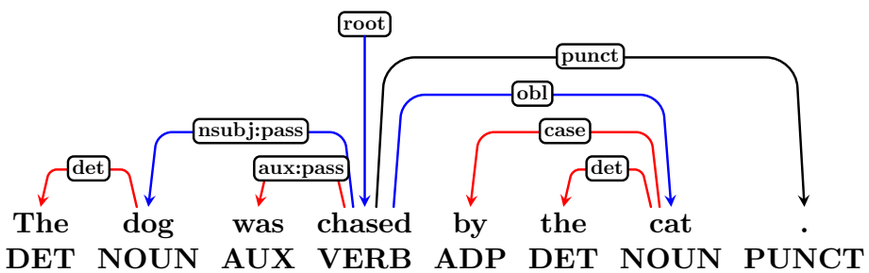
UD Syntax

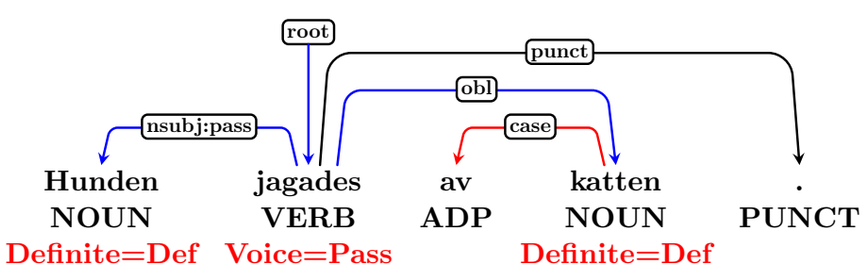
UD representations are mono-stratal – single tree

- Facilitates annotation, parsing and downstream tasks

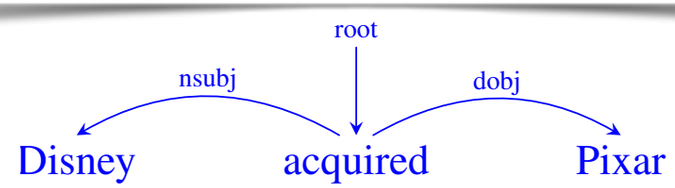
Tree structure primarily reflects lexical dependencies

- 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



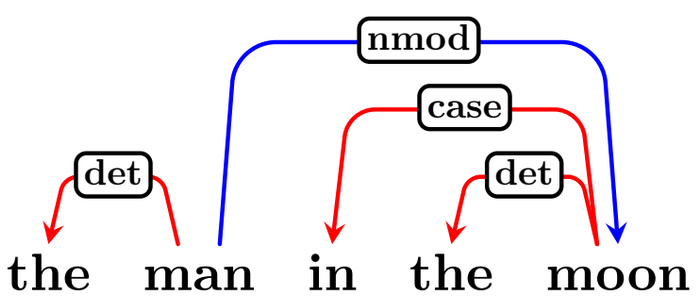
$$(\textbf{nsubj}\ \ (\textbf{dobj}\ \ \textbf{acquired}\ \ \textbf{Pixar})\ \ \textbf{Disney})$$

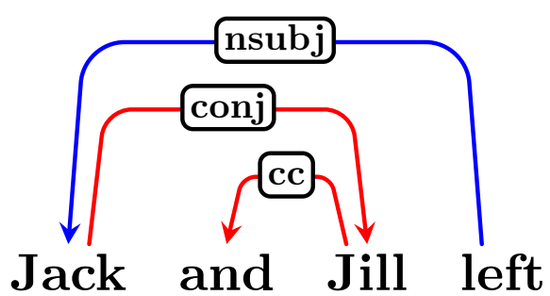
$$\lambda z.\exists xy.\text{acquired}(z_e) \wedge \text{Pixar}(y_a) \wedge \text{Disney}(x_a) \wedge \text{arg}_1(z_e, x_a) \wedge \text{arg}_2(z_e, y_a)$$

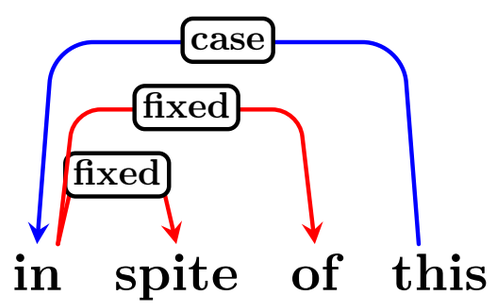
UD Syntax

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



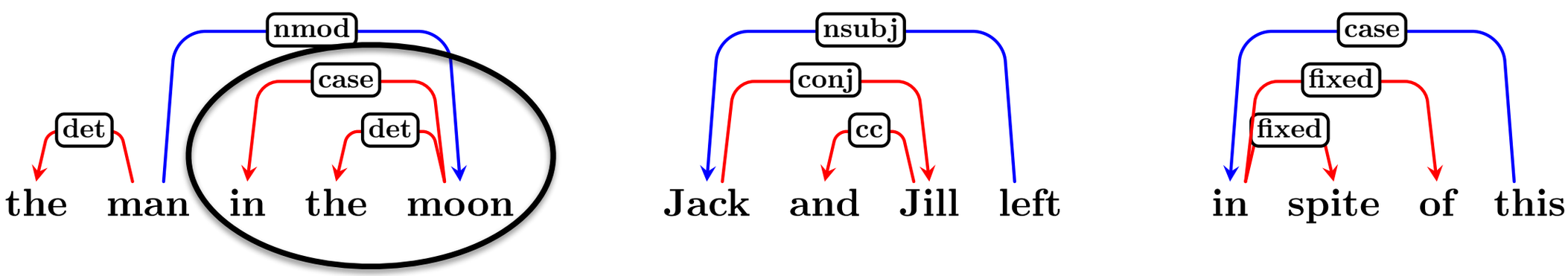




UD Syntax

Other relations encoded in labels – not tree structure

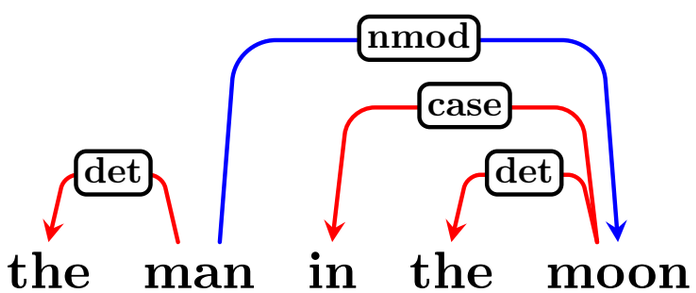
- Functional relations link functional heads to lexical heads
- Coordination relations link equivalent heads/dependents
- Multiword relations link elements of lexicalized expressions

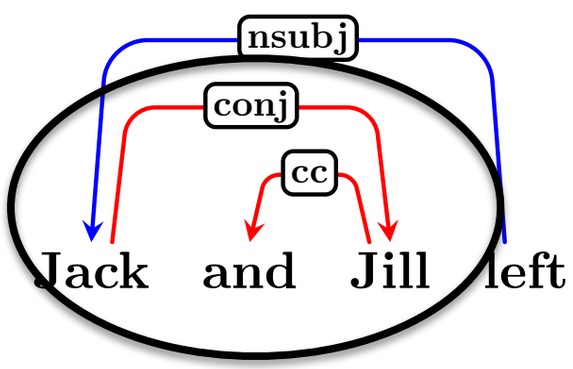


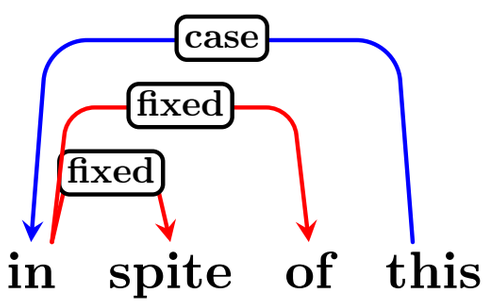
UD Syntax

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



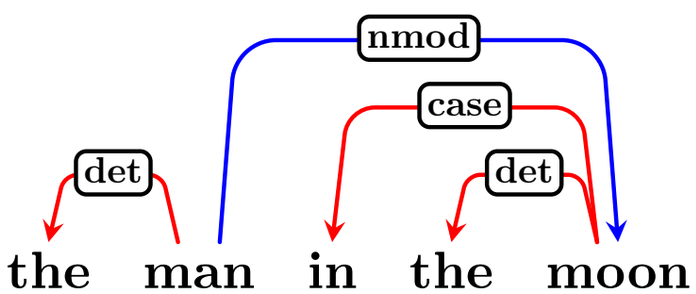


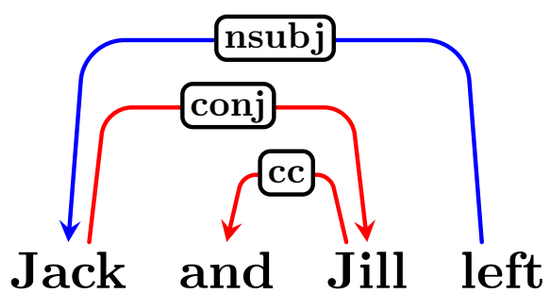


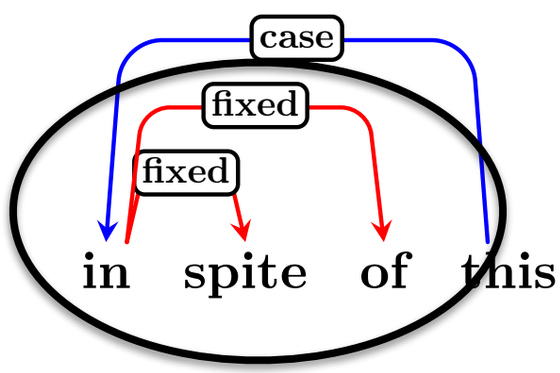
UD Syntax

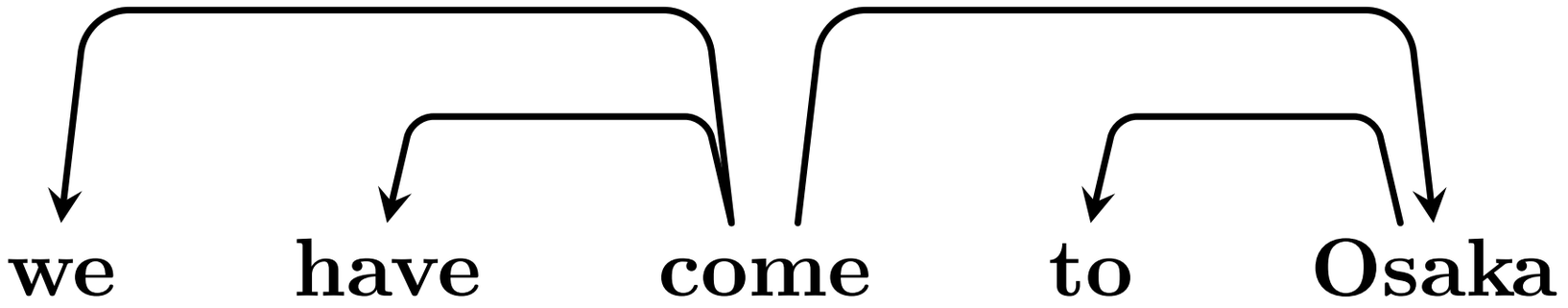
Other relations encoded in labels – not tree structure

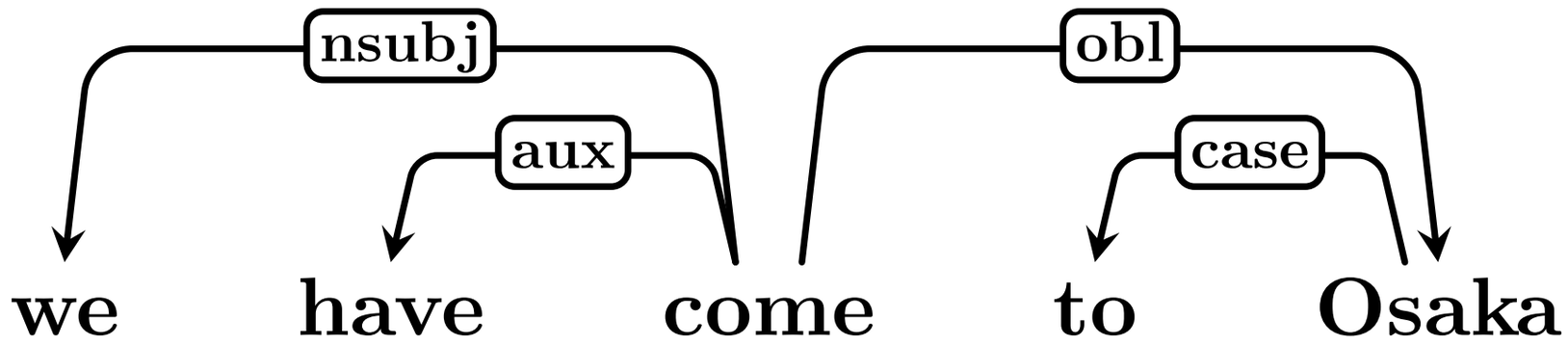
- Functional relations link functional heads to lexical heads
- Coordination relations link equivalent heads/dependents
- Multiword relations link elements of lexicalized expressions

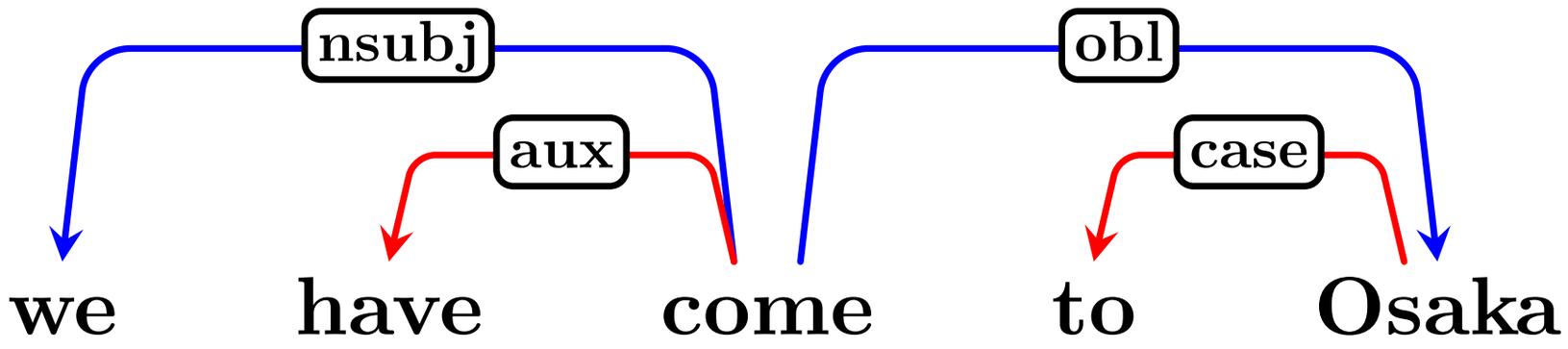


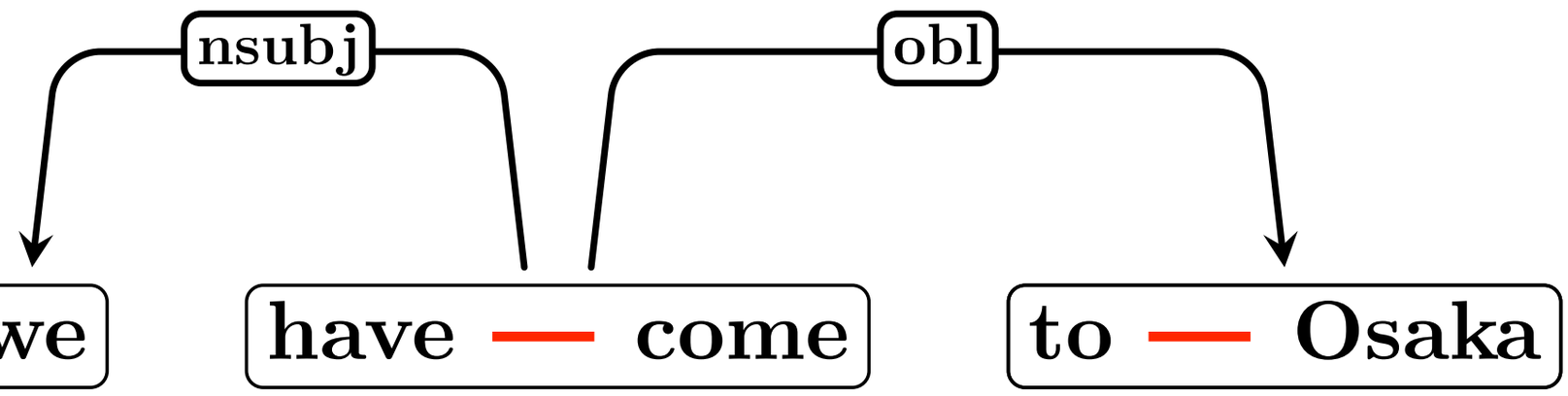




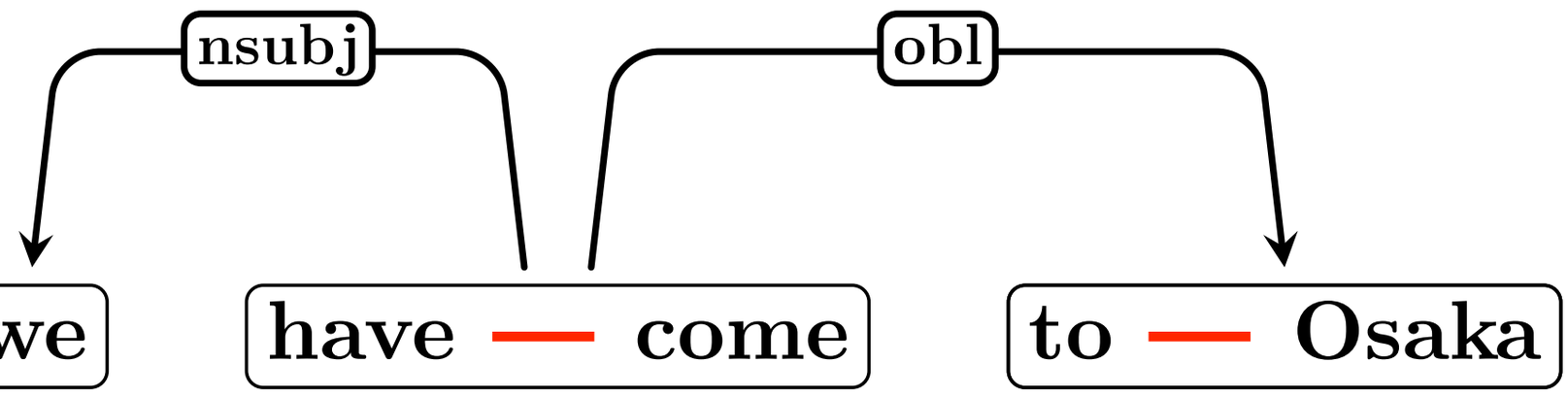

we
have
come
to
Osaka


nsubj
obl
aux
case
we
have
come
to
Osaka

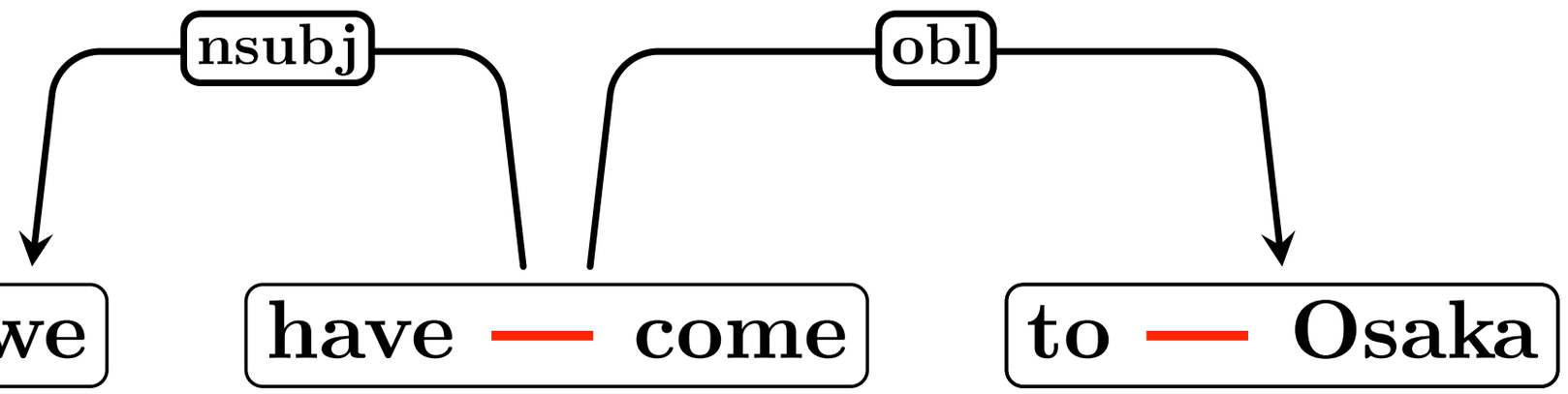

nsubj
obl
aux
case
we
have
come
to
Osaka



dependency	nucleus

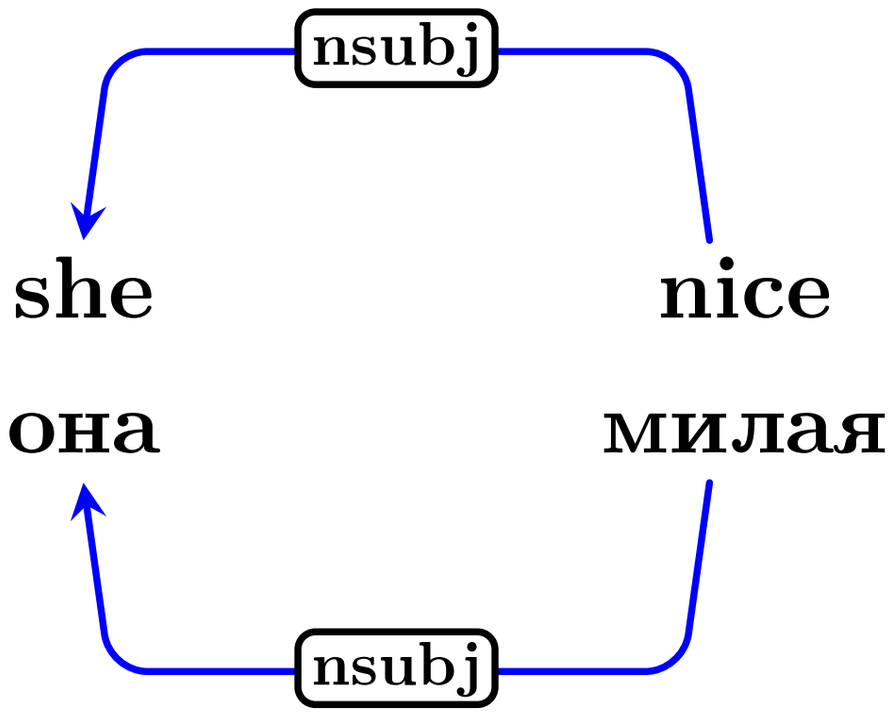


dependency	nucleus
karaka	vibhakti

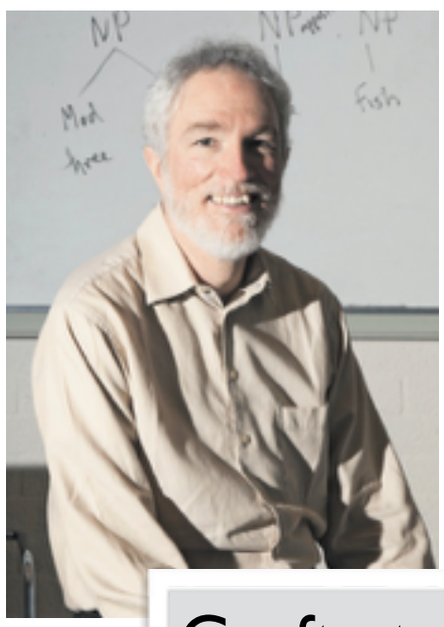


dependency	nucleus
karaka	vibhakti
kakariuke	bunsetsu

Linguistic Typology

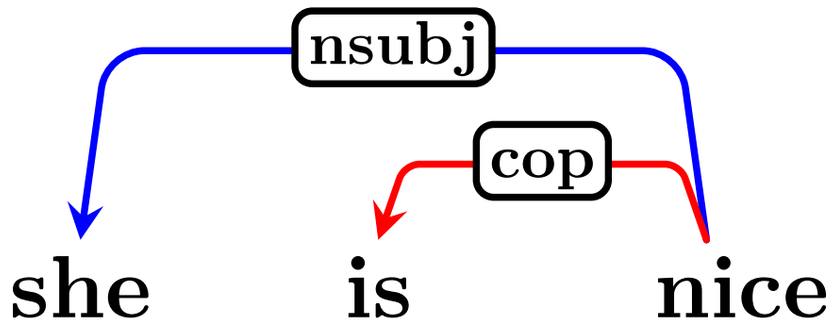


nonverbal predication



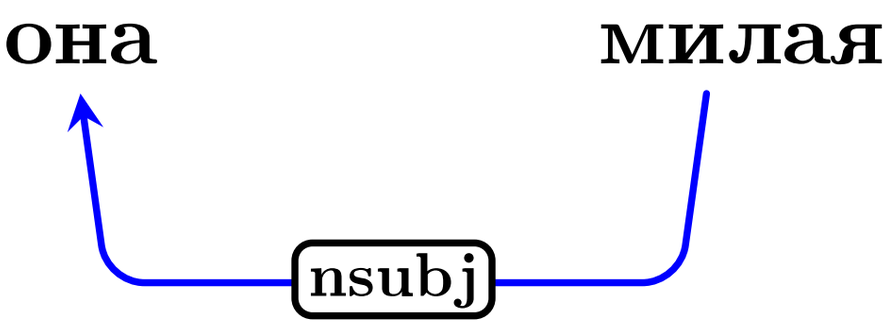
Croft et al. (2017) Linguistic Typology Meets Universal Dependencies

Linguistic Typology

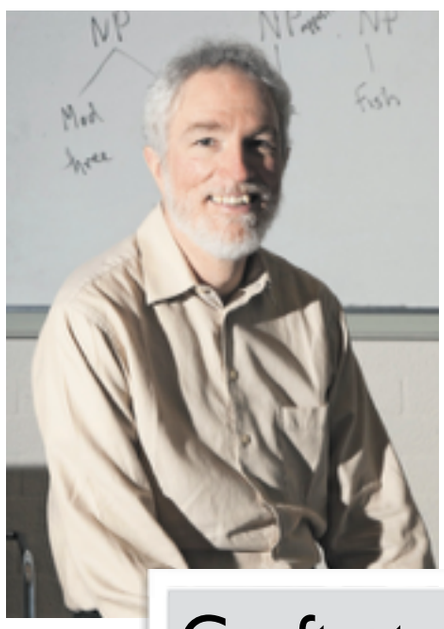


copula strategy

nonverbal predication

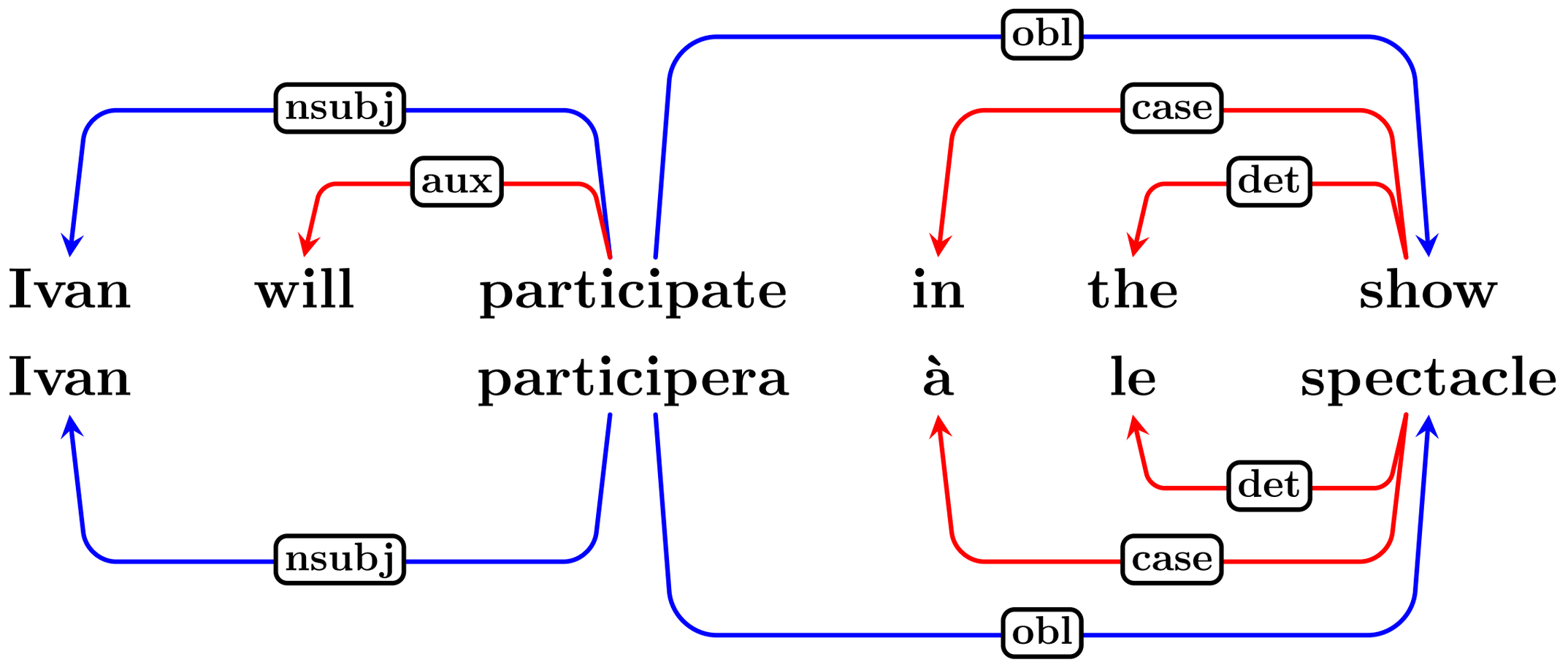


null strategy

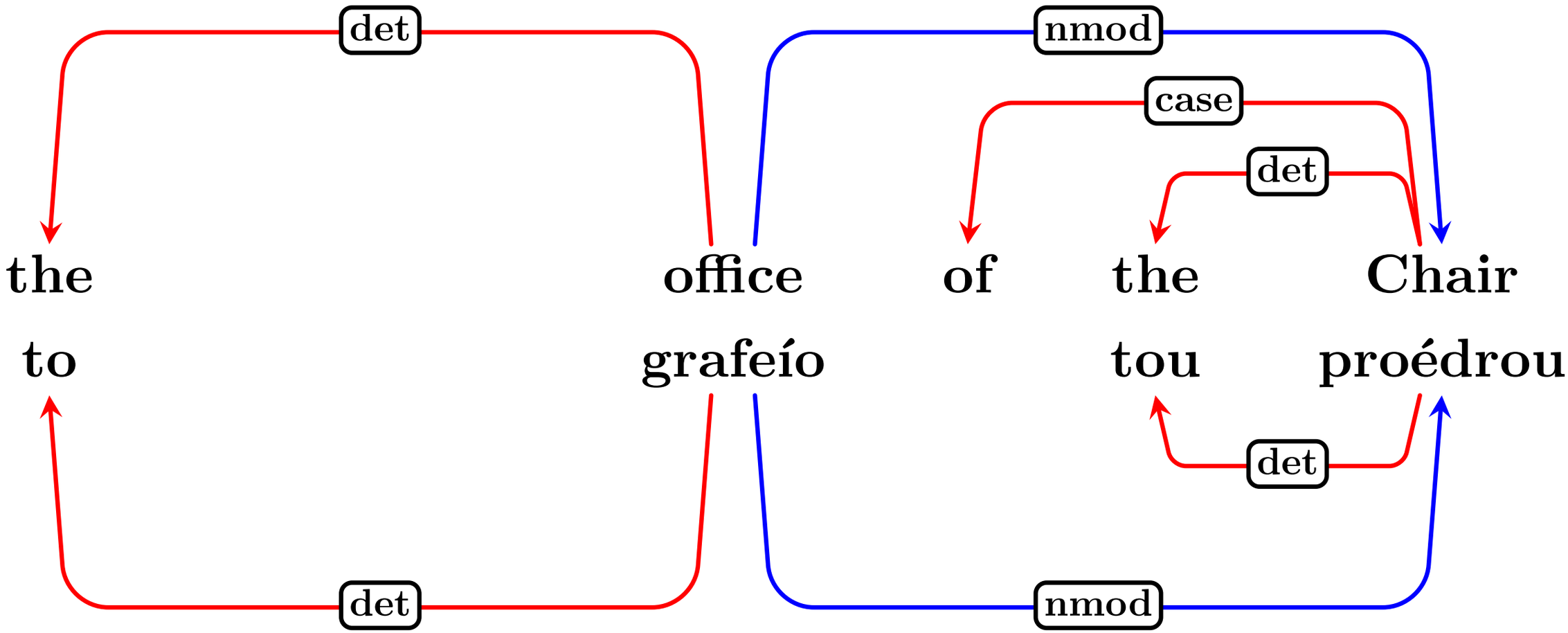


Croft et al. (2017) Linguistic Typology Meets Universal Dependencies

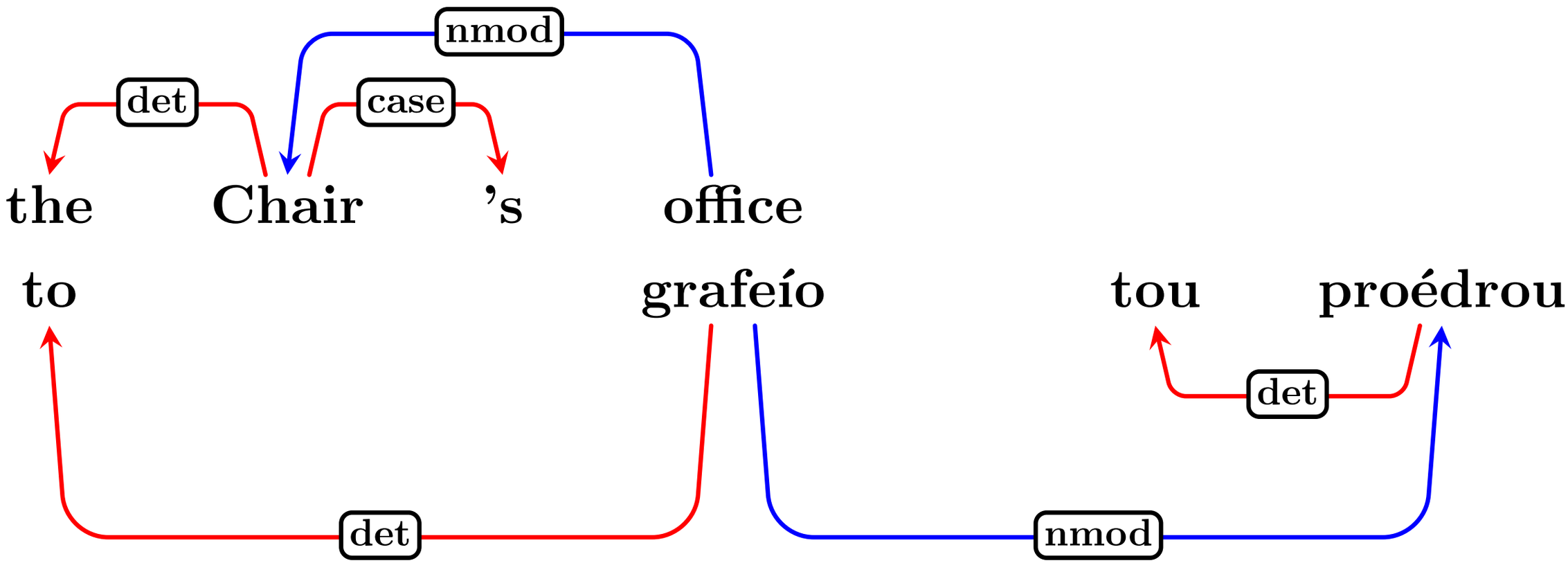
Verb Groups


nsubj
aux
obl
case
det
Ivan will participate in the show
Ivan participera à le spectacle
det
nsubj
case
obl

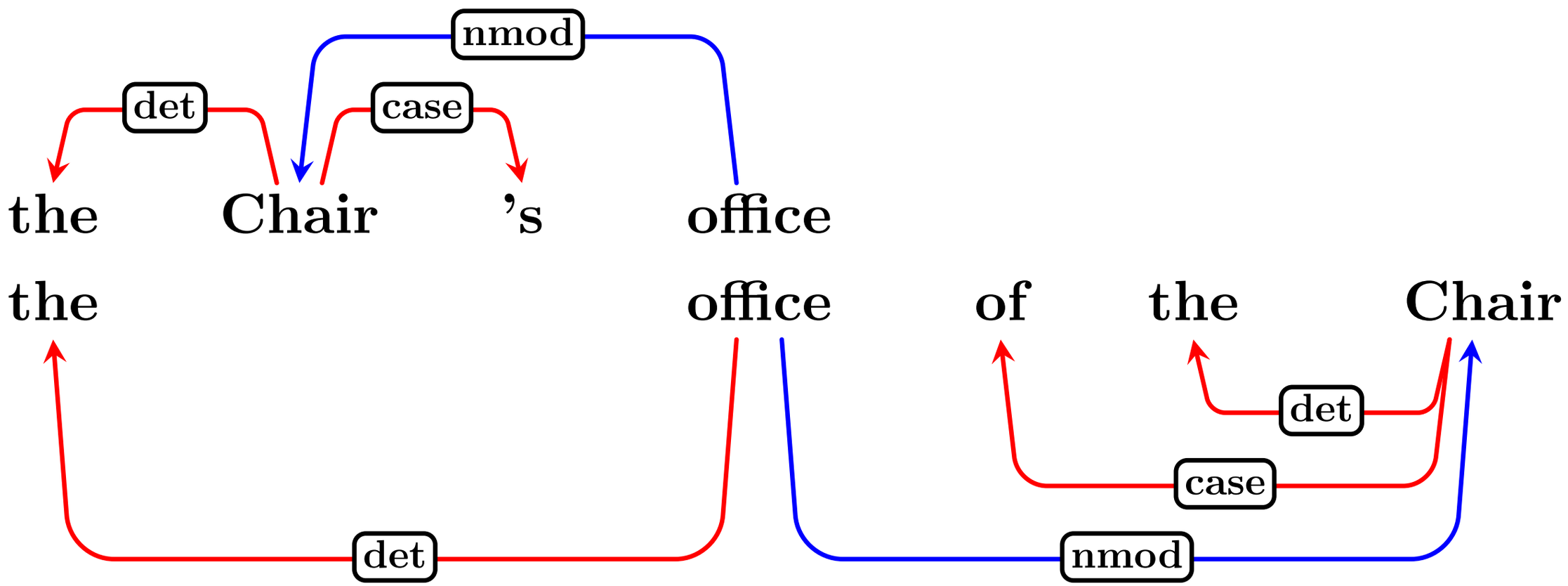
Adpositions and Case



Adpositions and Case



Adpositions and Case



Syntax or Semantics?

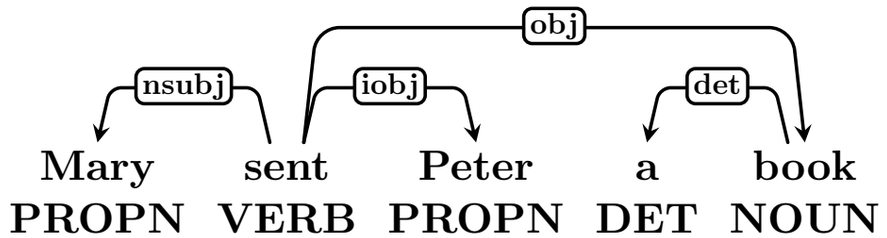
Are not UD dependencies semantic rather than syntactic?

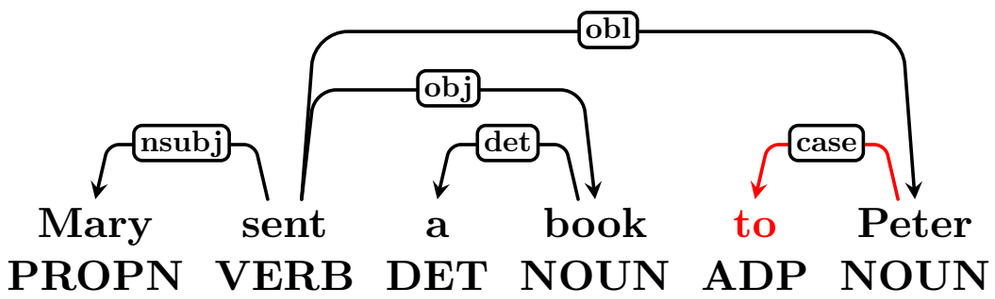
- Dependencies capture predicate-argument relations
- Dependencies do not (directly) capture agreement and government

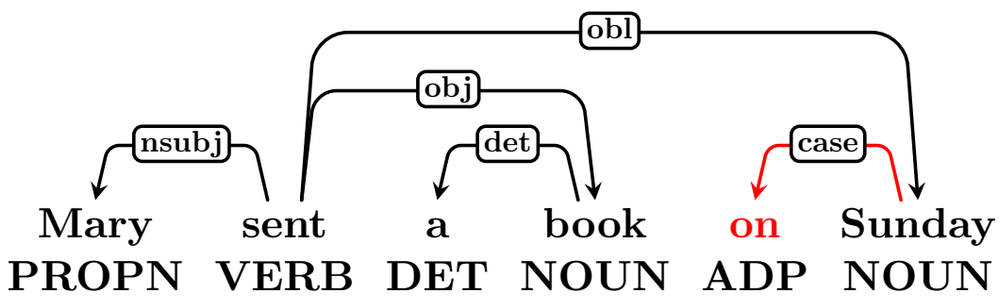
A functional view of syntax:

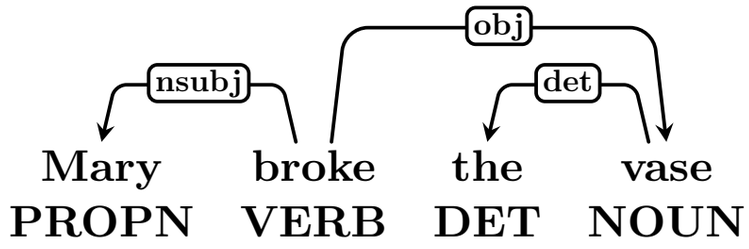
- UD is based on grammatical functions, not semantic roles
- UD models relations that are encoded morphosyntactically
- UD does not model semantic role alternations

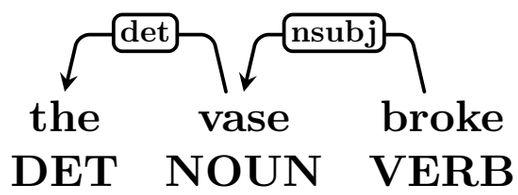
Grammatical Functions



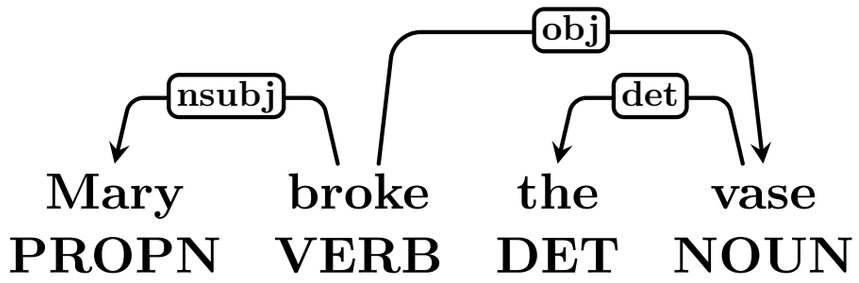


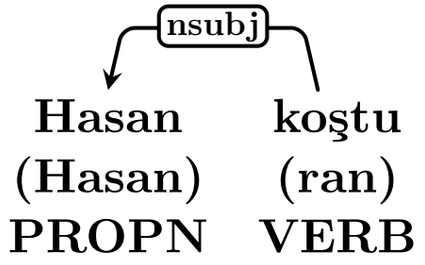


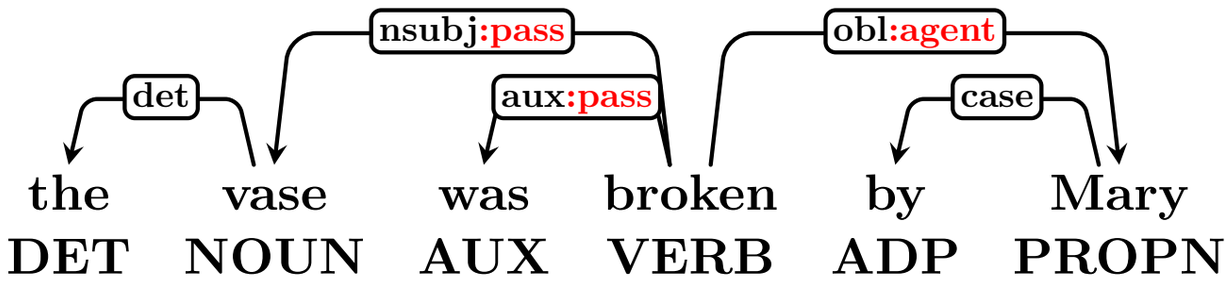


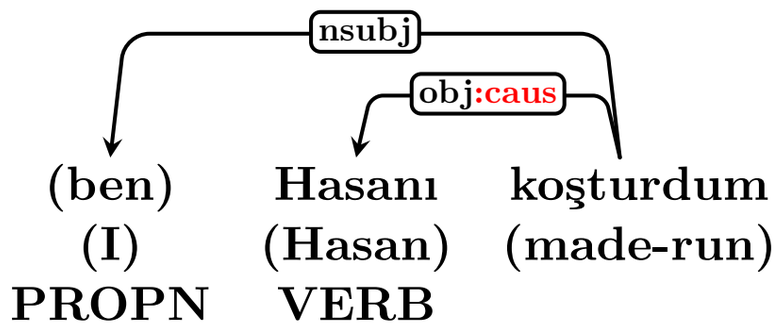


Valency-Changing Operations



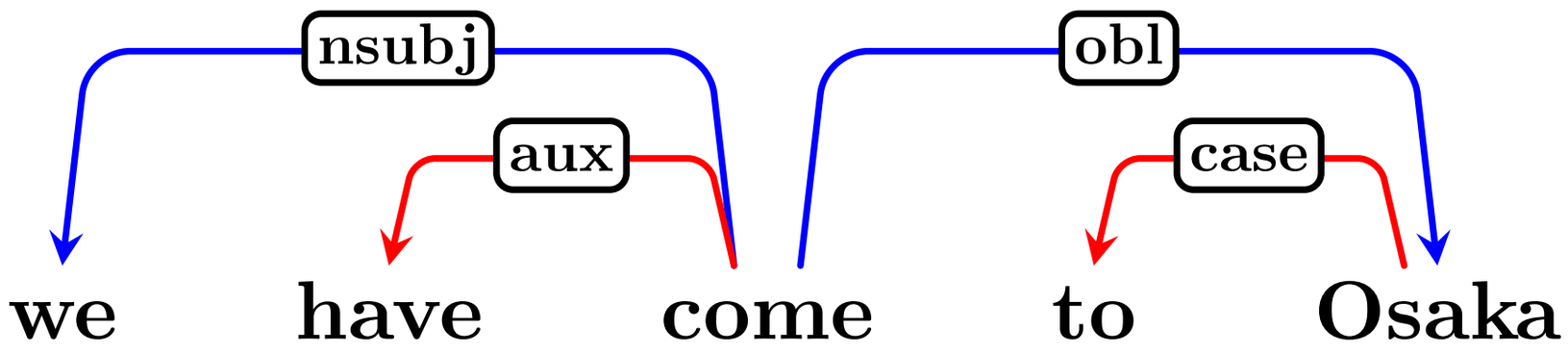






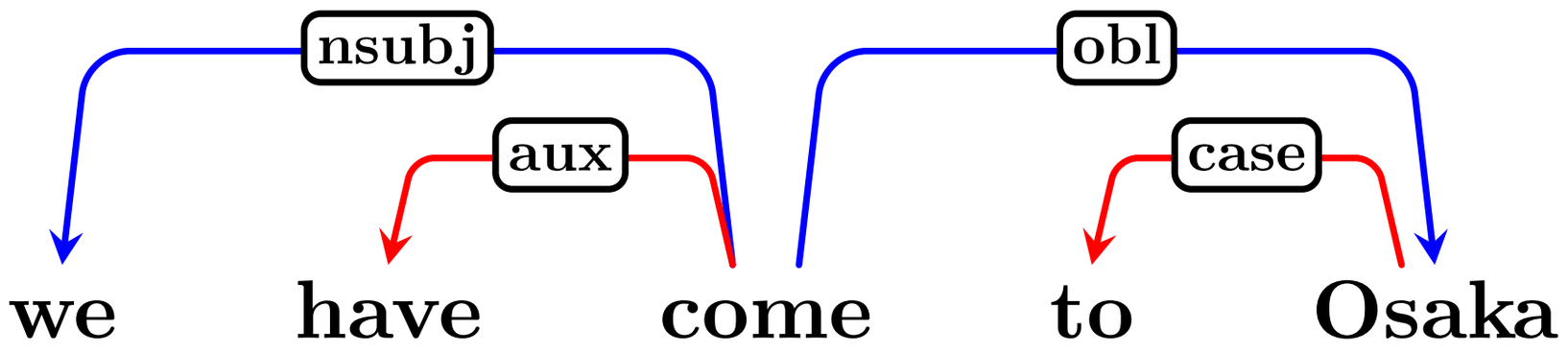
UD Representations

- Mono-stratal but multi-relational representations
- Grammatical functions take priority
- Both lexical and functional heads can be extracted



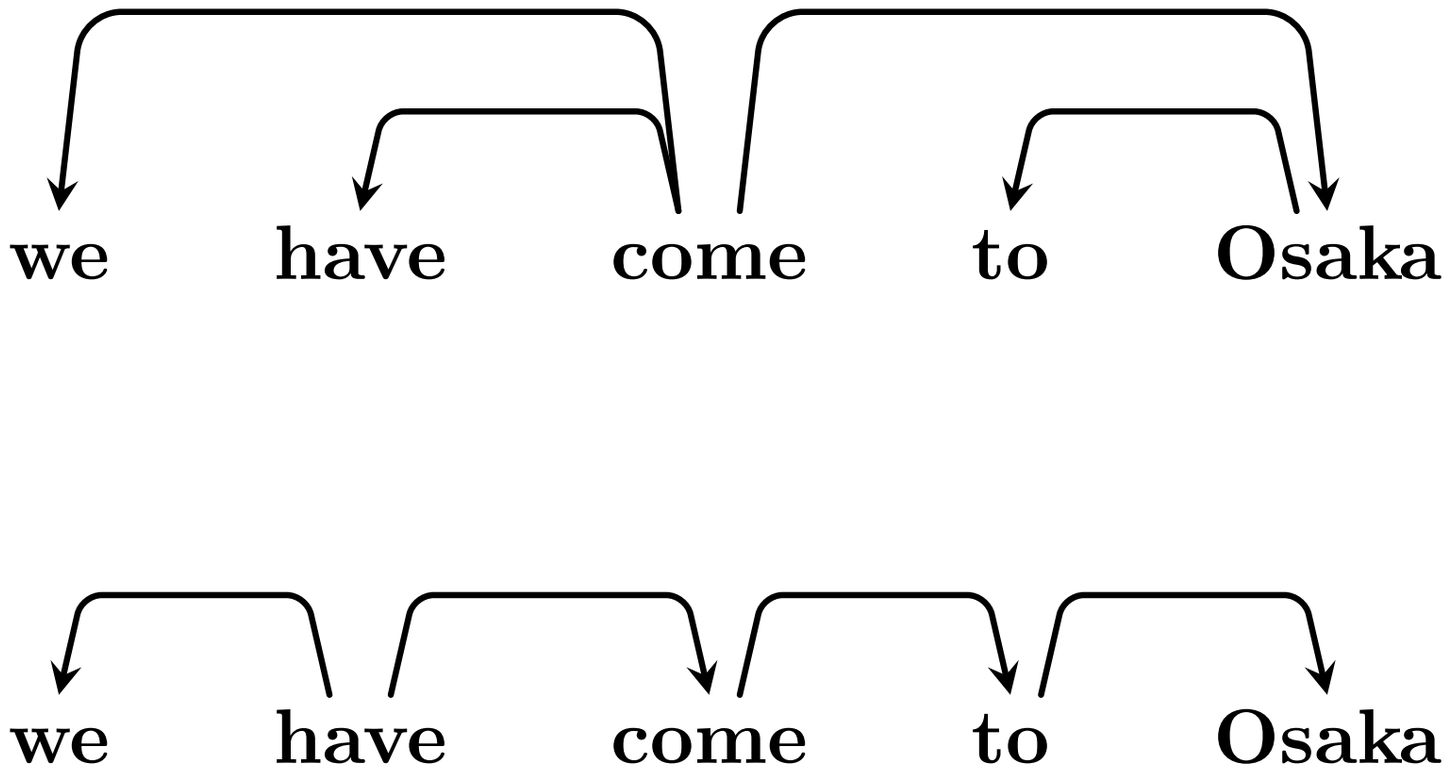
UD Representations

- Mono-stratal but multi-relational representations
- Grammatical functions take priority
- Both lexical and functional heads can be extracted



But you need to be aware of this!

Head-Initial or Head-Final?



Data

UD Treebanks

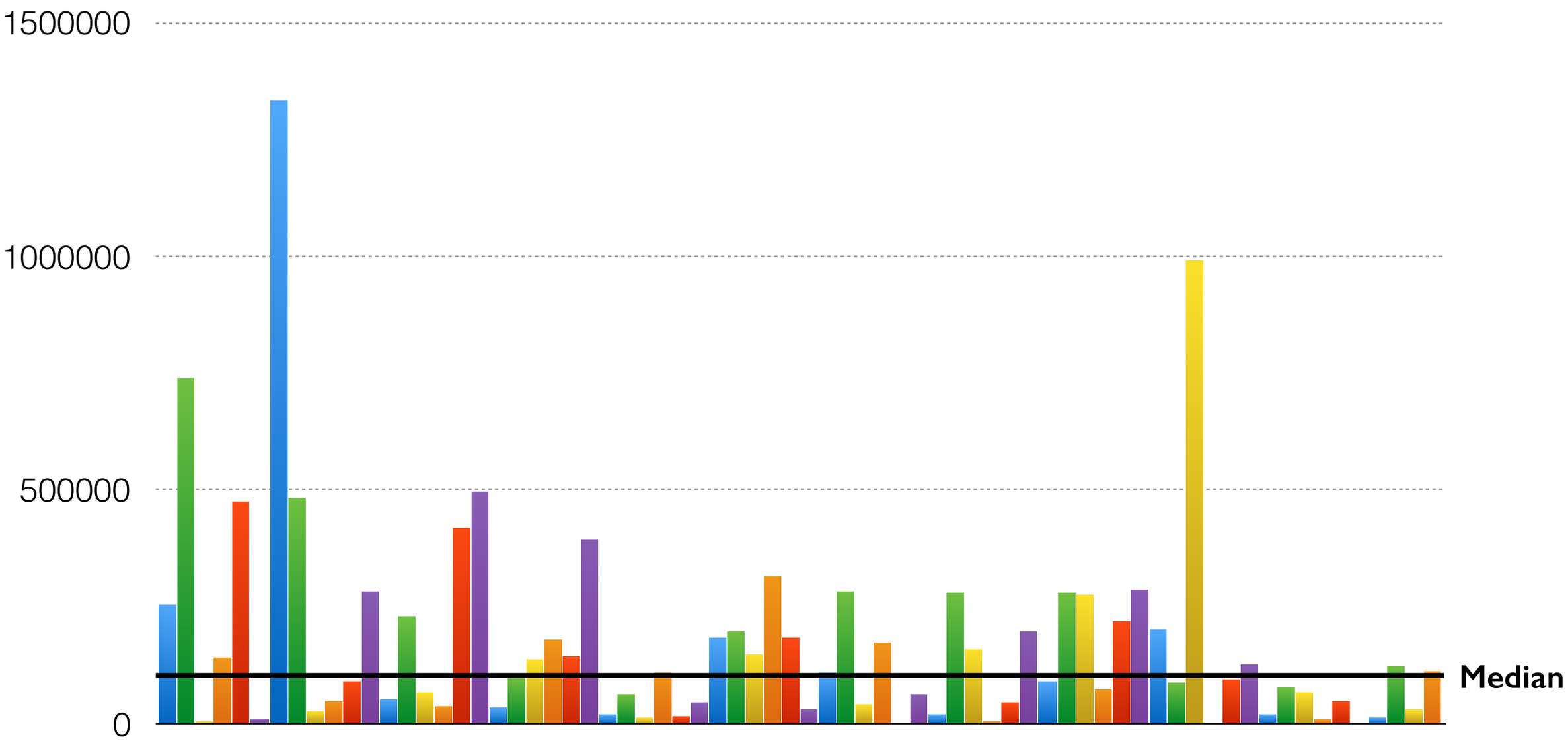
Ancient Greek	182K						
Ancient Greek-PROIEL	198K						
Arabic	217K						
Arabic-NYUAD	629K						
Basque	97K						
Belarusian	6K						
Bulgarian	140K						
Catalan	472K						
Chinese	111K						
Coptic	3K						
Croatian	183K						
Czech	1,330K						
Czech-CAC	482K						
Czech-CLTT	26K						
Danish	94K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

50 languages

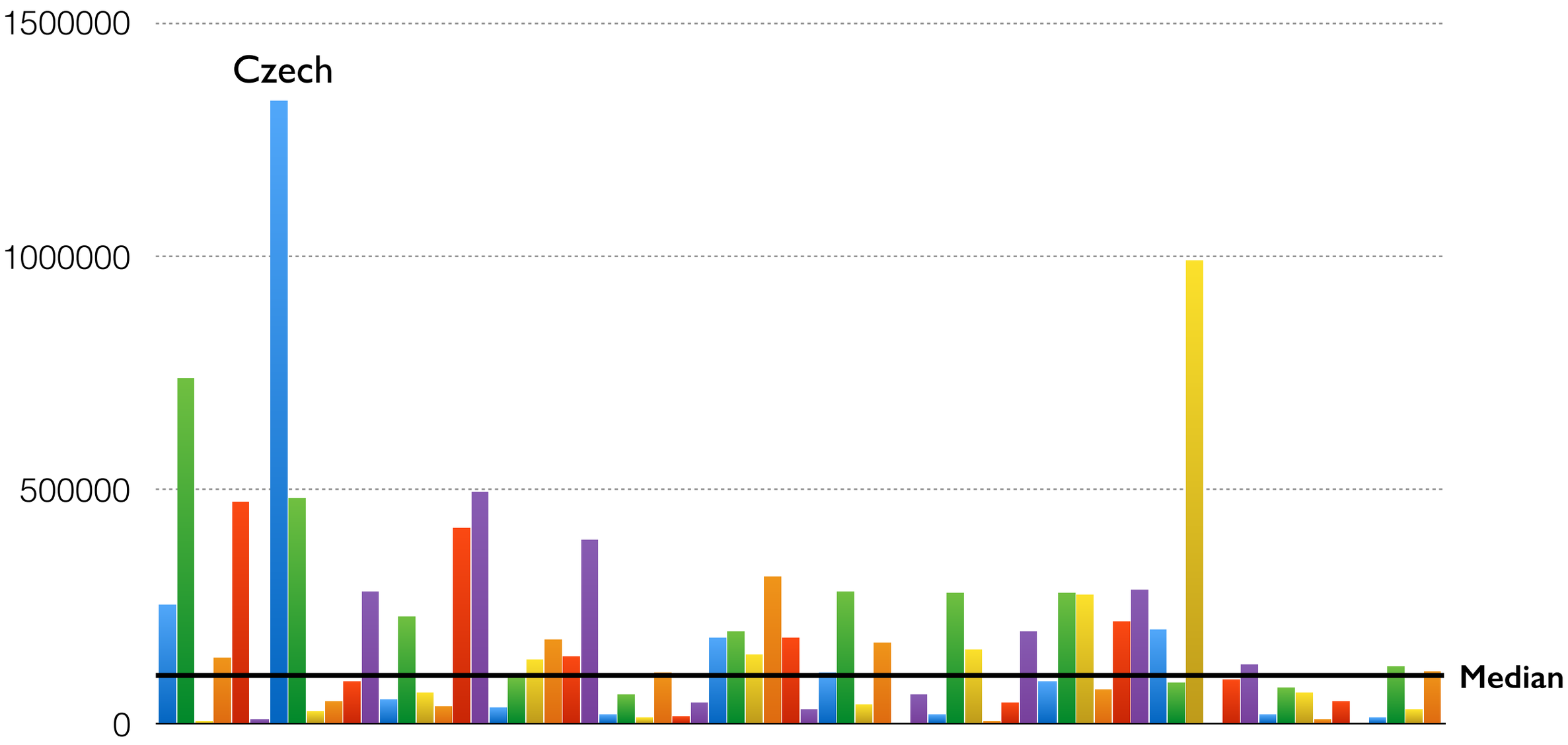
70 treebanks

http://universaldependencies.org

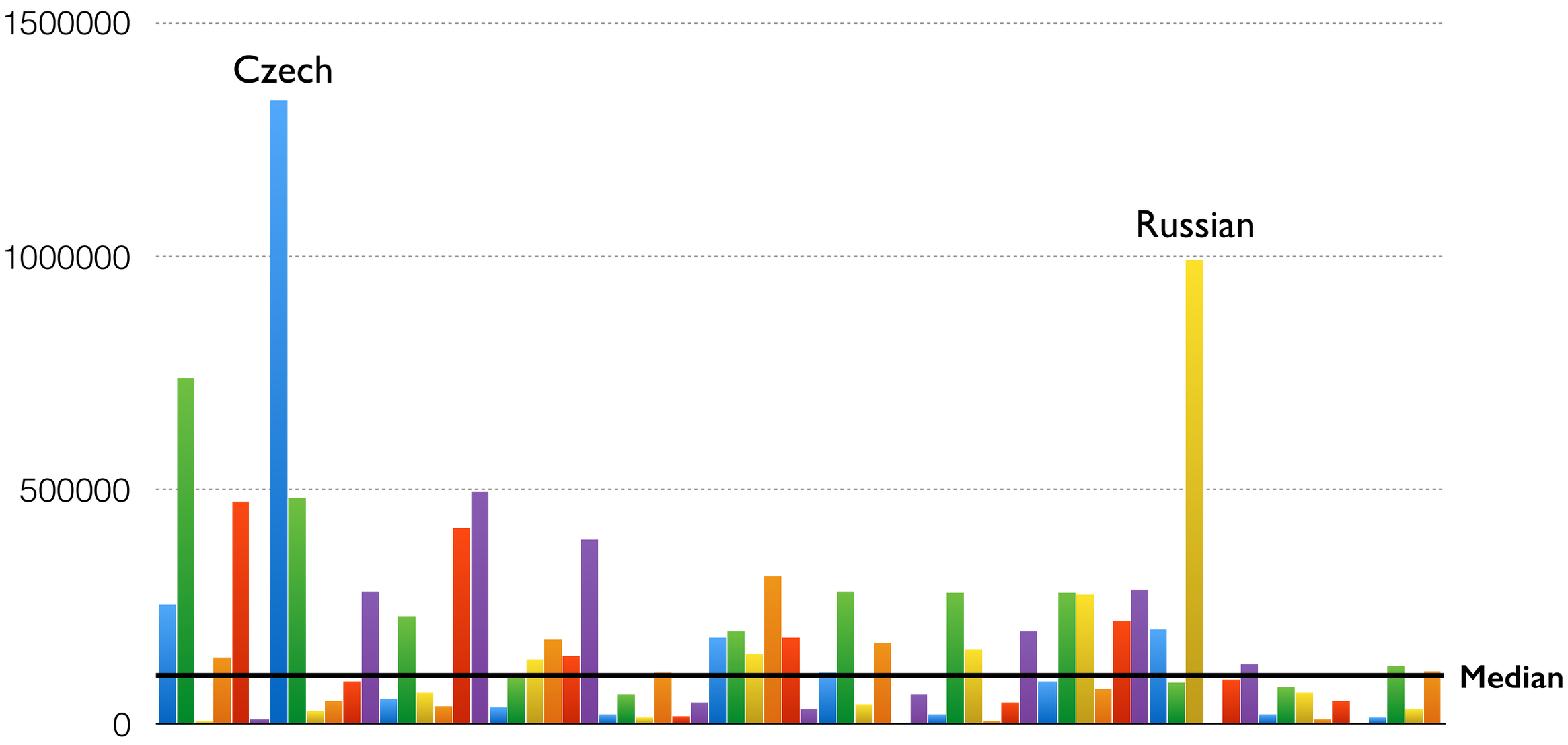
Treebank Size



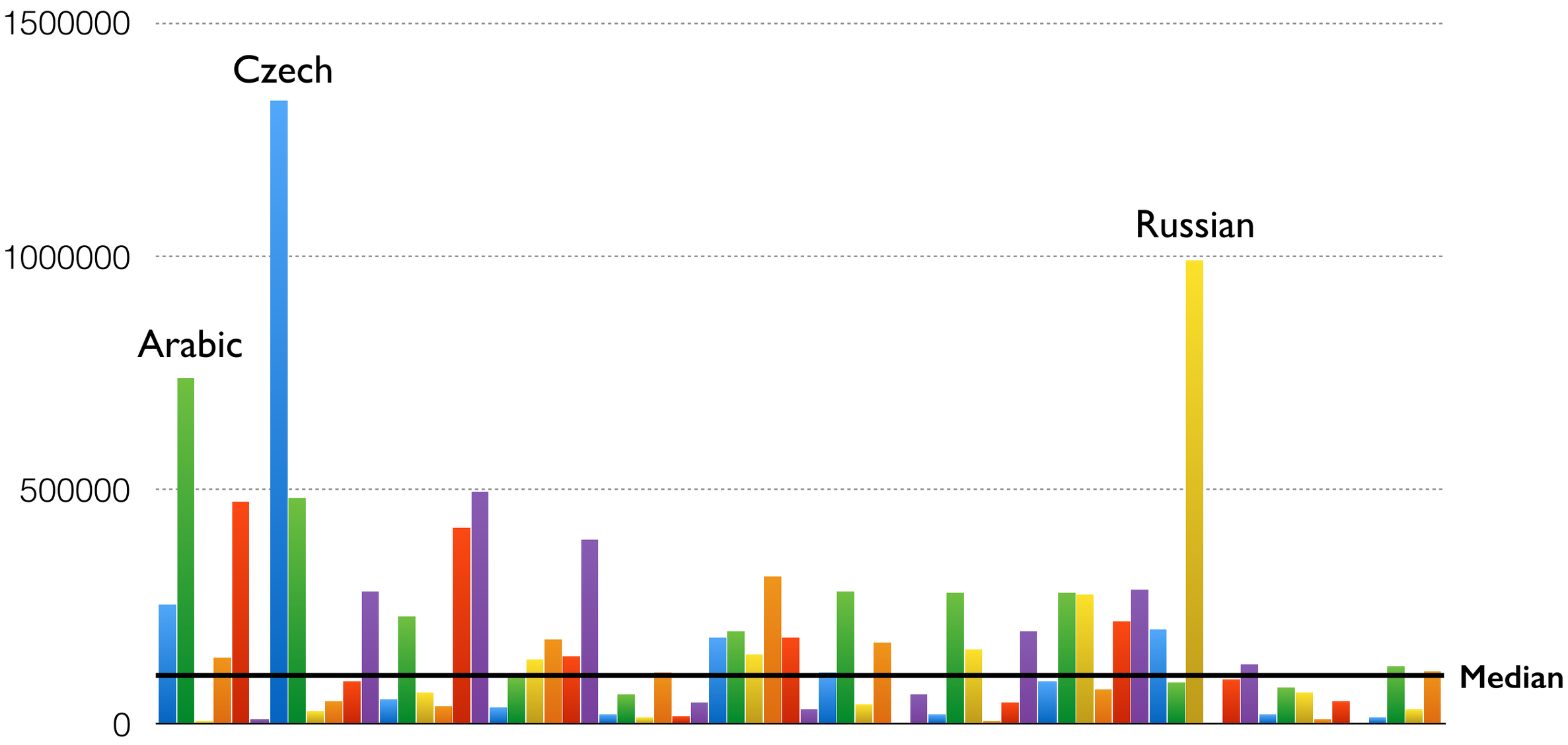
Treebank Size



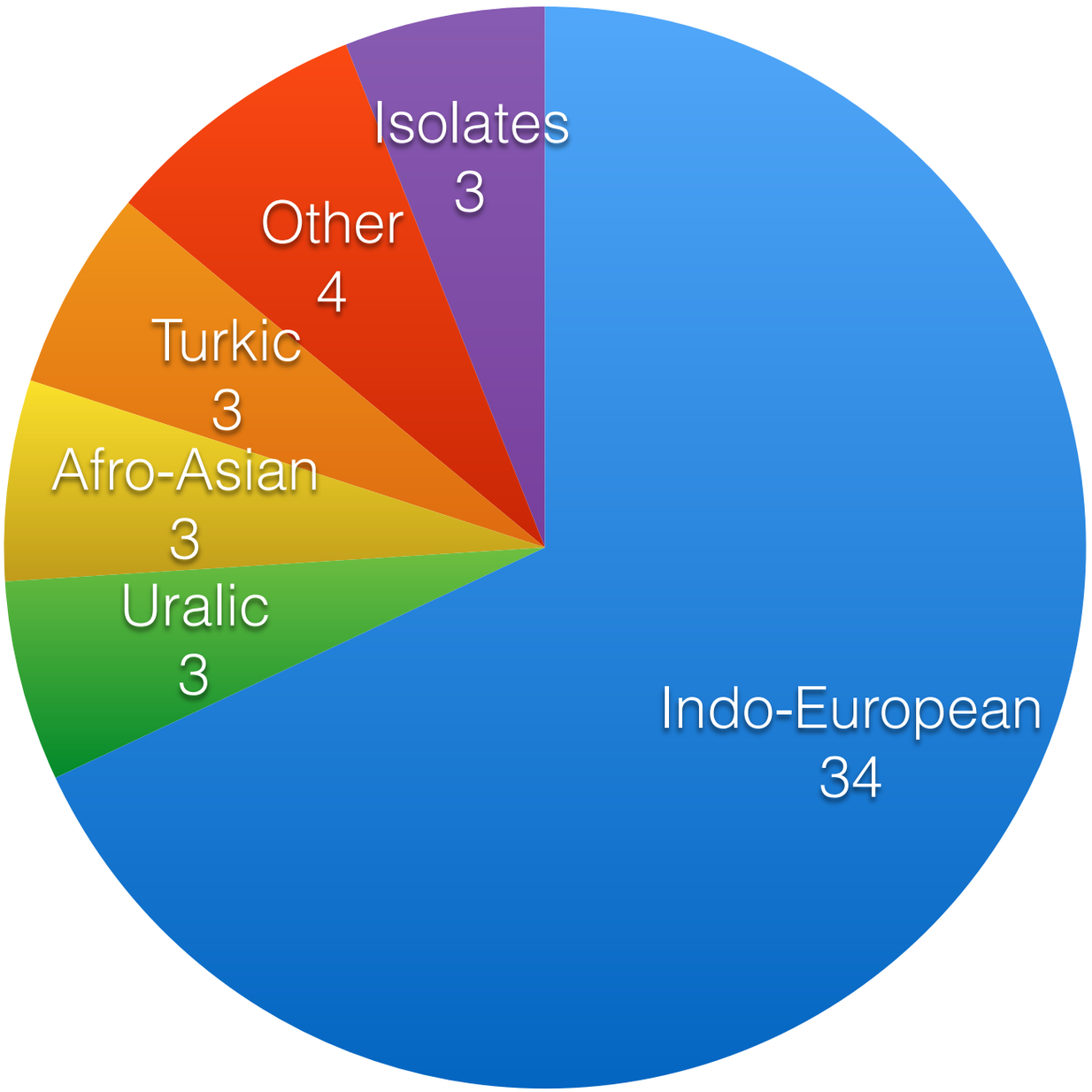
Treebank Size



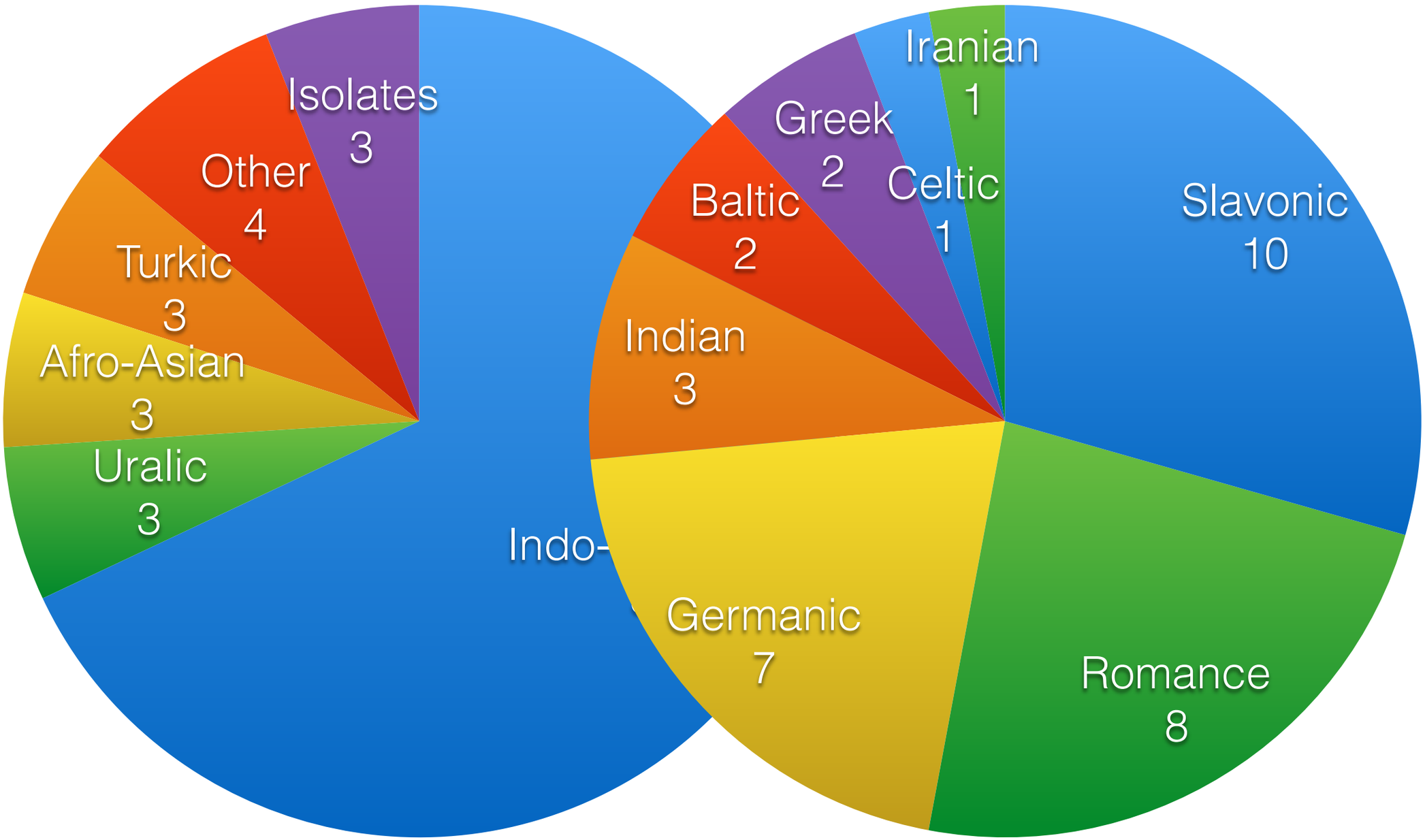
Treebank Size



Language Family

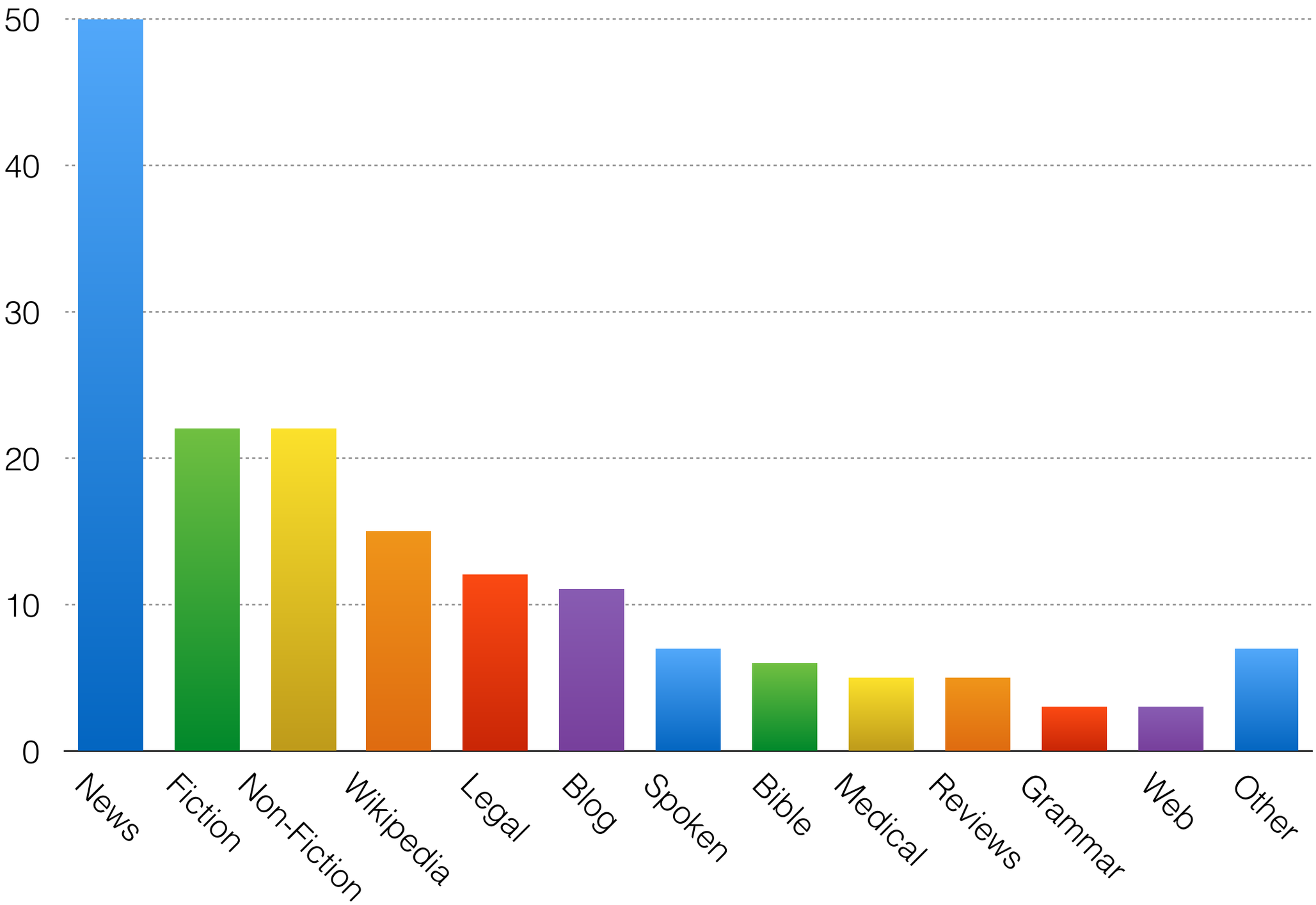


Language Family



Indo-European

Genre


50
40
30
20
10
0
News
Fiction
Non-Fiction
Wikipedia
Legal
Blog
Spoken
Bible
Medical
Reviews
Grammar
Web
Other

Studies

Quantifying Word Order Freedom in Dependency Corpora

Richard Futrell, Kyle Mahowald, and Edward Gibson
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
{futrell, kylemaho, egibson}@mit.edu

- Word order freedom studied in UD treebanks
- Conditional entropy of order given dependencies
- Test hypotheses about case and word order freedom

Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pages 91–100,
Uppsala, Sweden, August 24–26 2015.

Quantifying Word Order Freedom in Dependency Corpora

Richard Futrell, Kyle Mahowald, and Edward Gibson
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
{futrell, kylemaho, egibson}@mit.edu

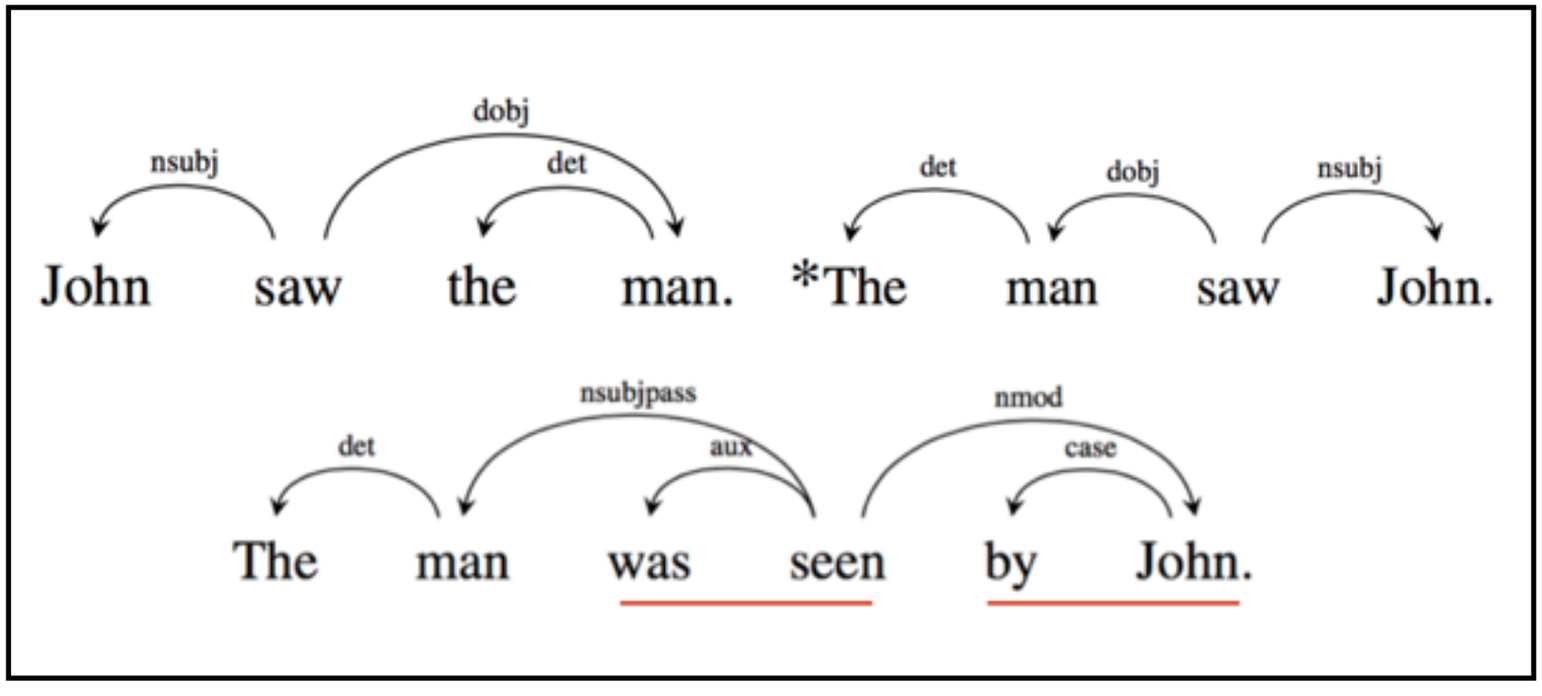
- Word order freedom studied in UD treebanks
- Conditional entropy of order given dependencies
- Test hypotheses about case and word order freedom

Thanks to Richard for sharing slides!

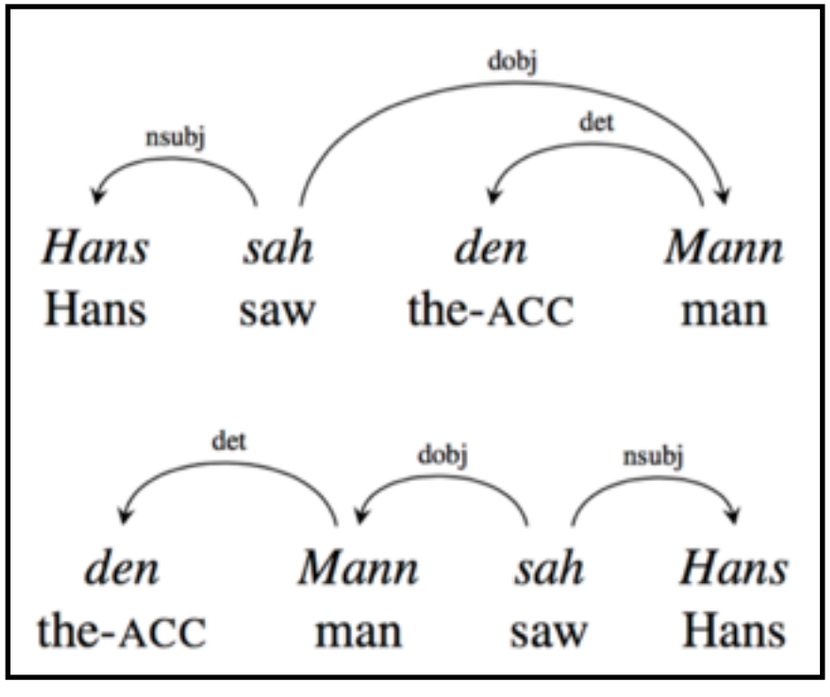
Word Order Freedom

We define **word order freedom** as the extent to which the same word or phrase in the same form can appear in multiple positions in a sentence while retaining the same propositional meaning.

English

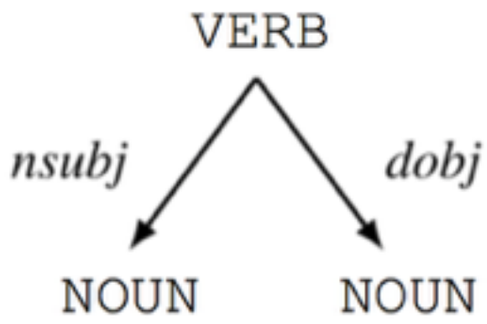


German



Conditional Entropy

Word order freedom is roughly *variability in linear order of words conditional on unordered dependency trees* (with relation type labels).



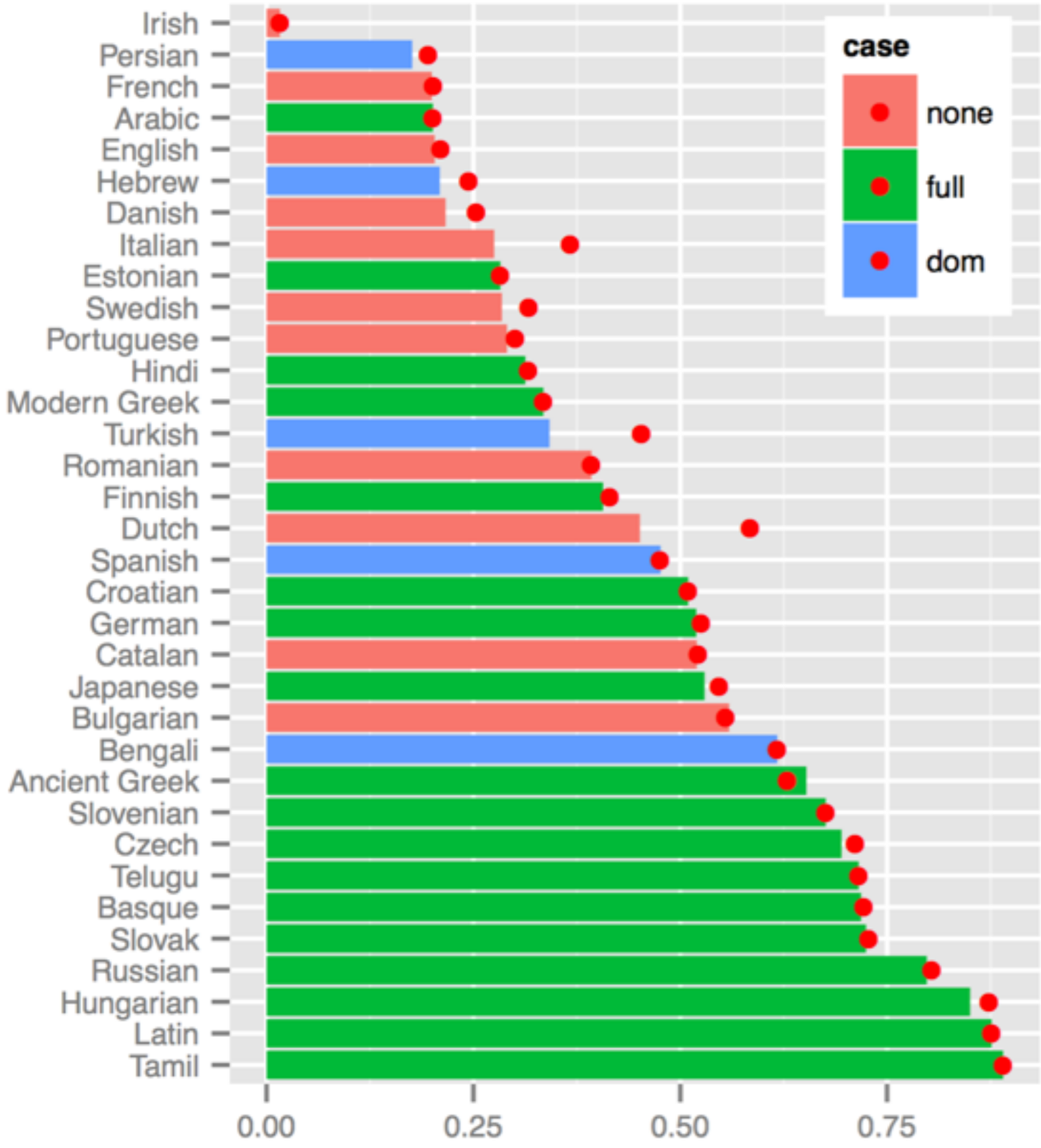
NOUN/*nsubj* **VERB**/*head* **NOUN**/*dobj: 55*
NOUN/*dobj* **VERB**/*head* **NOUN**/*nsubj: 25*

Local delexicalized trees to avoid data sparsity

Word Order and Morphology

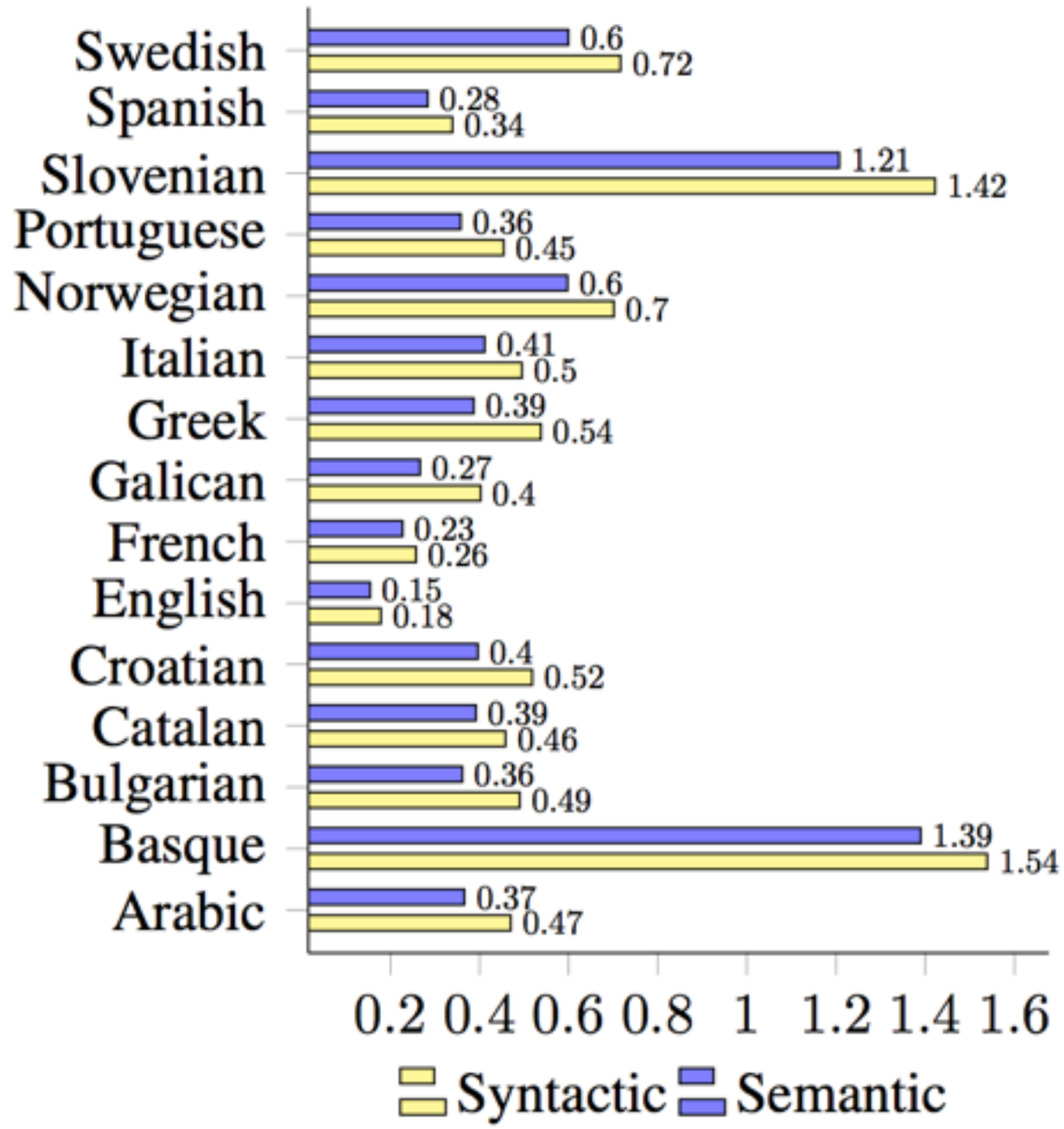
- The idea that there is a tradeoff between word order freedom and morphological marking is old (Sapir, 1921:66, Jakobson, 1936:28).
- A more recent typological claim (Kiparsky, 1997:461; cf. McFadden, 2003):
  - word order freedom ⇒ morphological marking
  - morphological marking ⇏ word order freedom
- But these generalizations are based on categorical descriptions and judgments.
  - Some quantitative work on word order freedom within languages (e.g. Pintzuk & Taylor, 2007), or between a few languages (e.g. Seo, 2001:92; Gulordava & Merlo, 2015)

Relation Order Entropy of Subject and Object



Semantics and Word Order

verb.contact
verb.competition
verb.consumption
verb.social
verb.body
verb.creation
verb.cognition
verb.motion
verb.perception
verb.weather
verb.change
verb.stative
verb.possession
verb.communication
verb.emotion



Word Order Typology through Multilingual Word Alignment

Robert Östling
Department of Linguistics
Stockholm University
SE-106 91 Stockholm, Sweden
robert@ling.su.se

- Word order typology based on Bible translations
- Massively parallel alignment and annotation projection
- UD tags and dependencies projected from English

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 205–211,
Beijing, China, July 26-31, 2015.

Word Order Typology through Multilingual Word Alignment

Robert Östling
Department of Linguistics
Stockholm University
SE-106 91 Stockholm, Sweden
robert@ling.su.se

- Word order typology based on Bible translations
- Massively parallel alignment and annotation projection
- UD tags and dependencies projected from English

Thanks to Robert for sharing slides!

Methodology

Word alignment of parallel texts:

- New Testament in 986 languages (1144 translations)
- Bayesian word alignment with interlingua (Östling, 2015)

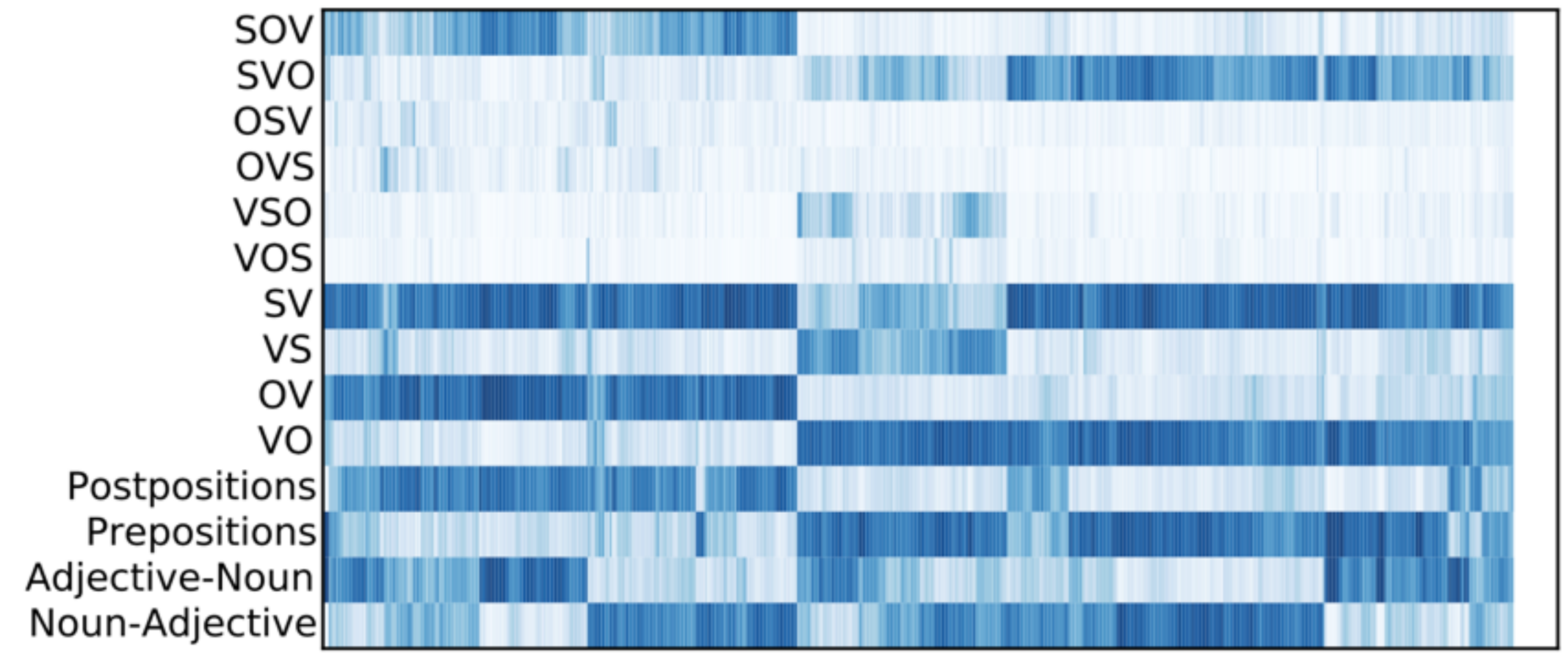
Annotation projection:

- UD tags and dependencies from 5 English translations
- Sparse, high-precision projection (80% of links must agree)

Word order statistics:

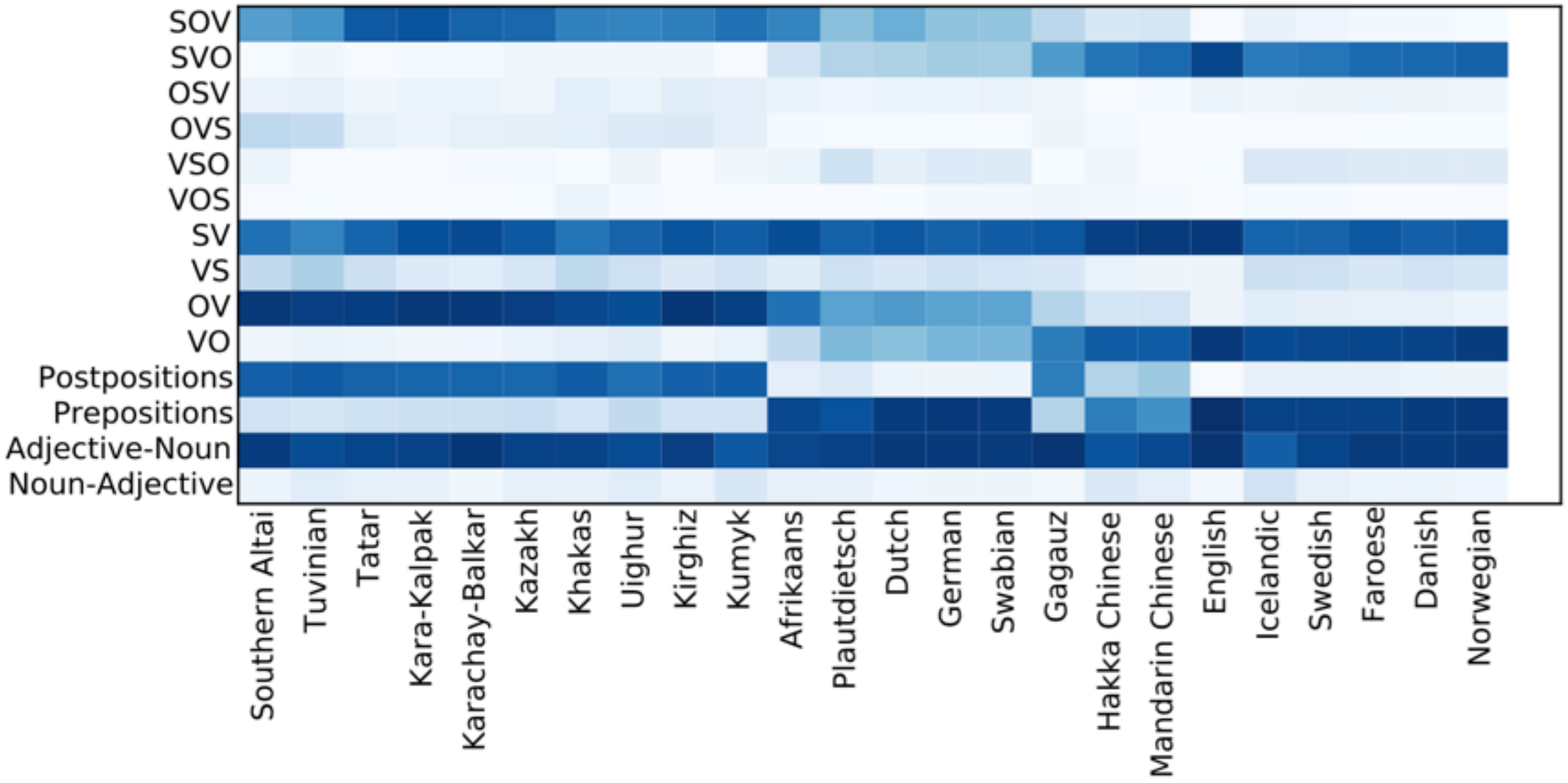
- Count frequency of different word order variants (tokens or types)
- Constructions: SOV – SV – OV – Adp-Noun – Adj-Noun

Results at a glance



The X axis consists of 986 individual languages,
clustered by their word order vector.

A closer look



Comparison to WALS

Feature	Languages	Types	Tokens	Most common
Subj,Obj,V	342	85.4%	85.7%	SOV: 43.3%
Subj,V	376	89.4%	90.4%	SV: 79.8%
Obj,V	387	96.4%	96.4%	VO: 54.8%
Adp, NP	329	94.8%	95.1%	Prep: 50.4%
Adj, Noun	334	85.9%	88.0%	AdjN: 68.9%

Comparison to WALS

Feature	Languages	Types	Tokens	Most common
Subj,Obj,V	342	85.4%	85.7%	SOV: 43.3%
Subj,V	376	89.4%	90.4%	SV: 79.8%
Obj,V	387	96.4%	96.4%	VO: 54.8%
Adp, NP	329	94.8%	95.1%	Prep: 50.4%
Adj, Noun	334	85.9%	88.0%	AdjN: 68.9%

New information for ≈ 600 languages

Conclusion

- Large collection of languages – but biased sample and mostly small corpora
- Cross-linguistically consistent grammatical annotation – but you have to know the quirks

http://universaldependencies.org