Nucleus Composition in Transition-Based Dependency Parsing

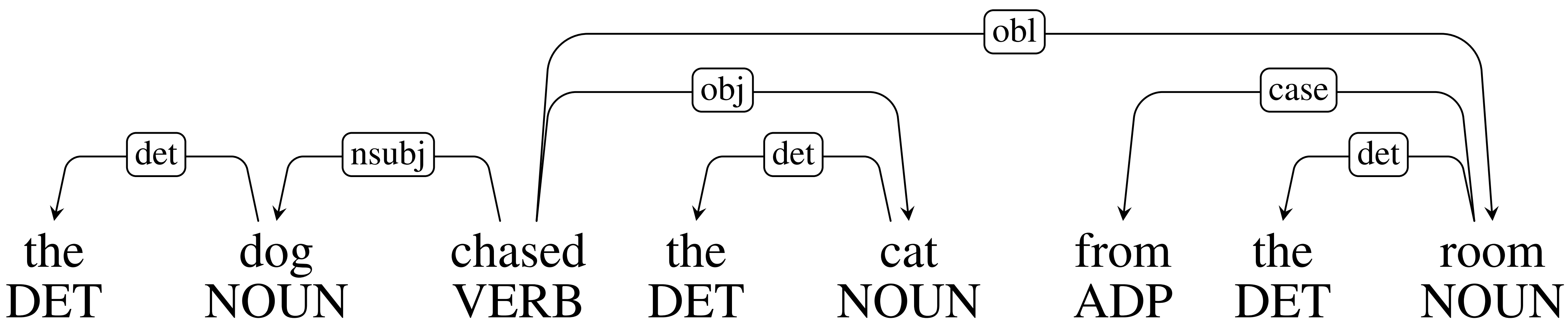
Joakim Nivre

RISE Research Institutes of Sweden

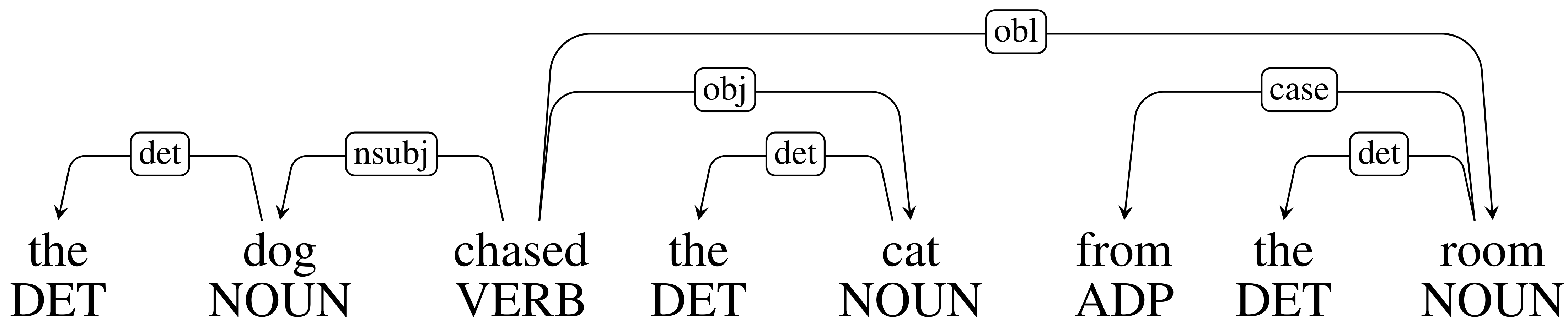
Uppsala University
Department of Linguistics and Philology

Joint work with Ali Basirat, Luise Dürlich and Adam Moss

Dependency Parsing

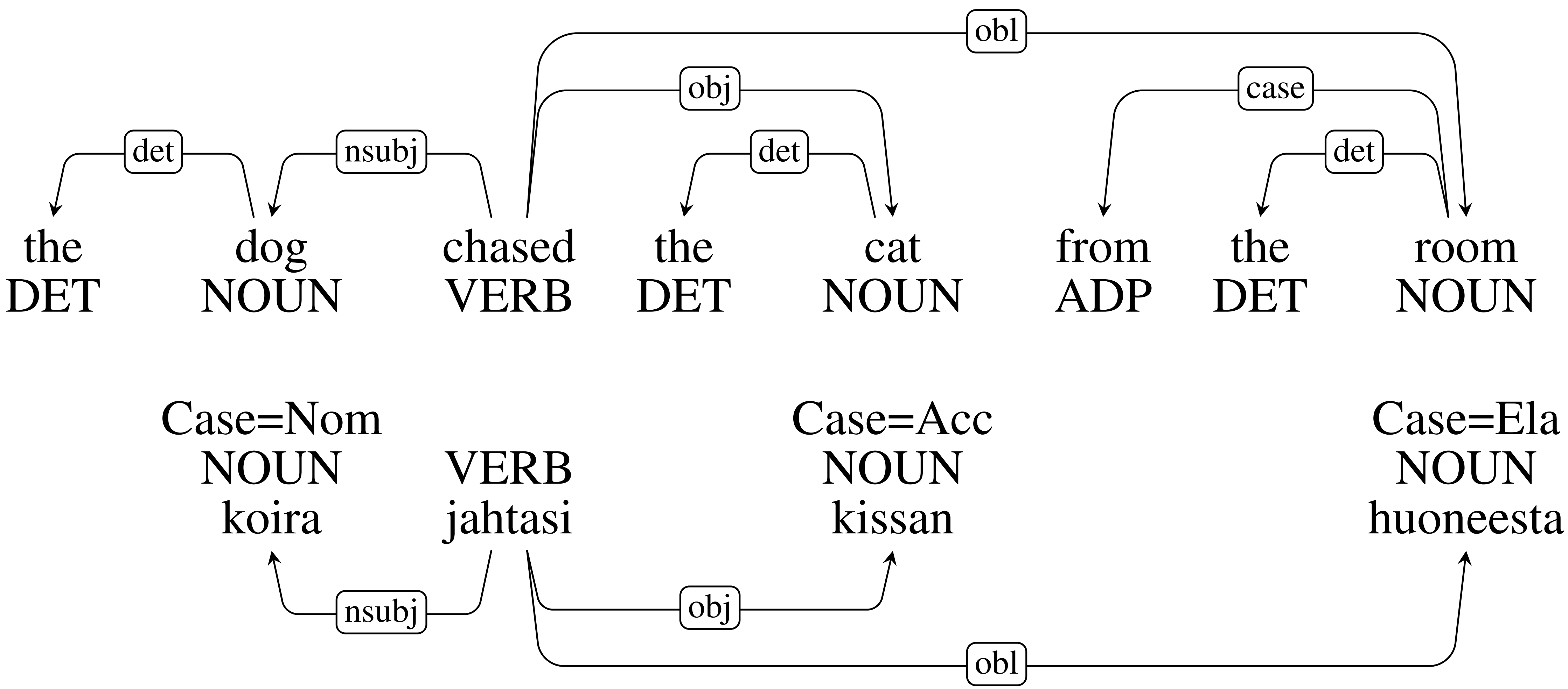


Dependency Parsing



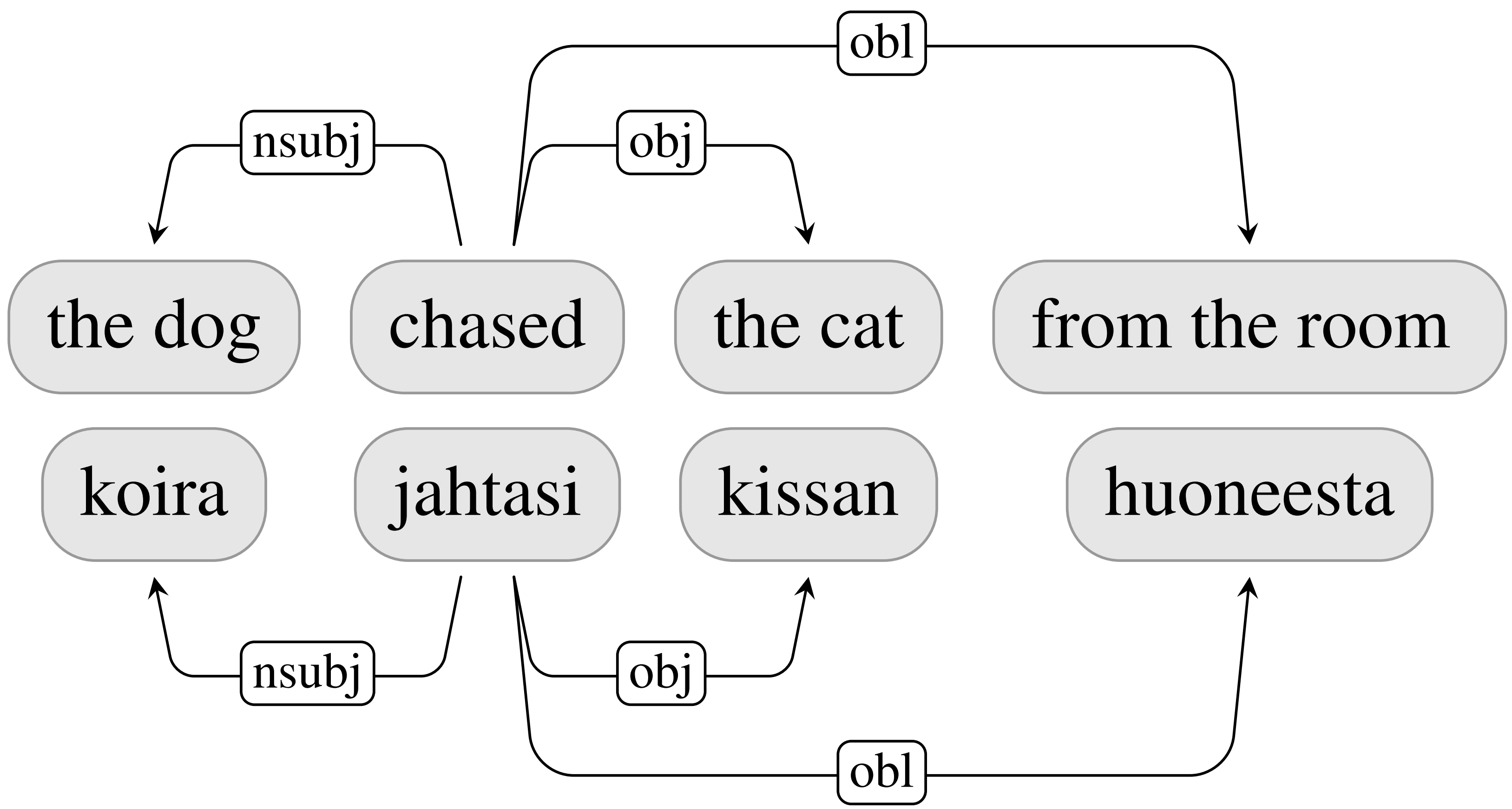
elementary syntactic unit = word

Dependency Parsing



elementary syntactic unit = word

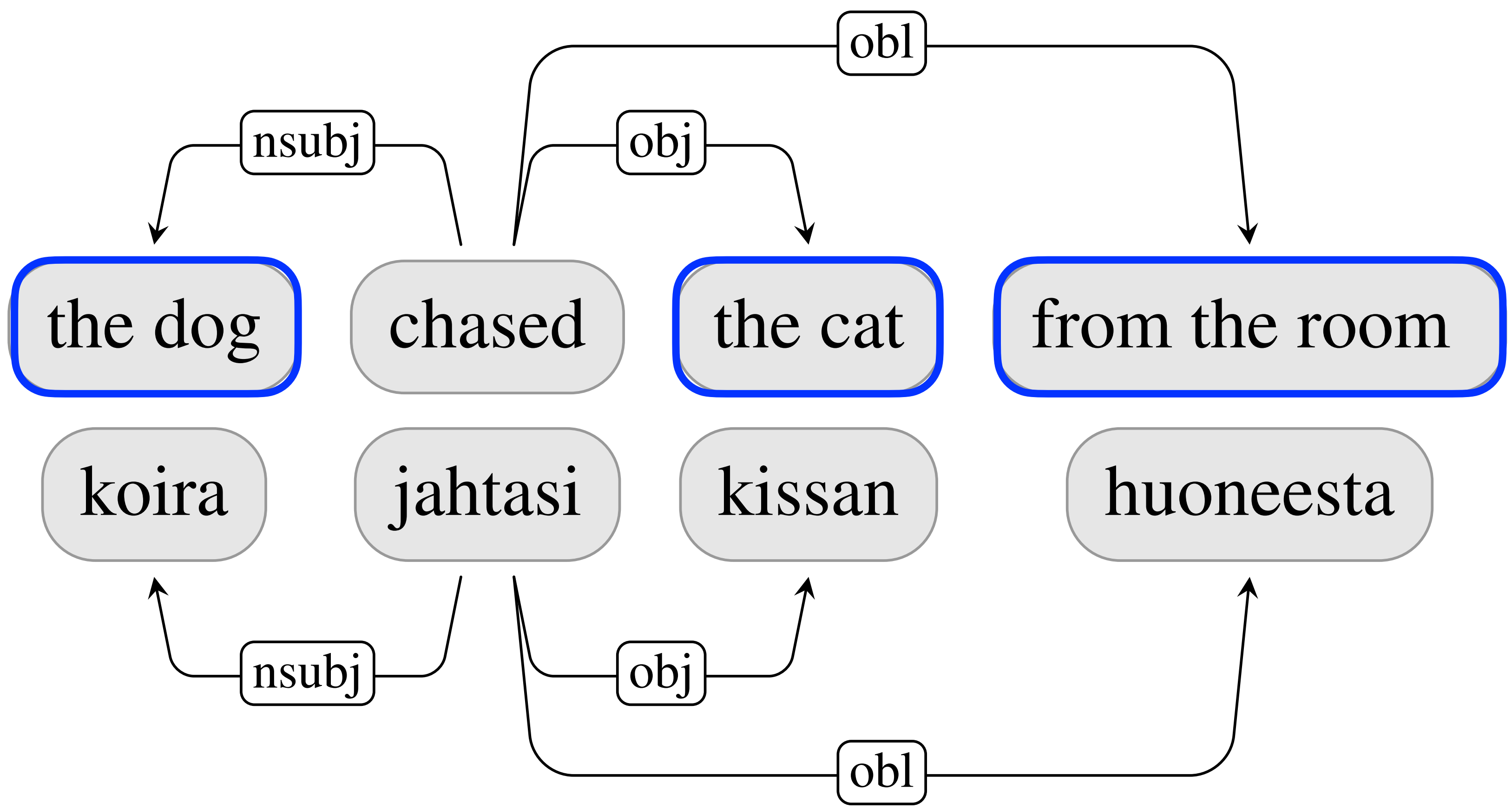
Dependency Parsing



elementary syntactic unit = nucleus

Lucien Tesnière (1959) *Éléments de syntaxe structurale*. Klincksieck

Dependency Parsing



elementary syntactic unit = nucleus

Lucien Tesnière (1959) *Éléments de syntaxe structurale*. Klincksieck

This Talk

- Define the notion of nucleus in Universal Dependencies
- Add nucleus representations to a dependency parser
- Analyse the impact of this technique across languages

This Talk

- Define the notion of nucleus in Universal Dependencies
- Add nucleus representations to a dependency parser
- Analyse the impact of this technique across languages

Ali Basirat and Joakim Nivre (2021) Syntactic Nuclei in Dependency Parsing – A Multilingual Exploration. In *Proceedings of EACL*, 1376–1387.

Joakim Nivre, Ali Basirat, Luise Dürlich and Adam Moss (2022) Nucleus Composition in Transition-Based Dependency Parsing. *Computational Linguistics* 48:4.

Historical Backdrop

Historical Backdrop

Towards an implementable dependency grammar

Timo Järvinen and **Pasi Tapanainen**
Research Unit for Multilingual Language Technology
P.O. Box 4, FIN-00014 University of Helsinki, Finland

Historical Backdrop

Towards an implementable dependency grammar

Timo Järvinen and **Pasi Tapanainen**
Research Unit for Multilingual Language
P.O. Box 4, FIN-00014 University of Helsi

A Statistical Theory of Dependency Syntax

Christer Samuelsson
Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan, FRANCE
Christer.Samuelsson@xrce.xerox.com

Historical Backdrop

Towards an implementable dependency grammar

Timo Järvinen and **Pasi Tapanainen**
Research Unit for Multilingual Language
P.O. Box 4, FIN-00014 University of Helsi

A Statistical Theory of Dependency Syntax

Christer Samuelsson
Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan, FRANCE
rister.Samuelsson@xrce.xerox.com

An English Dependency Treebank
à la Tesnière

Federico Sangati
University of Amsterdam
f.sangati@uva.nl

Chiara Mazza
University of Pisa
chiara.mazza@gmail.com

Historical Backdrop

Towards an implementable dependency grammar

Timo Järvinen and **Pasi Tapanainen**
Research Unit for Multilingual Language
P.O. Box 4, FIN-00014 University of Helsi

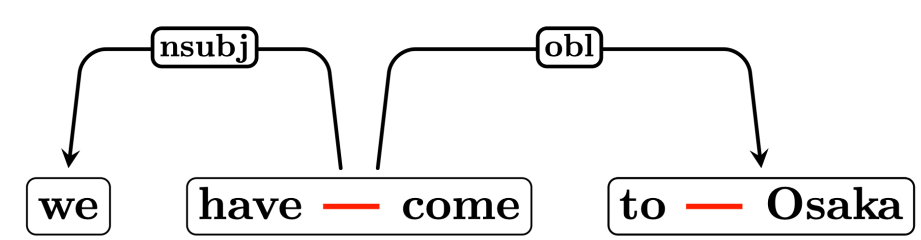
A Statistical Theory of Dependency Syntax

Christer Samuelsson
Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan, FRANCE
rister.Samuelsson@xrce.xerox.com

An English Dependency Treebank
à la Tesnière

Federico Sangati
University of Amsterdam
f.sangati@uva.nl

Chiara Mazza
University of Pisa
chiara.mazza@gmail.com



dependency	nucleus
karaka	vibhakti
kakariuke	bunsetsu

Historical Backdrop

Towards an implementable dependency grammar

Timo Järvinen and Pasi Tapanainen
Research Unit for Multilingual Language
P.O. Box 4, FIN-00014 University of Helsi

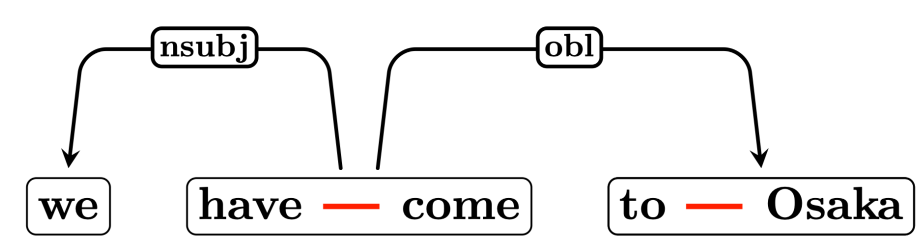
A Statistical Theory of Dependency Syntax

Christer Samuelsson
Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan, FRANCE
rister.Samuelsson@xrce.xerox.com

An English Dependency Treebank
à la Tesnière

Federico Sangati
University of Amsterdam
f.sangati@uva.nl

Chiara Mazza
University of Pisa
chiara.mazza@gmail.com



- Lack of annotated corpora

dependency	nucleus
karaka	vibhakti
kakariuke	bunsetsu

Historical Backdrop

Towards an implementable dependency grammar

Timo Järvinen and **Pasi Tapanainen**
Research Unit for Multilingual Language
P.O. Box 4, FIN-00014 University of Helsi

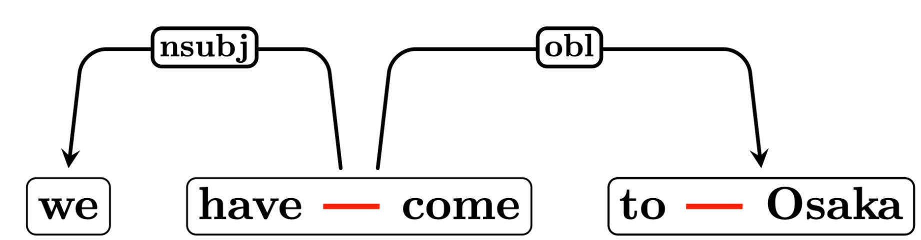
A Statistical Theory of Dependency Syntax

Christer Samuelsson
Xerox Research Centre Europe
6, chemin de Maupertuis
38240 Meylan, FRANCE
rister.Samuelsson@xrce.xerox.com

An English Dependency Treebank
à la Tesnière

Federico Sangati
University of Amsterdam
f.sangati@uva.nl

Chiara Mazza
University of Pisa
chiara.mazza@gmail.com



dependency	nucleus
karaka	vibhakti
kakariuke	bunsetsu

- Lack of annotated corpora
- Lack of appropriate parsers

Universal Dependencies

- Framework for morphosyntactic annotation
- Designed to promote cross-linguistic consistency
- UD v2.11: 243 treebanks, 138 languages, 29 families

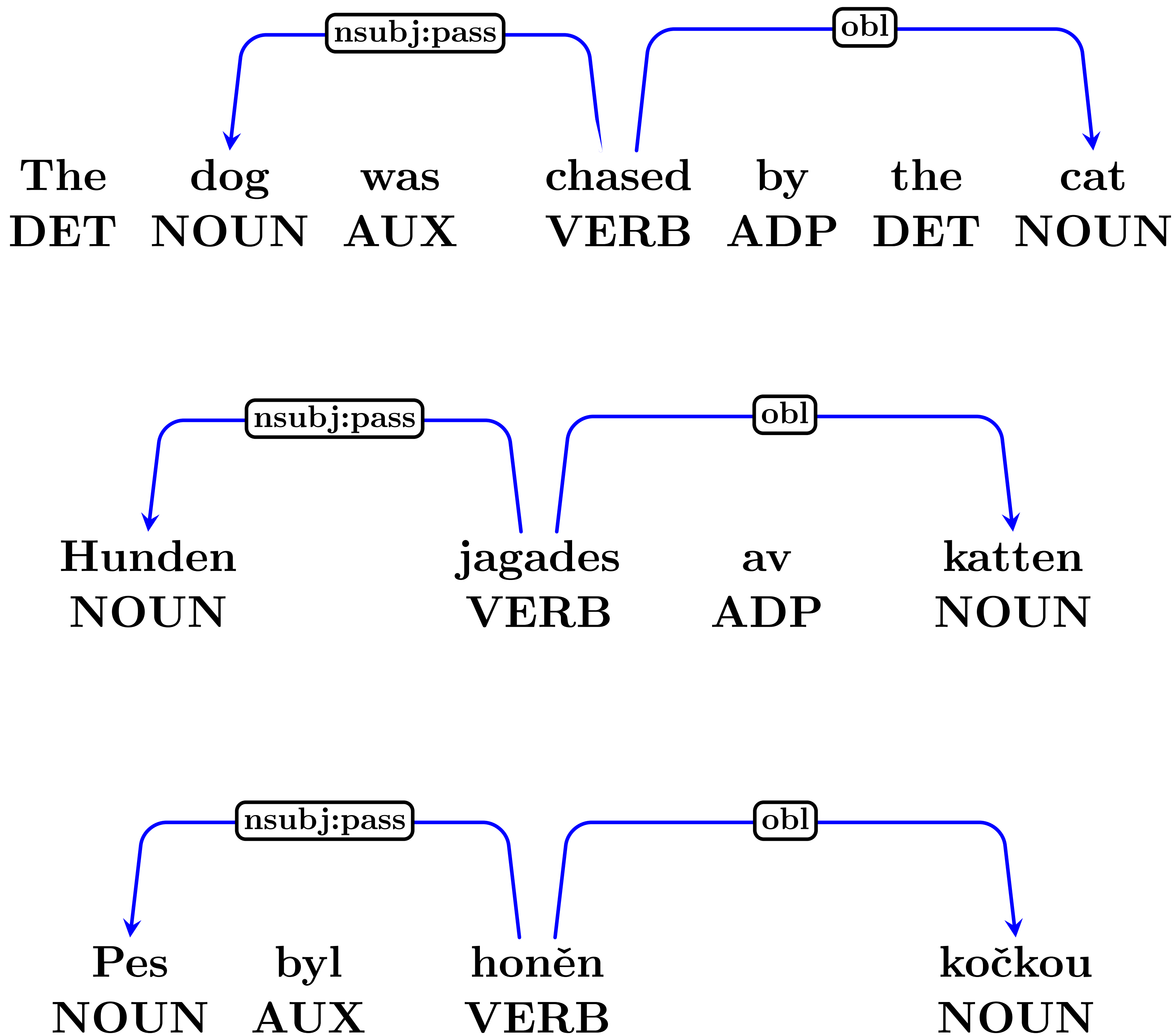
Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of *LREC*.

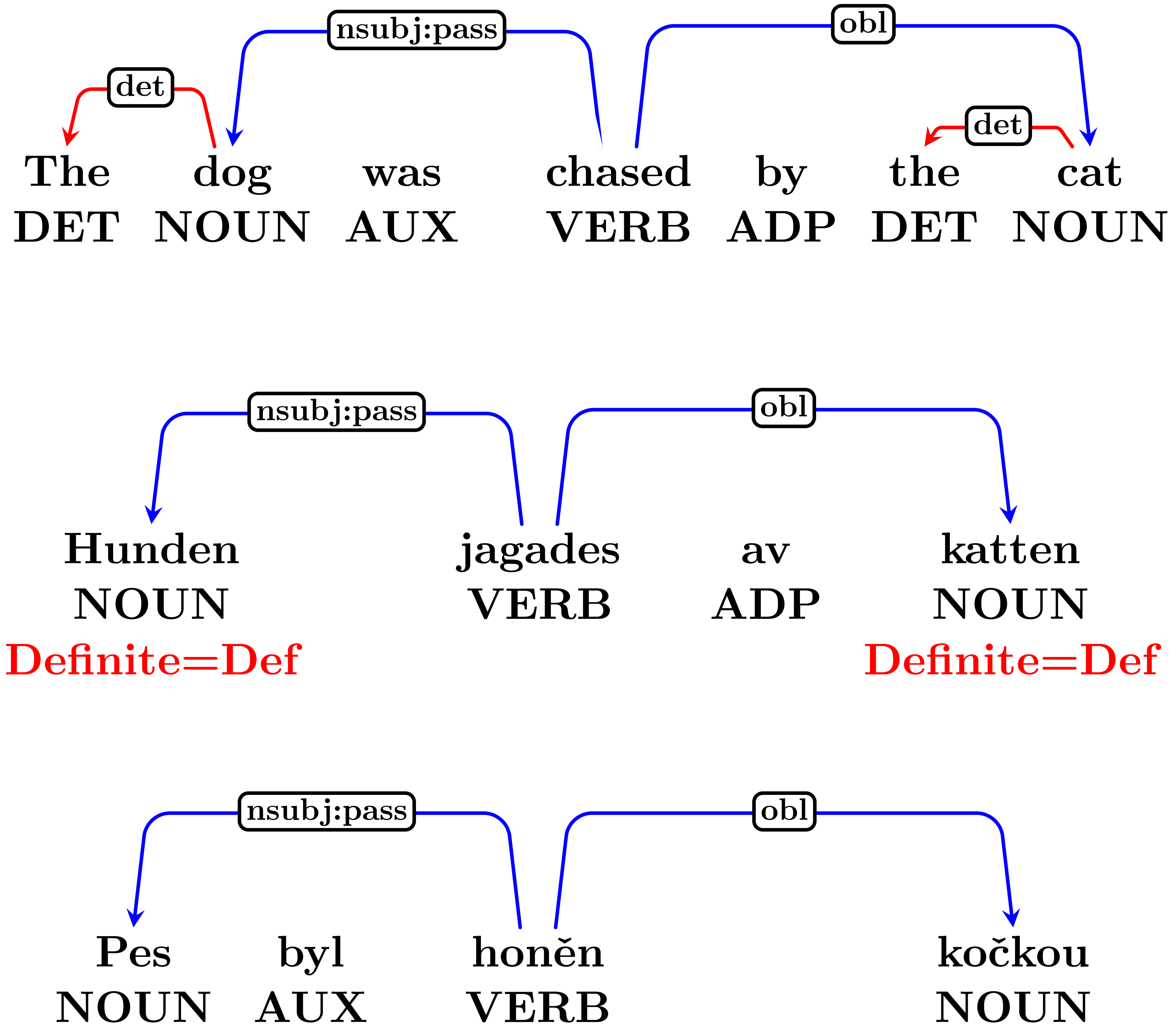
Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In Proceedings *LREC*, 4034–4043

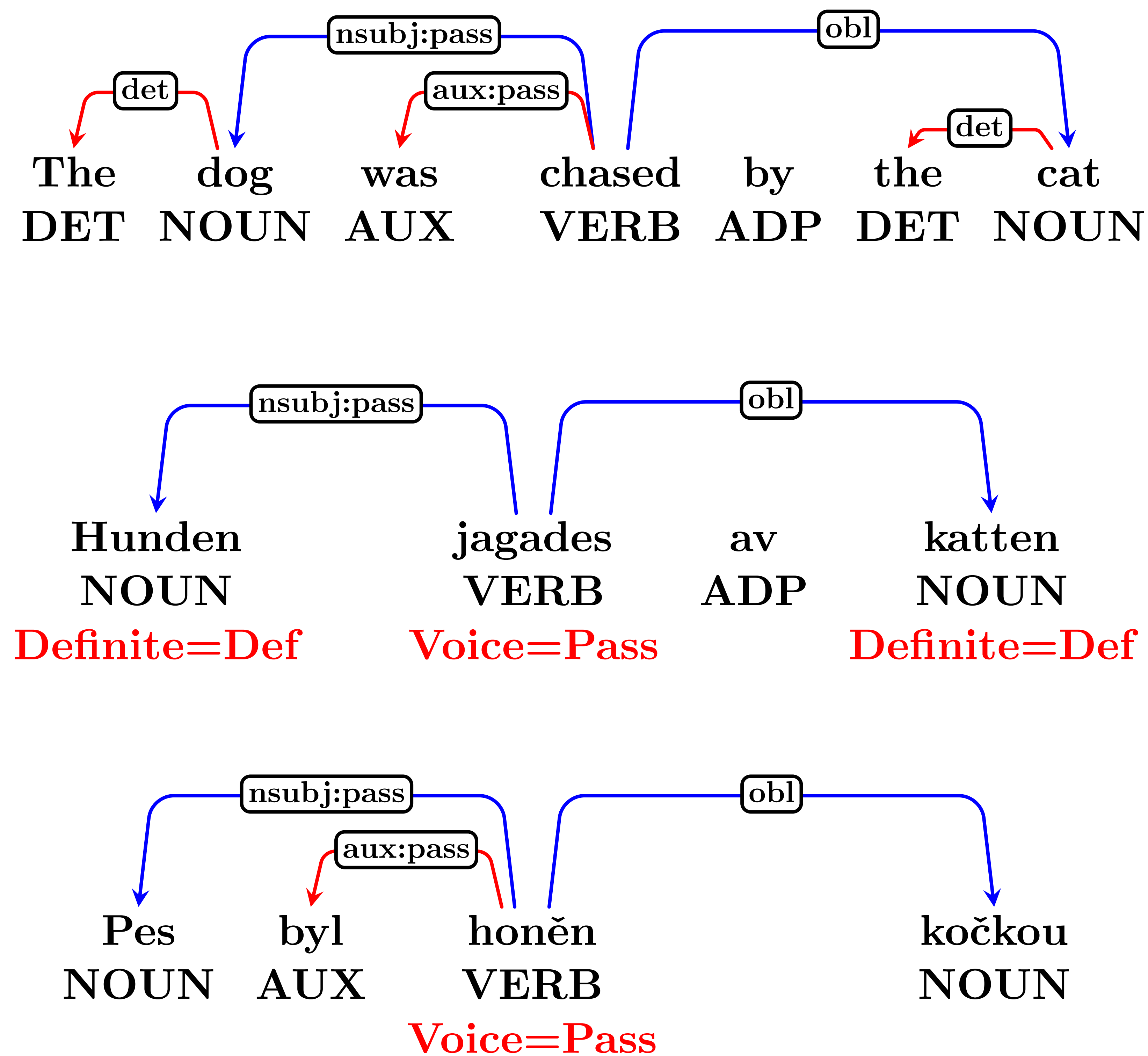
Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, Daniel Zeman (2021): Universal Dependencies. *Computational Linguistics*, 47(2): 255–308.

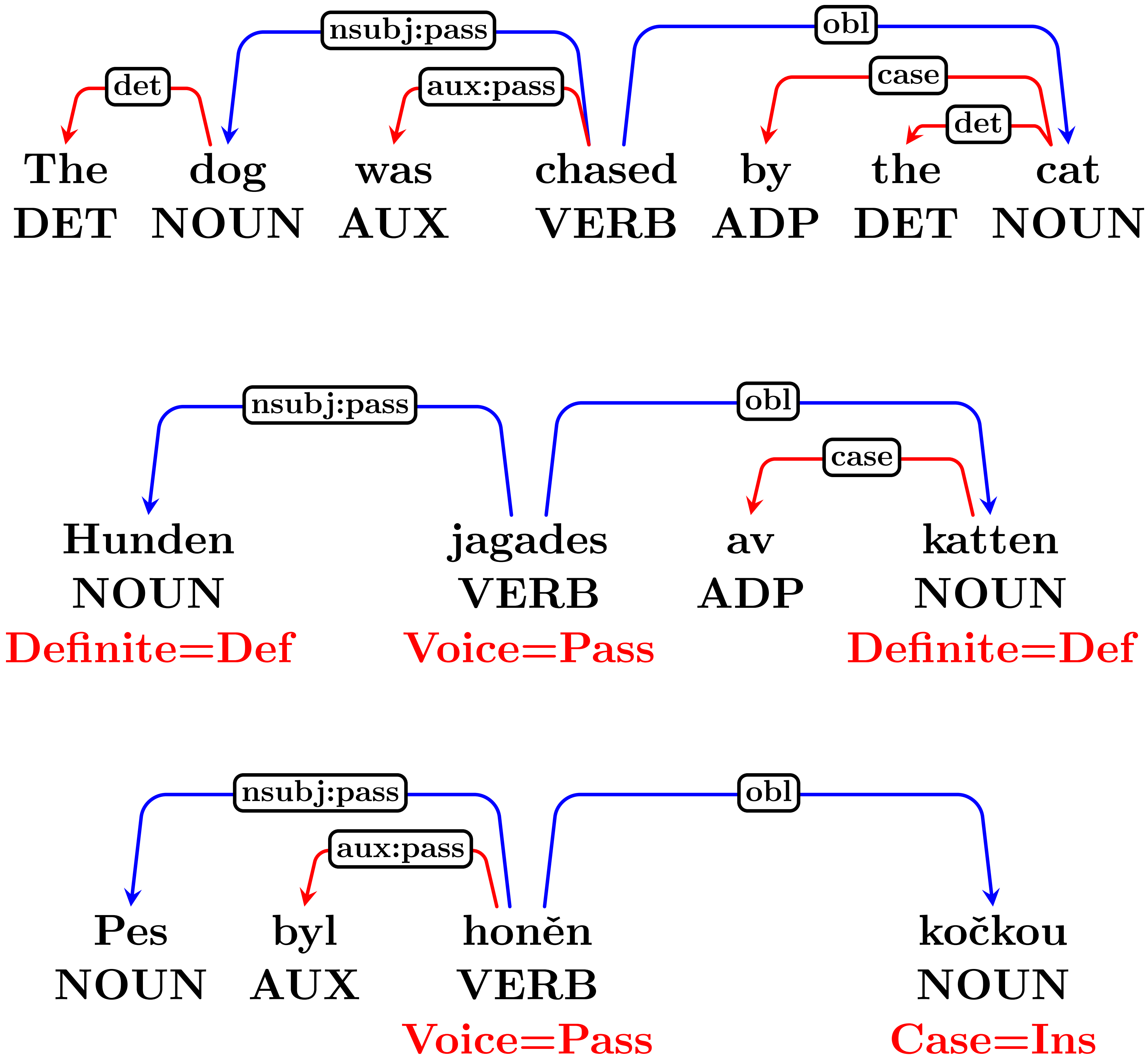
Universal Dependencies

- UD representations are word-based – but nucleus-aware
- UD prioritizes direct relations between content words
- UD treats function words as grammatical markers


nsubj:pass
obl
The dog was chased by the cat
DET NOUN AUX VERB ADP DET NOUN
nsubj:pass
obl
Hunden jagades av katten
NOUN VERB ADP NOUN
nsubj:pass
obl
Pes byl honěn kočkou
NOUN AUX VERB NOUN

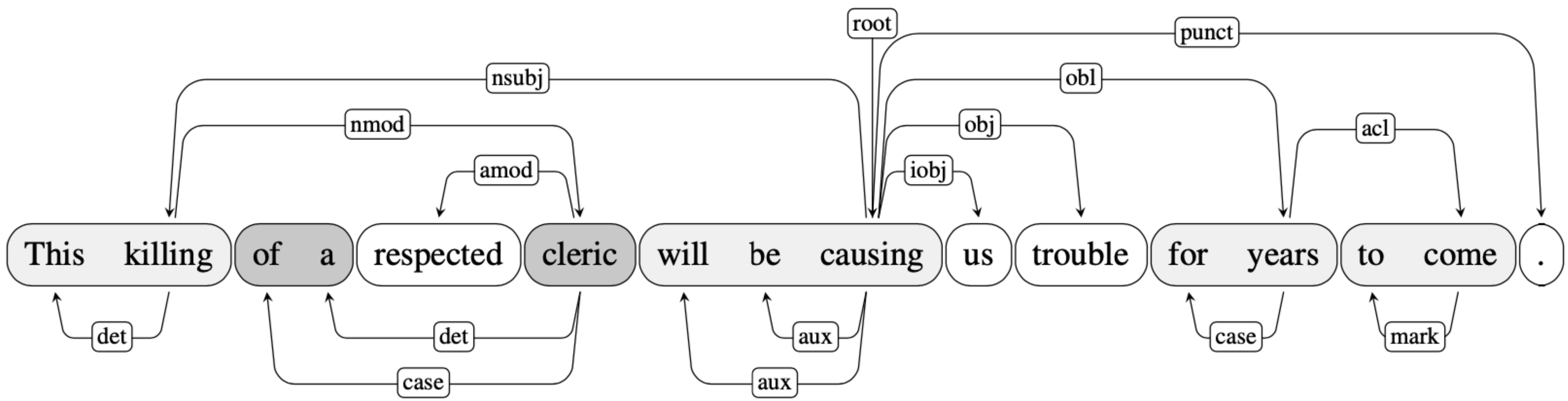

det
nsubj:pass
obl
det
The dog was chased by the cat
DET NOUN AUX VERB ADP DET NOUN
nsubj:pass
obl
Hunden jagades av katten
NOUN VERB ADP NOUN
Definite=Def
Definite=Def
nsubj:pass
obl
Pes byl honěn kočkou
NOUN AUX VERB NOUN


nsubj:pass
obl
det
aux:pass
det
The dog was chased by the cat
DET NOUN AUX VERB ADP DET NOUN
nsubj:pass
obl
Hunden jagades av katten
NOUN VERB ADP NOUN
Definite=Def Voice=Pass Definite=Def
nsubj:pass
obl
aux:pass
Pes byl honěn kočkou
NOUN AUX VERB NOUN
Voice=Pass

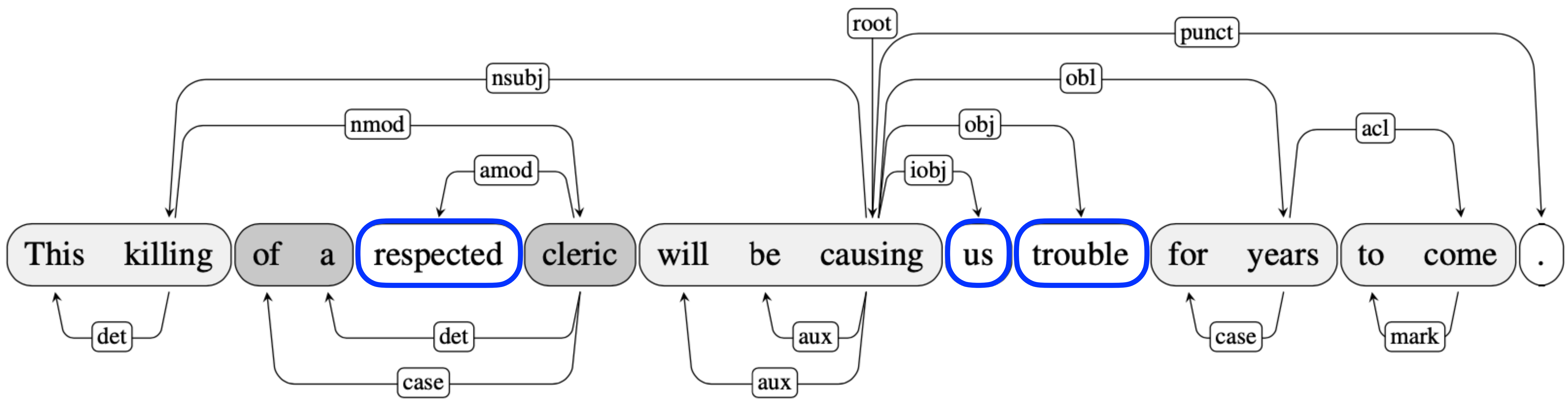

nsubj:pass
obl
det
aux:pass
case
det
The DET
dog NOUN
was AUX
chased VERB
by ADP
the DET
cat NOUN
nsubj:pass
obl
case
Hunden NOUN Definite=Def
jagades VERB Voice=Pass
av ADP
katten NOUN Definite=Def
nsubj:pass
obl
aux:pass
Pes NOUN
byl AUX
honěn VERB Voice=Pass
kočkou NOUN Case=Ins

Syntactic Nuclei in UD

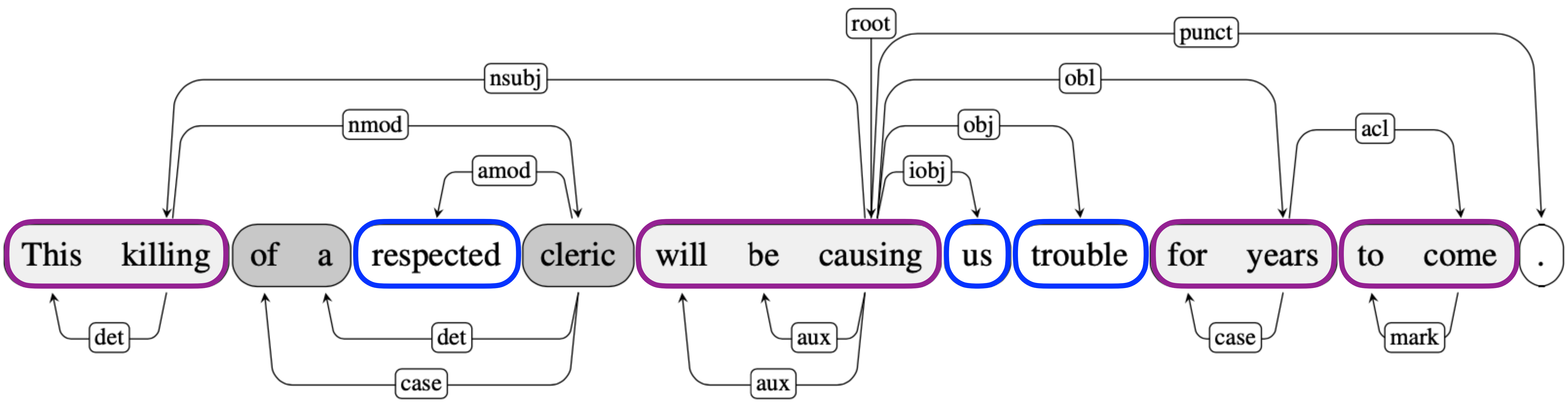
- Content word ≈ lexical core of a nucleus
- Function word ≈ non-lexical part of dissociated nucleus
- Nucleus ≈ subtree containing only functional relations



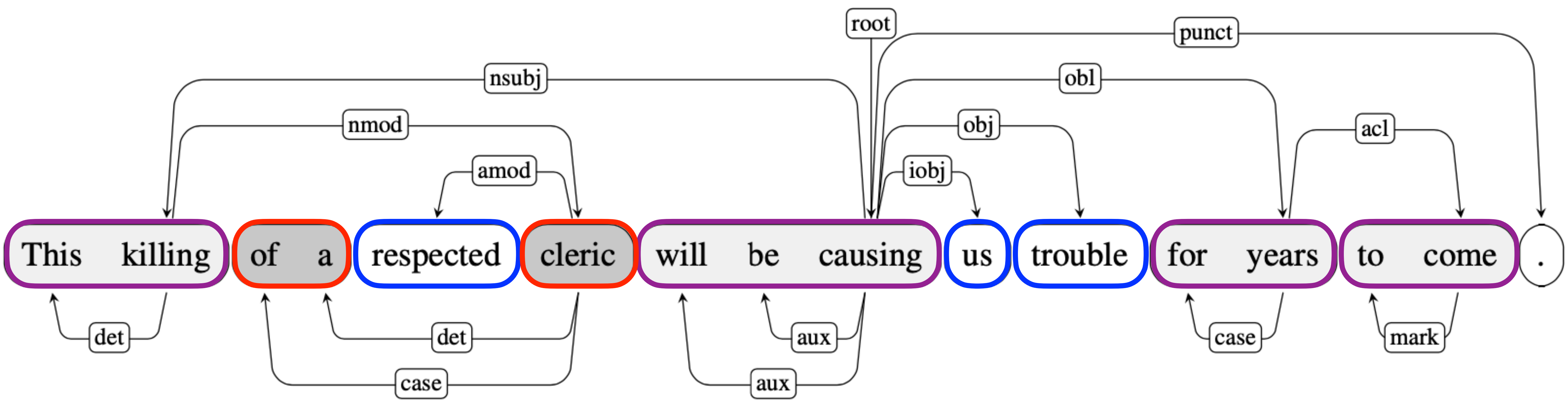
root
punct
nsubj
obl
nmod
obj
acl
amod
iobj
This killing
of a
respected
cleric
will be causing
us
trouble
for years
to come
.
det
det
case
aux
aux
case
mark



root
punct
nsubj
obl
nmod
obj
acl
amod
iobj
This killing
of a
respected
cleric
will be causing
us
trouble
for years
to come
.
det
det
aux
case
mark
case
aux



root
punct
nsubj
obl
nmod
obj
acl
amod
iobj
This killing
of a
respected
cleric
will be causing
us
trouble
for years
to come
.
det
det
aux
case
mark
case
aux



root
punct
nsubj
obl
nmod
obj
acl
amod
iobj
This
killing
of
a
respected
cleric
will
be
causing
us
trouble
for
years
to
come
.
det
det
aux
case
mark
case
aux

Functional Relations

- Determiner (det)
- Case marker (case)
- Classifier (clf)
- Auxiliary (aux)
- Copula (cop)
- Subordination marker (mark)
- Coordinating conjunction (cc)

Functional Relations

- Determiner (det)
- Case marker (case)
- Classifier (clf)
- Auxiliary (aux)
- Copula (cop)
- Subordination marker (mark)
- Coordinating conjunction (cc)

} Nominals

Functional Relations

- Determiner (det)
- Case marker (case)
- Classifier (clf)
- Auxiliary (aux)
- Copula (cop)
- Subordination marker (mark)
- Coordinating conjunction (cc)

} Nominals (det, case, clf)

} Predicates (aux, cop, mark)

Functional Relations

- Determiner (det)
- Case marker (case)
- Classifier (clf)
- Auxiliary (aux)
- Copula (cop)
- Subordination marker (mark)
- Coordinating conjunction (cc) – Tesnière's junction

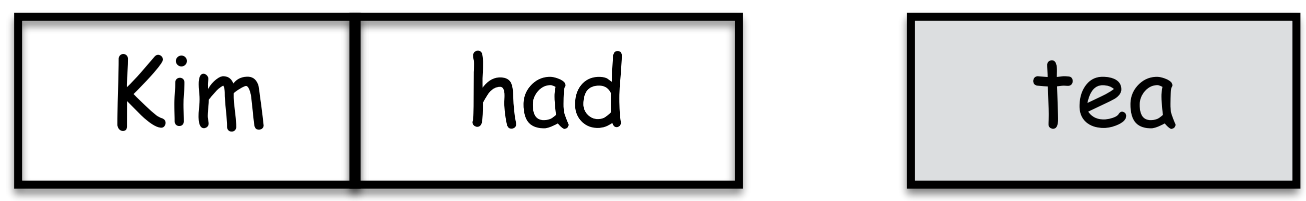
Nominals: det, case, clf

Predicates: aux, cop, mark

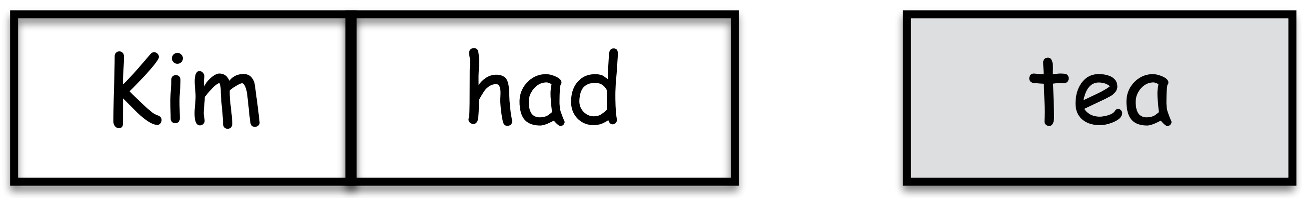
From UD to Parsing

- How can we use our nuclei with standard parsers?
- Evaluation: Content Labeled Attachment Score (CLAS)
- Composition: Parser-internal representations of nuclei

Transition-Based Parsing

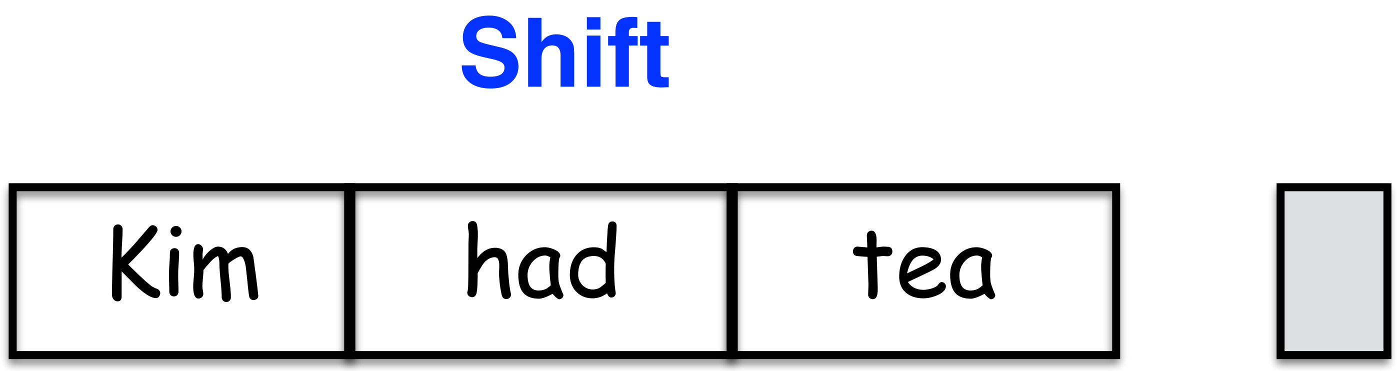


Transition-Based Parsing



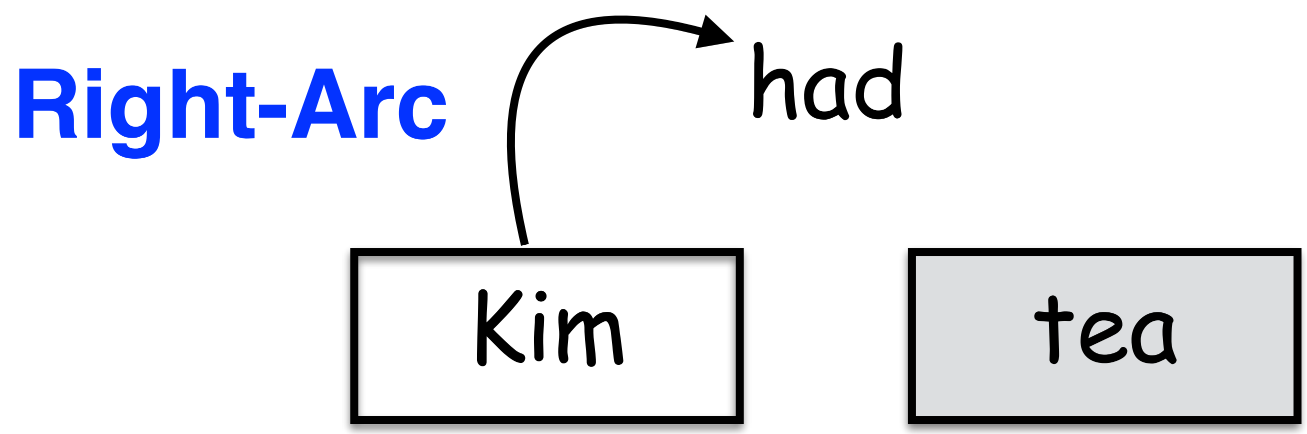
- Dependency trees ≈ derivations in a transition system

Transition-Based Parsing



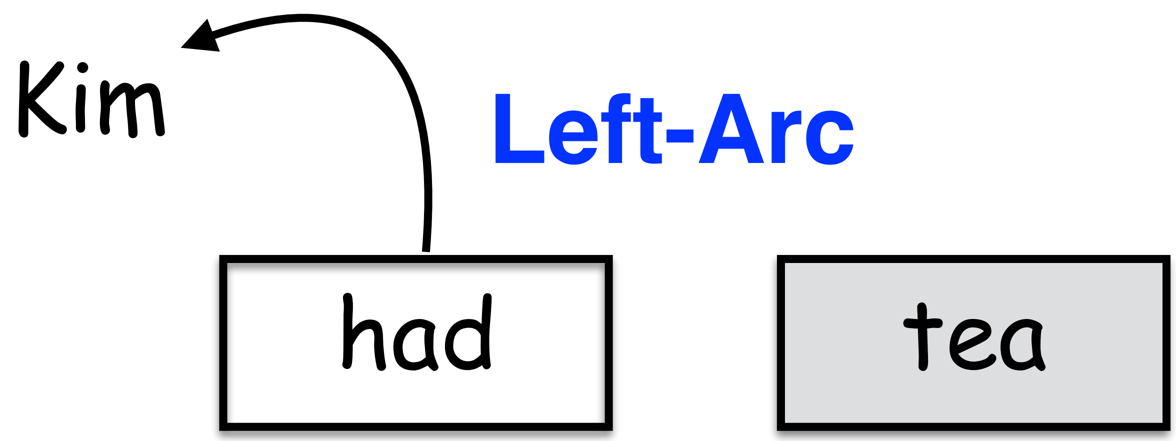
- Dependency trees ≈ derivations in a transition system

Transition-Based Parsing



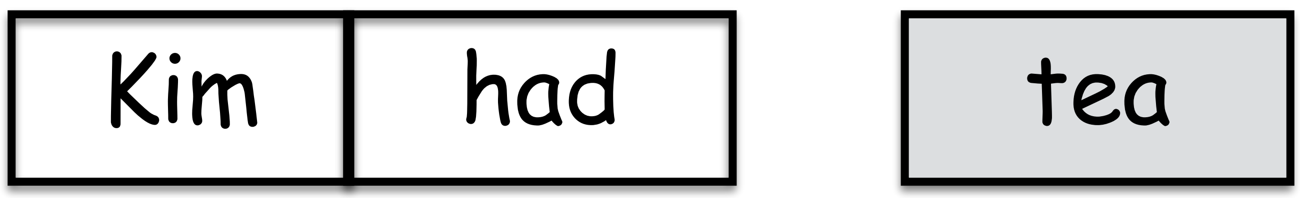
- Dependency trees ≈ derivations in a transition system

Transition-Based Parsing



- Dependency trees ≈ derivations in a transition system

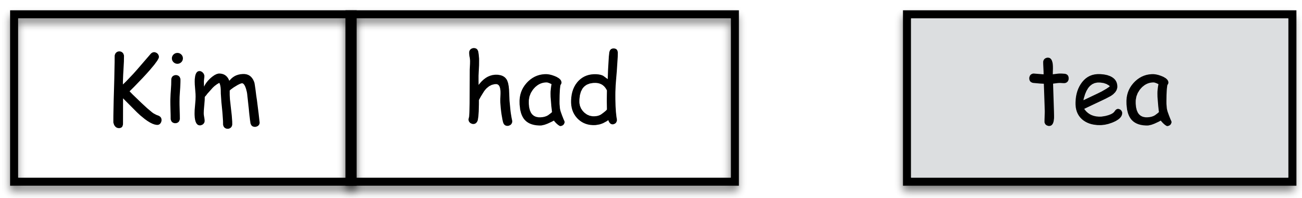
Transition-Based Parsing



$$S(T) = S(D)_{D \Rightarrow T} = \sum_{(c,t) \in D} S(c, t)$$

- Dependency trees ≈ derivations in a transition system
- Learn model *M* to score derivations by transitions

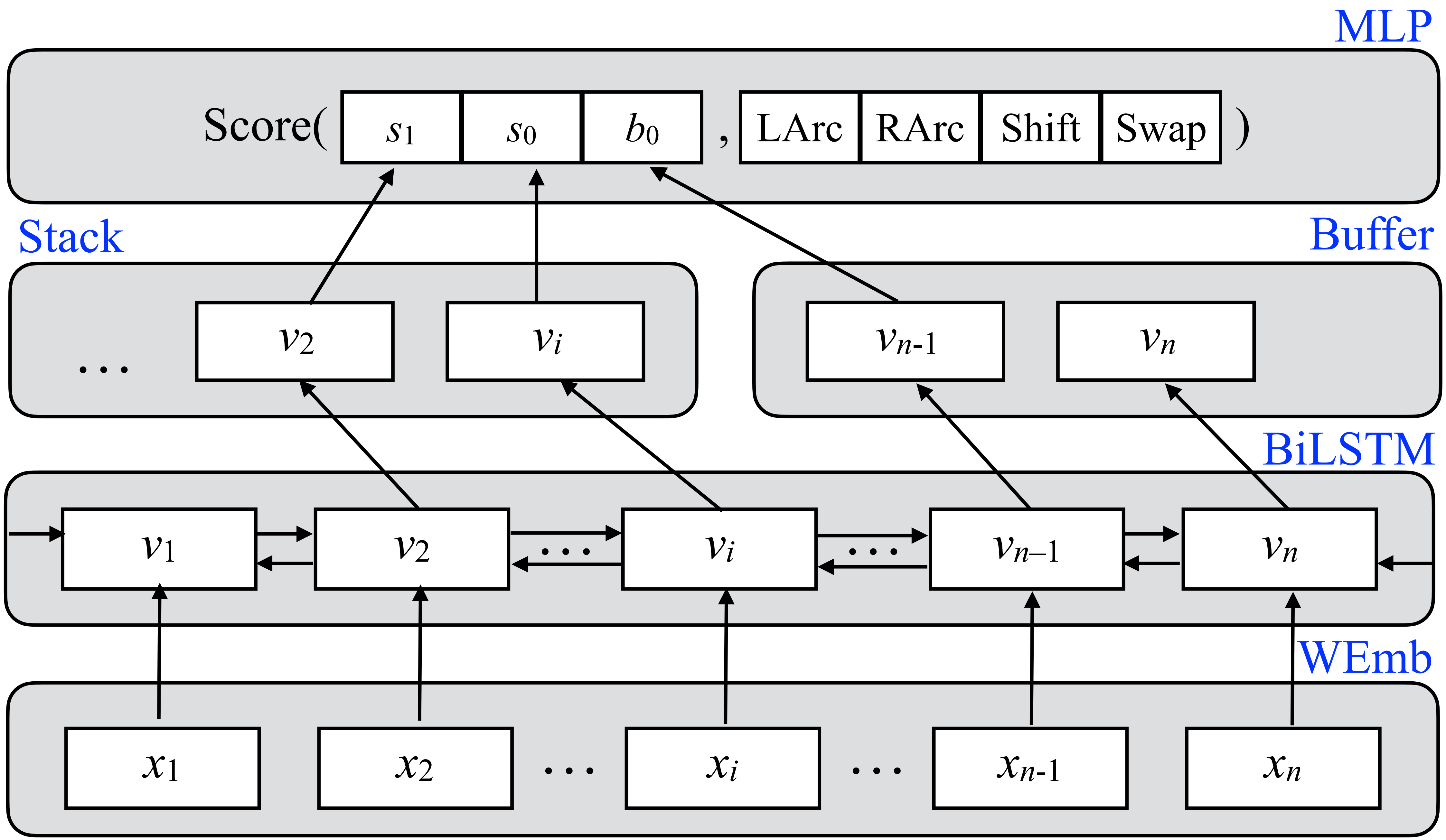
Transition-Based Parsing



$$T^* = T : \arg\max_{D} S(D) \Rightarrow T$$

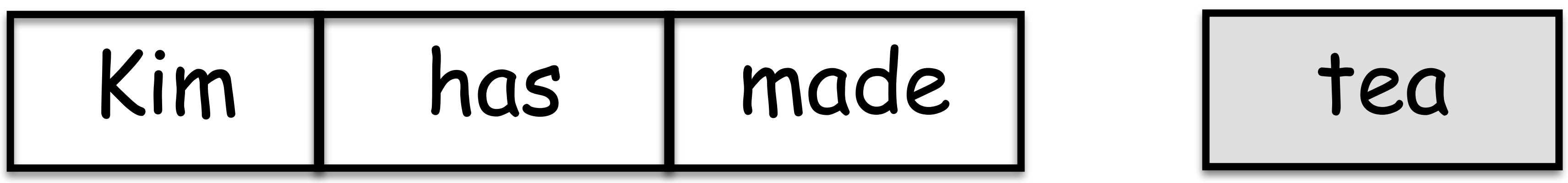
- Dependency trees ≈ derivations in a transition system
- Learn model *M* to score derivations by transitions
- Find highest scoring derivation *D* under the model *M*

Parsing Architecture

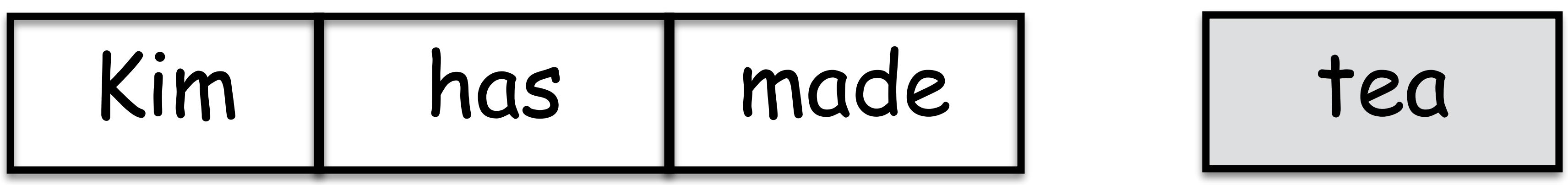


Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representation Networks. *TACL* 4: 313–327.

Adding Nuclei



Adding Nuclei

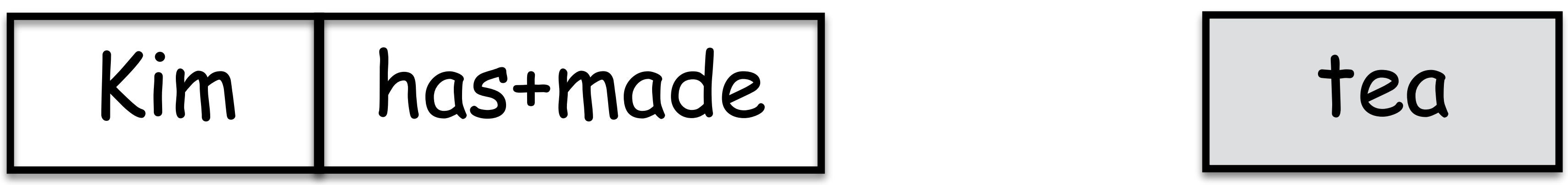


- Subtrees are represented by their root

Adding Nuclei

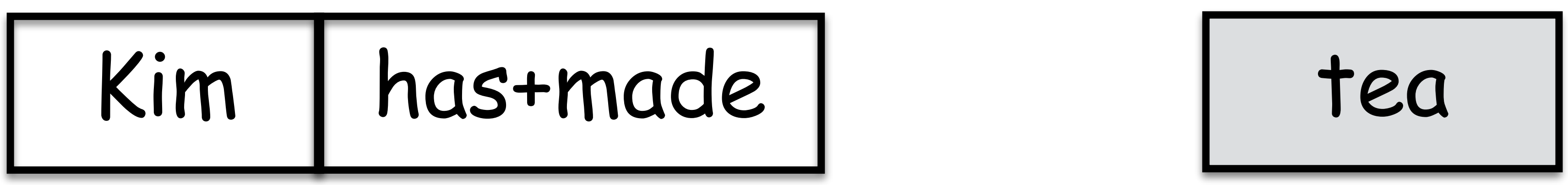
- Subtrees are represented by their root
- Old model: root word

Adding Nuclei



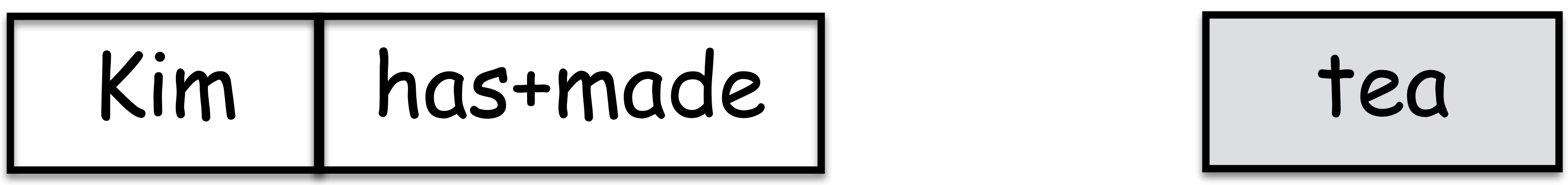
- Subtrees are represented by their root
- Old model: root word
- New model: root nucleus

Adding Nuclei



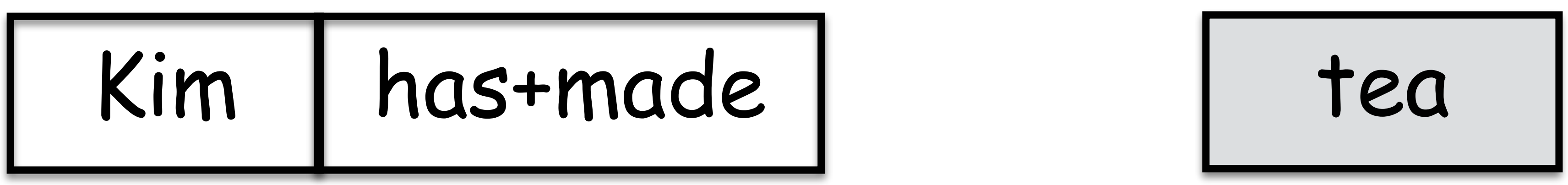
- Subtrees are represented by their root
- Old model: root word
- New model: root nucleus
- Alternative 1: new transition for nucleus creation

Adding Nuclei



- Subtrees are represented by their root
- Old model: root word
- New model: root nucleus
- Alternative 1: new transition for nucleus creation
- Alternative 2: nucleus composition at arc creation

Adding Nuclei



- Subtrees are represented by their root
- Old model: root word
- New model: root nucleus
- Alternative 1: new transition for nucleus creation
- Alternative 2: nucleus composition at arc creation
- Possible thanks to incremental history-based parsing

Recursive Composition

Recursive Composition

Pontus Stenetorp. 2013. Transition-Based Dependency Parsing Using Recursive Neural Networks. In *Proceedings of the Deep Learning Workshop at NIPS*.

Recursive Composition

Pontus Stenetorp. 2013. Transition-Based Dependency Parsing Using Recursive Neural Networks. In *Proceedings of the Deep Learning Workshop at NIPS*.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews and Noah A. Smith. 2015. Transition-Based Dependency Parsing with Stack Long Short-Term Memory. In *Proceedings of ACL*, 334–343.

Recursive Composition

Pontus Stenetorp. 2013. Transition-Based Dependency Parsing Using Recursive Neural Networks. In *Proceedings of the Deep Learning Workshop at NIPS*.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews and Noah A. Smith. 2015. Transition-Based Dependency Parsing with Stack Long Short-Term Memory. In *Proceedings of ACL*, 334–343.

Miryam de Lhoneux, Miguel Ballesteros, Sara Stymne and Joakim Nivre. 2019. Recursive Subtree Composition in LSTM-Based Dependency Parsing. In *Proceedings of NAACL*, 1566–1576.

Recursive Composition

Pontus Stenetorp. 2013. Transition-Based Dependency Parsing Using Recursive Neural Networks. In *Proceedings of the Deep Learning Workshop at NIPS*.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews and Noah A. Smith. 2015. Transition-Based Dependency Parsing with Stack Long Short-Term Memory. In *Proceedings of ACL*, 334–343.

Miryam de Lhoneux, Miguel Ballesteros, Sara Stymne and Joakim Nivre. 2019. Recursive Subtree Composition in LSTM-Based Dependency Parsing. In *Proceedings of NAACL*, 1566–1576.

Miryam de Lhoneux, Sara Stymne and Joakim Nivre. 2019. What Should/Do/Can LSTMs Learn When Parsing Auxiliary-Verb Constructions. *Computational Linguistics* 46(4): 763–784.

Nucleus Composition

Nucleus Composition

- Nucleus representation: $f(h, d, l)$

h = head
d = dependent
l = label

Nucleus Composition

- Nucleus representation: $f(h, d, l)$
- Baseline model: $f(h, d, l) = h$

h = head
d = dependent
l = label

Nucleus Composition

- Nucleus representation: $f(h, d, l)$
- Baseline model: $f(h, d, l) = h$
- Nucleus composition model:

h = head
d = dependent
l = label

$$f(h, d, l) = \begin{cases} h + g(h, d, l) & \text{if } l \in F \\ h & \text{otherwise} \end{cases}$$

$$g(h, d, l) = \sigma(W(h \circ d \circ l) + b)$$

Data Sets

Language	Treebank	Family	Genus	Size	aux	case	cc	clf	cop	det	mark	Func
Arabic	PADT	Afro-Asiatic	Semitic	242K	0.60	14.29	5.11	0.00	0.16	0.76	2.71	23.63
Armenian	ArmTDP	Indo-European	Armenian	52K	5.04	3.03	4.10	0.00	2.01	3.46	1.67	19.30
Basque	BDT	Basque	Basque	121K	8.54	1.56	3.85	0.00	2.02	2.50	0.18	18.65
Chinese	GSD	Sino-Tibetan	Chinese	121K	1.83	6.31	1.42	1.82	1.45	1.35	5.75	19.93
Finnish	TDT	Uralic-Finnic	Finnish	202K	3.26	1.48	4.13	0.00	2.72	1.72	1.95	15.27
Greek	GDT	Indo-European	Greek	62K	3.81	8.47	3.19	0.00	0.94	19.12	1.83	37.37
Hebrew	HTB	Afro-Asiatic	Semitic	116K	0.45	16.26	2.93	0.00	0.69	11.55	3.32	35.19
Hindi	HDTB	Indo-European	Indic	352K	6.41	19.27	1.87	0.00	1.00	2.05	4.11	34.70
Indonesian	GSD	Austronesian	Malayo-Sumbawan	121K	0.00	9.87	2.96	0.00	0.87	3.71	1.31	18.72
Irish	IDT	Indo-European	Celtic	116K	0.00	13.44	3.14	0.00	1.32	8.15	5.79	31.84
Italian	ISDT	Indo-European	Romance	278K	2.77	14.01	2.73	0.00	1.15	16.30	2.11	39.08
Japanese	GSD	Japanese	Japanese	194K	8.90	21.34	0.42	0.00	1.26	0.49	4.06	36.47
Korean	GSD	Korean	Korean	80K	0.08	2.03	0.28	0.00	0.13	3.83	0.46	6.81
Latvian	LVTB	Indo-European	Baltic	252K	1.26	4.68	4.01	0.00	1.39	2.63	1.91	15.87
Persian	PerDT	Indo-European	Iranian	494K	2.73	14.17	4.24	0.00	1.27	2.05	2.39	26.85
Russian	Taiga	Indo-European	Slavic	197K	0.30	8.56	4.12	0.00	0.41	2.49	1.63	17.51
Swedish	Talbanken	Indo-European	Germanic	97K	2.65	10.02	3.70	0.00	1.77	5.08	4.01	27.23
Turkish	Kenet	Turkic	Southwestern	179K	0.49	2.11	1.68	0.01	0.00	4.33	0.35	8.97
Vietnamese	VTB	Austro-Asiatic	Viet-Muong	44K	1.34	5.35	3.80	0.00	0.95	3.60	0.49	15.52
Wolof	WTB	Niger-Congo	Northern-Atlantic	43K	7.46	5.46	3.09	0.00	1.36	7.09	4.14	28.59
Average				168K	2.90	9.08	3.04	0.09	1.14	5.11	2.51	23.88

Data Sets

Language	Treebank	Family	Genus	Size	aux	case	cc	clf	cop	det	mark	Func
Arabic	PADT	Afro-Asiatic	Semitic	242K	0.60	14.29	5.11	0.00	0.16	0.76	2.71	23.63
Armenian	ArmTDP	Indo-European	Armenian	52K	5.04	3.03	4.10	0.00	2.01	3.46	1.67	19.30
Basque	BDT	Basque	Basque	121K	8.54	1.56	3.85	0.00	2.02	2.50	0.18	18.65
Chinese	GSD	Sino-Tibetan	Chinese	121K	1.83	6.31	1.42	1.82	1.45	1.35	5.75	19.93
Finnish	TDT	Uralic-Finnic	Finnish	202K	3.26	1.48	4.13	0.00	2.72	1.72	1.95	15.27
Greek	GDT	Indo-European	Greek	62K	3.81	8.47	3.19	0.00	0.94	19.12	1.83	37.37
Hebrew	HTB	Afro-Asiatic	Semitic	116K	0.45	16.26	2.93	0.00	0.69	11.55	3.32	35.19
Hindi	HDTB	Indo-European	Indic	352K	6.41	19.27	1.87	0.00	1.00	2.05	4.11	34.70
Indonesian	GSD	Austronesian	Malayo-Sumbawan	121K	0.00	9.87	2.96	0.00	0.87	3.71	1.31	18.72
Irish	IDT	Indo-European	Celtic	116K	0.00	13.44	3.14	0.00	1.32	8.15	5.79	31.84
Italian	ISDT	Indo-European	Romance	278K	2.77	14.01	2.73	0.00	1.15	16.30	2.11	39.08
Japanese	GSD	Japanese	Japanese	194K	8.90	21.34	0.42	0.00	1.26	0.49	4.06	36.47
Korean	GSD	Korean	Korean	80K	0.08	2.03	0.28	0.00	0.13	3.83	0.46	6.81
Latvian	LVTB	Indo-European	Baltic	252K	1.26	4.68	4.01	0.00	1.39	2.63	1.91	15.87
Persian	PerDT	Indo-European	Iranian	494K	2.73	14.17	4.24	0.00	1.27	2.05	2.39	26.85
Russian	Taiga	Indo-European	Slavic	197K	0.30	8.56	4.12	0.00	0.41	2.49	1.63	17.51
Swedish	Talbanken	Indo-European	Germanic	97K	2.65	10.02	3.70	0.00	1.77	5.08	4.01	27.23
Turkish	Kenet	Turkic	Southwestern	179K	0.49	2.11	1.68	0.01	0.00	4.33	0.35	8.97
Vietnamese	VTB	Austro-Asiatic	Viet-Muong	44K	1.34	5.35	3.80	0.00	0.95	3.60	0.49	15.52
Wolof	WTB	Niger-Congo	Northern-Atlantic	43K	7.46	5.46	3.09	0.00	1.36	7.09	4.14	28.59
Average				168K	2.90	9.08	3.04	0.09	1.14	5.11	2.51	23.88

Data Sets

Language	Treebank	Family	Genus	Size	aux	case	cc	clf	cop	det	mark	Func
Arabic	PADT	Afro-Asiatic	Semitic	242K	0.60	14.29	5.11	0.00	0.16	0.76	2.71	23.63
Armenian	ArmTDP	Indo-European	Armenian	52K	5.04	3.03	4.10	0.00	2.01	3.46	1.67	19.30
Basque	BDT	Basque	Basque	121K	8.54	1.56	3.85	0.00	2.02	2.50	0.18	18.65
Chinese	GSD	Sino-Tibetan	Chinese	121K	1.83	6.31	1.42	1.82	1.45	1.35	5.75	19.93
Finnish	TDT	Uralic-Finnic	Finnish	202K	3.26	1.48	4.13	0.00	2.72	1.72	1.95	15.27
Greek	GDT	Indo-European	Greek	62K	3.81	8.47	3.19	0.00	0.94	19.12	1.83	37.37
Hebrew	HTB	Afro-Asiatic	Semitic	116K	0.45	16.26	2.93	0.00	0.69	11.55	3.32	35.19
Hindi	HDTB	Indo-European	Indic	352K	6.41	19.27	1.87	0.00	1.00	2.05	4.11	34.70
Indonesian	GSD	Austronesian	Malayo-Sumbawan	121K	0.00	9.87	2.96	0.00	0.87	3.71	1.31	18.72
Irish	IDT	Indo-European	Celtic	116K	0.00	13.44	3.14	0.00	1.32	8.15	5.79	31.84
Italian	ISDT	Indo-European	Romance	278K	2.77	14.01	2.73	0.00	1.15	16.30	2.11	39.08
Japanese	GSD	Japanese	Japanese	194K	8.90	21.34	0.42	0.00	1.26	0.49	4.06	36.47
Korean	GSD	Korean	Korean	80K	0.08	2.03	0.28	0.00	0.13	3.83	0.46	6.81
Latvian	LVTB	Indo-European	Baltic	252K	1.26	4.68	4.01	0.00	1.39	2.63	1.91	15.87
Persian	PerDT	Indo-European	Iranian	494K	2.73	14.17	4.24	0.00	1.27	2.05	2.39	26.85
Russian	Taiga	Indo-European	Slavic	197K	0.30	8.56	4.12	0.00	0.41	2.49	1.63	17.51
Swedish	Talbanken	Indo-European	Germanic	97K	2.65	10.02	3.70	0.00	1.77	5.08	4.01	27.23
Turkish	Kenet	Turkic	Southwestern	179K	0.49	2.11	1.68	0.01	0.00	4.33	0.35	8.97
Vietnamese	VTB	Austro-Asiatic	Viet-Muong	44K	1.34	5.35	3.80	0.00	0.95	3.60	0.49	15.52
Wolof	WTB	Niger-Congo	Northern-Atlantic	43K	7.46	5.46	3.09	0.00	1.36	7.09	4.14	28.59
Average				168K	2.90	9.08	3.04	0.09	1.14	5.11	2.51	23.88

Data Sets

Language	Treebank	Family	Genus	Size	aux	case	cc	clf	cop	det	mark	Func
Arabic	PADT	Afro-Asiatic	Semitic	242K	0.60	14.29	5.11	0.00	0.16	0.76	2.71	23.63
Armenian	ArmTDP	Indo-European	Armenian	52K	5.04	3.03	4.10	0.00	2.01	3.46	1.67	19.30
Basque	BDT	Basque	Basque	121K	8.54	1.56	3.85	0.00	2.02	2.50	0.18	18.65
Chinese	GSD	Sino-Tibetan	Chinese	121K	1.83	6.31	1.42	1.82	1.45	1.35	5.75	19.93
Finnish	TDT	Uralic-Finnic	Finnish	202K	3.26	1.48	4.13	0.00	2.72	1.72	1.95	15.27
Greek	GDT	Indo-European	Greek	62K	3.81	8.47	3.19	0.00	0.94	19.12	1.83	37.37
Hebrew	HTB	Afro-Asiatic	Semitic	116K	0.45	16.26	2.93	0.00	0.69	11.55	3.32	35.19
Hindi	HDTB	Indo-European	Indic	352K	6.41	19.27	1.87	0.00	1.00	2.05	4.11	34.70
Indonesian	GSD	Austronesian	Malayo-Sumbawan	121K	0.00	9.87	2.96	0.00	0.87	3.71	1.31	18.72
Irish	IDT	Indo-European	Celtic	116K	0.00	13.44	3.14	0.00	1.32	8.15	5.79	31.84
Italian	ISDT	Indo-European	Romance	278K	2.77	14.01	2.73	0.00	1.15	16.30	2.11	39.08
Japanese	GSD	Japanese	Japanese	194K	8.90	21.34	0.42	0.00	1.26	0.49	4.06	36.47
Korean	GSD	Korean	Korean	80K	0.08	2.03	0.28	0.00	0.13	3.83	0.46	6.81
Latvian	LVTB	Indo-European	Baltic	252K	1.26	4.68	4.01	0.00	1.39	2.63	1.91	15.87
Persian	PerDT	Indo-European	Iranian	494K	2.73	14.17	4.24	0.00	1.27	2.05	2.39	26.85
Russian	Taiga	Indo-European	Slavic	197K	0.30	8.56	4.12	0.00	0.41	2.49	1.63	17.51
Swedish	Talbanken	Indo-European	Germanic	97K	2.65	10.02	3.70	0.00	1.77	5.08	4.01	27.23
Turkish	Kenet	Turkic	Southwestern	179K	0.49	2.11	1.68	0.01	0.00	4.33	0.35	8.97
Vietnamese	VTB	Austro-Asiatic	Viet-Muong	44K	1.34	5.35	3.80	0.00	0.95	3.60	0.49	15.52
Wolof	WTB	Niger-Congo	Northern-Atlantic	43K	7.46	5.46	3.09	0.00	1.36	7.09	4.14	28.59
Average				168K	2.90	9.08	3.04	0.09	1.14	5.11	2.51	23.88

Data Sets

Language	Treebank	Family	Genus	Size	aux	case	cc	clf	cop	det	mark	Func
Arabic	PADT	Afro-Asiatic	Semitic	242K	0.60	14.29	5.11	0.00	0.16	0.76	2.71	23.63
Armenian	ArmTDP	Indo-European	Armenian	52K	5.04	3.03	4.10	0.00	2.01	3.46	1.67	19.30
Basque	BDT	Basque	Basque	121K	8.54	1.56	3.85	0.00	2.02	2.50	0.18	18.65
Chinese	GSD	Sino-Tibetan	Chinese	121K	1.83	6.31	1.42	1.82	1.45	1.35	5.75	19.93
Finnish	TDT	Uralic-Finnic	Finnish	202K	3.26	1.48	4.13	0.00	2.72	1.72	1.95	15.27
Greek	GDT	Indo-European	Greek	62K	3.81	8.47	3.19	0.00	0.94	19.12	1.83	37.37
Hebrew	HTB	Afro-Asiatic	Semitic	116K	0.45	16.26	2.93	0.00	0.69	11.55	3.32	35.19
Hindi	HDTB	Indo-European	Indic	352K	6.41	19.27	1.87	0.00	1.00	2.05	4.11	34.70
Indonesian	GSD	Austronesian	Malayo-Sumbawan	121K	0.00	9.87	2.96	0.00	0.87	3.71	1.31	18.72
Irish	IDT	Indo-European	Celtic	116K	0.00	13.44	3.14	0.00	1.32	8.15	5.79	31.84
Italian	ISDT	Indo-European	Romance	278K	2.77	14.01	2.73	0.00	1.15	16.30	2.11	39.08
Japanese	GSD	Japanese	Japanese	194K	8.90	21.34	0.42	0.00	1.26	0.49	4.06	36.47
Korean	GSD	Korean	Korean	80K	0.08	2.03	0.28	0.00	0.13	3.83	0.46	6.81
Latvian	LVTB	Indo-European	Baltic	252K	1.26	4.68	4.01	0.00	1.39	2.63	1.91	15.87
Persian	PerDT	Indo-European	Iranian	494K	2.73	14.17	4.24	0.00	1.27	2.05	2.39	26.85
Russian	Taiga	Indo-European	Slavic	197K	0.30	8.56	4.12	0.00	0.41	2.49	1.63	17.51
Swedish	Talbanken	Indo-European	Germanic	97K	2.65	10.02	3.70	0.00	1.77	5.08	4.01	27.23
Turkish	Kenet	Turkic	Southwestern	179K	0.49	2.11	1.68	0.01	0.00	4.33	0.35	8.97
Vietnamese	VTB	Austro-Asiatic	Viet-Muong	44K	1.34	5.35	3.80	0.00	0.95	3.60	0.49	15.52
Wolof	WTB	Niger-Congo	Northern-Atlantic	43K	7.46	5.46	3.09	0.00	1.36	7.09	4.14	28.59
Average				168K	2.90	9.08	3.04	0.09	1.14	5.11	2.51	23.88

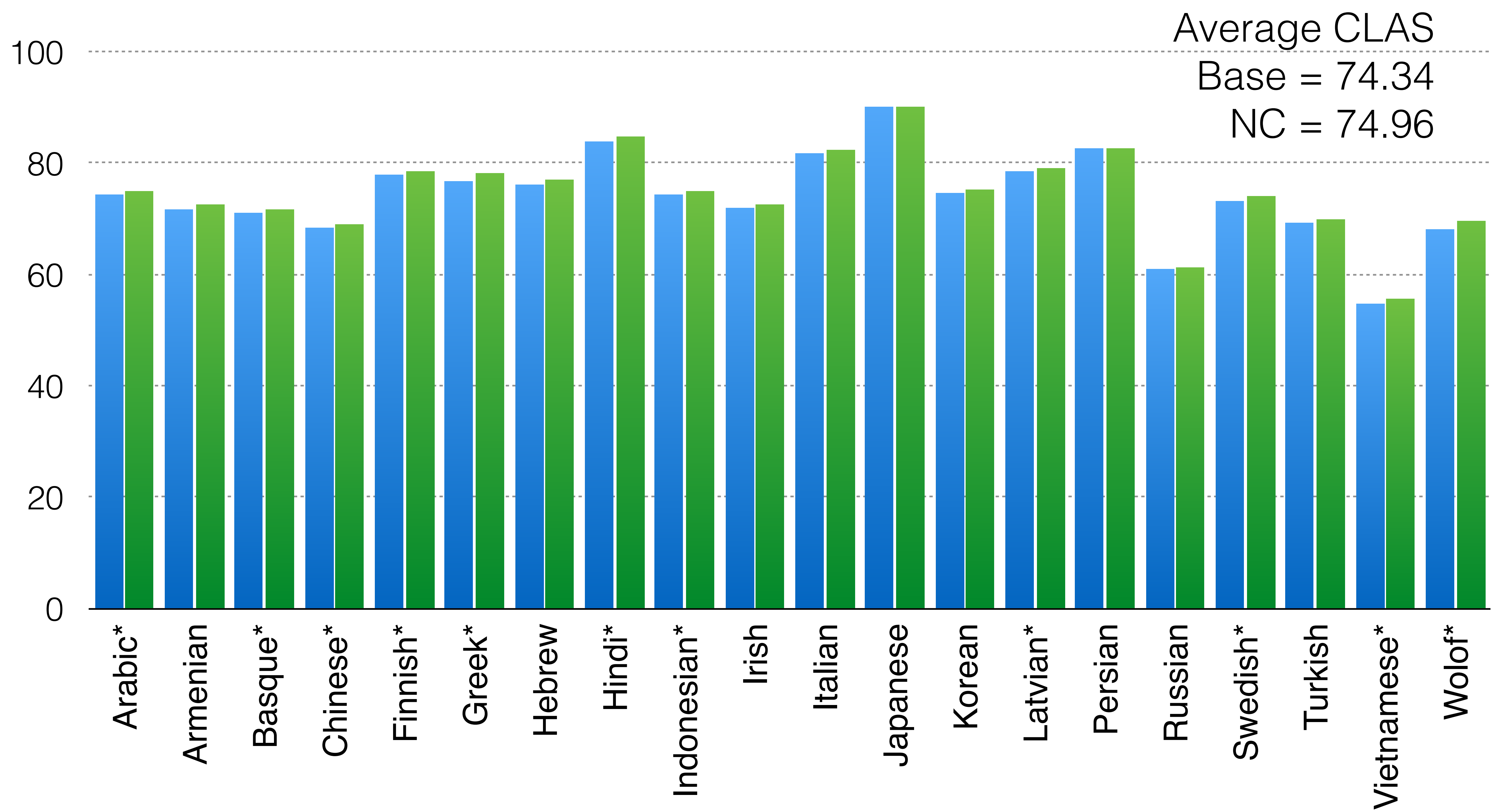
Data Sets

Language	Treebank	Family	Genus	Size	aux	case	cc	clf	cop	det	mark	Func
Arabic	PADT	Afro-Asiatic	Semitic	242K	0.60	14.29	5.11	0.00	0.16	0.76	2.71	23.63
Armenian	ArmTDP	Indo-European	Armenian	52K	5.04	3.03	4.10	0.00	2.01	3.46	1.67	19.30
Basque	BDT	Basque	Basque	121K	8.54	1.56	3.85	0.00	2.02	2.50	0.18	18.65
Chinese	GSD	Sino-Tibetan	Chinese	121K	1.83	6.31	1.42	1.82	1.45	1.35	5.75	19.93
Finnish	TDT	Uralic-Finnic	Finnish	202K	3.26	1.48	4.13	0.00	2.72	1.72	1.95	15.27
Greek	GDT	Indo-European	Greek	62K	3.81	8.47	3.19	0.00	0.94	19.12	1.83	37.37
Hebrew	HTB	Afro-Asiatic	Semitic	116K	0.45	16.26	2.93	0.00	0.69	11.55	3.32	35.19
Hindi	HDTB	Indo-European	Indic	352K	6.41	19.27	1.87	0.00	1.00	2.05	4.11	34.70
Indonesian	GSD	Austronesian	Malayo-Sumbawan	121K	0.00	9.87	2.96	0.00	0.87	3.71	1.31	18.72
Irish	IDT	Indo-European	Celtic	116K	0.00	13.44	3.14	0.00	1.32	8.15	5.79	31.84
Italian	ISDT	Indo-European	Romance	278K	2.77	14.01	2.73	0.00	1.15	16.30	2.11	39.08
Japanese	GSD	Japanese	Japanese	194K	8.90	21.34	0.42	0.00	1.26	0.49	4.06	36.47
Korean	GSD	Korean	Korean	80K	0.08	2.03	0.28	0.00	0.13	3.83	0.46	6.81
Latvian	LVTB	Indo-European	Baltic	252K	1.26	4.68	4.01	0.00	1.39	2.63	1.91	15.87
Persian	PerDT	Indo-European	Iranian	494K	2.73	14.17	4.24	0.00	1.27	2.05	2.39	26.85
Russian	Taiga	Indo-European	Slavic	197K	0.30	8.56	4.12	0.00	0.41	2.49	1.63	17.51
Swedish	Talbanken	Indo-European	Germanic	97K	2.65	10.02	3.70	0.00	1.77	5.08	4.01	27.23
Turkish	Kenet	Turkic	Southwestern	179K	0.49	2.11	1.68	0.01	0.00	4.33	0.35	8.97
Vietnamese	VTB	Austro-Asiatic	Viet-Muong	44K	1.34	5.35	3.80	0.00	0.95	3.60	0.49	15.52
Wolof	WTB	Niger-Congo	Northern-Atlantic	43K	7.46	5.46	3.09	0.00	1.36	7.09	4.14	28.59
Average				168K	2.90	9.08	3.04	0.09	1.14	5.11	2.51	23.88

Data Sets

Language	Treebank	Family	Genus	Size	aux	case	cc	clf	cop	det	mark	Func
Arabic	PADT	Afro-Asiatic	Semitic	242K	0.60	14.29	5.11	0.00	0.16	0.76	2.71	23.63
Armenian	ArmTDP	Indo-European	Armenian	52K	5.04	3.03	4.10	0.00	2.01	3.46	1.67	19.30
Basque	BDT	Basque	Basque	121K	8.54	1.56	3.85	0.00	2.02	2.50	0.18	18.65
Chinese	GSD	Sino-Tibetan	Chinese	121K	1.83	6.31	1.42	1.82	1.45	1.35	5.75	19.93
Finnish	TDT	Uralic-Finnic	Finnish	202K	3.26	1.48	4.13	0.00	2.72	1.72	1.95	15.27
Greek	GDT	Indo-European	Greek	62K	3.81	8.47	3.19	0.00	0.94	19.12	1.83	37.37
Hebrew	HTB	Afro-Asiatic	Semitic	116K	0.45	16.26	2.93	0.00	0.69	11.55	3.32	35.19
Hindi	HDTB	Indo-European	Indic	352K	6.41	19.27	1.87	0.00	1.00	2.05	4.11	34.70
Indonesian	GSD	Austronesian	Malayo-Sumbawan	121K	0.00	9.87	2.96	0.00	0.87	3.71	1.31	18.72
Irish	IDT	Indo-European	Celtic	116K	0.00	13.44	3.14	0.00	1.32	8.15	5.79	31.84
Italian	ISDT	Indo-European	Romance	278K	2.77	14.01	2.73	0.00	1.15	16.30	2.11	39.08
Japanese	GSD	Japanese	Japanese	194K	8.90	21.34	0.42	0.00	1.26	0.49	4.06	36.47
Korean	GSD	Korean	Korean	80K	0.08	2.03	0.28	0.00	0.13	3.83	0.46	6.81
Latvian	LVTB	Indo-European	Baltic	252K	1.26	4.68	4.01	0.00	1.39	2.63	1.91	15.87
Persian	PerDT	Indo-European	Iranian	494K	2.73	14.17	4.24	0.00	1.27	2.05	2.39	26.85
Russian	Taiga	Indo-European	Slavic	197K	0.30	8.56	4.12	0.00	0.41	2.49	1.63	17.51
Swedish	Talbanken	Indo-European	Germanic	97K	2.65	10.02	3.70	0.00	1.77	5.08	4.01	27.23
Turkish	Kenet	Turkic	Southwestern	179K	0.49	2.11	1.68	0.01	0.00	4.33	0.35	8.97
Vietnamese	VTB	Austro-Asiatic	Viet-Muong	44K	1.34	5.35	3.80	0.00	0.95	3.60	0.49	15.52
Wolof	WTB	Niger-Congo	Northern-Atlantic	43K	7.46	5.46	3.09	0.00	1.36	7.09	4.14	28.59
Average				168K	2.90	9.08	3.04	0.09	1.14	5.11	2.51	23.88

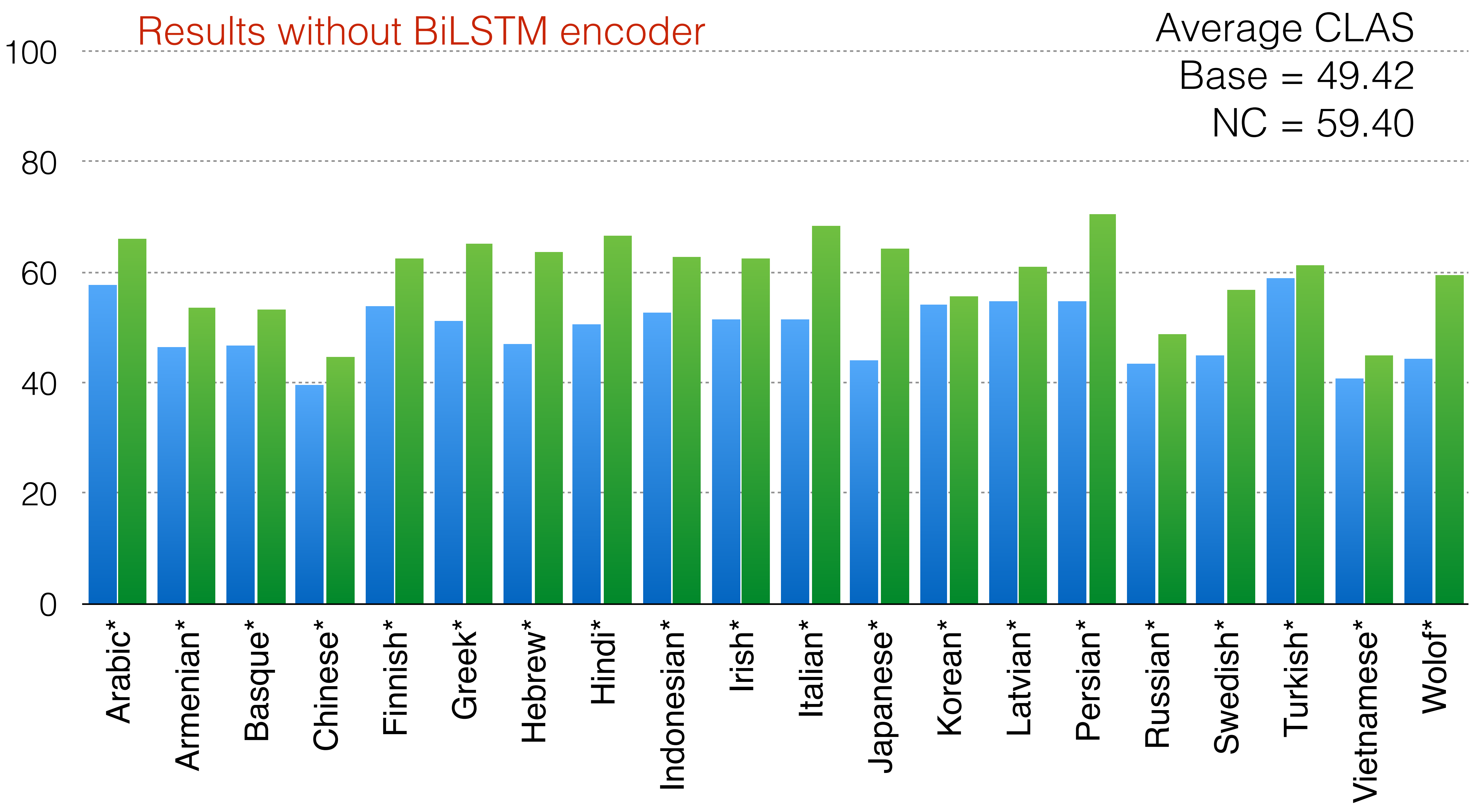
Experimental Results



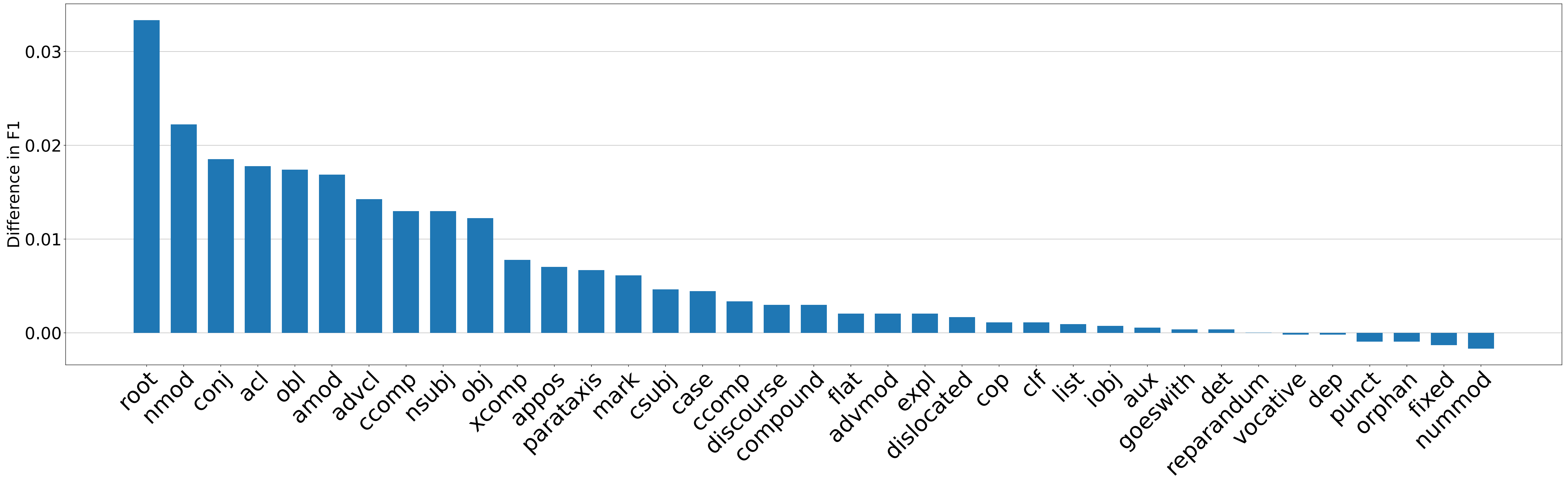
Analysis

- Why does composition give such modest improvements?
- Which linguistic relations benefit the most?
- Why is composition more effective in certain languages?
- What information is captured in composition?

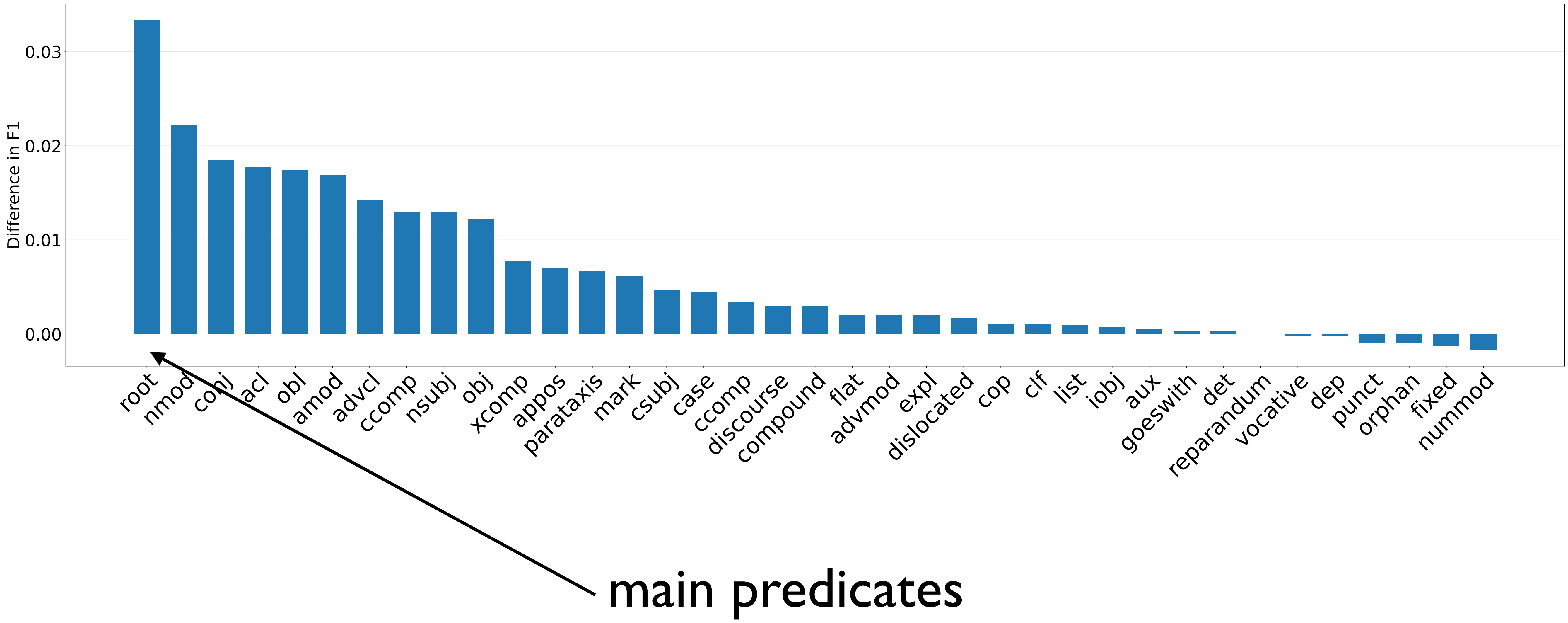
Ablation: No BiLSTM



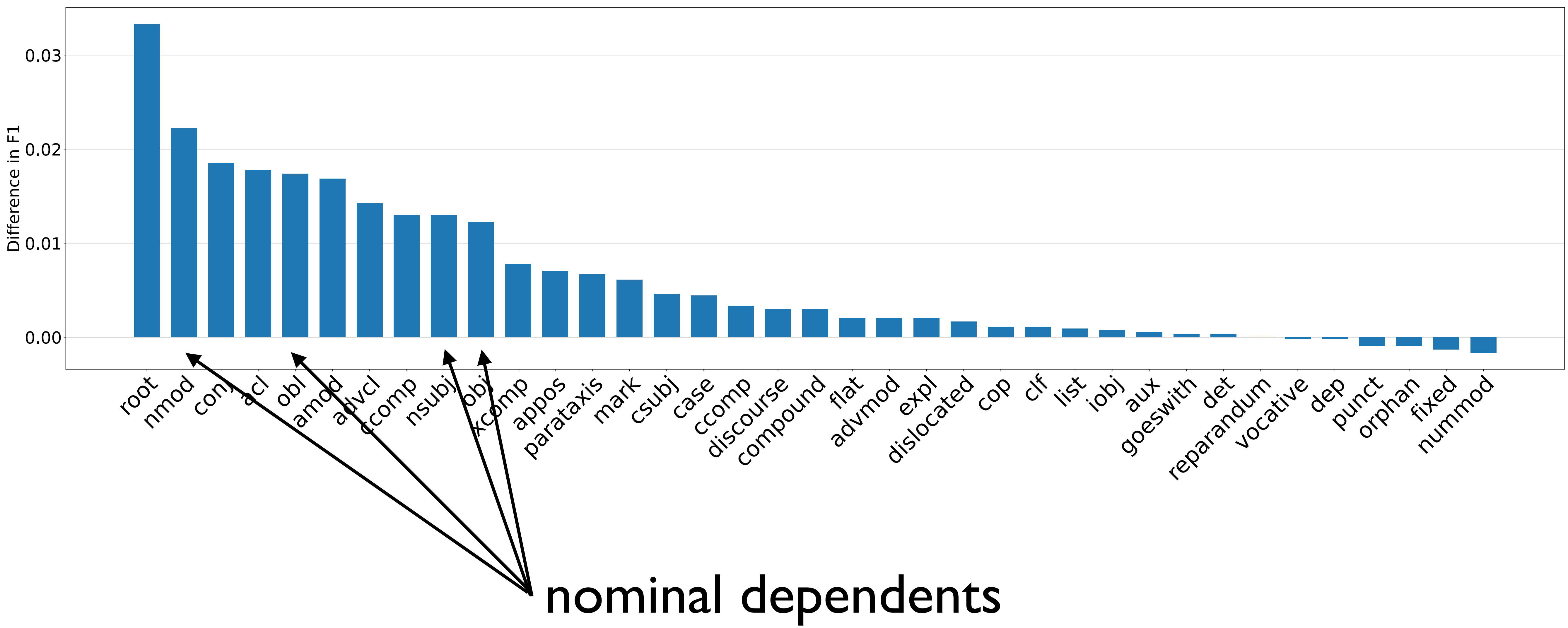
Improvement per Relation



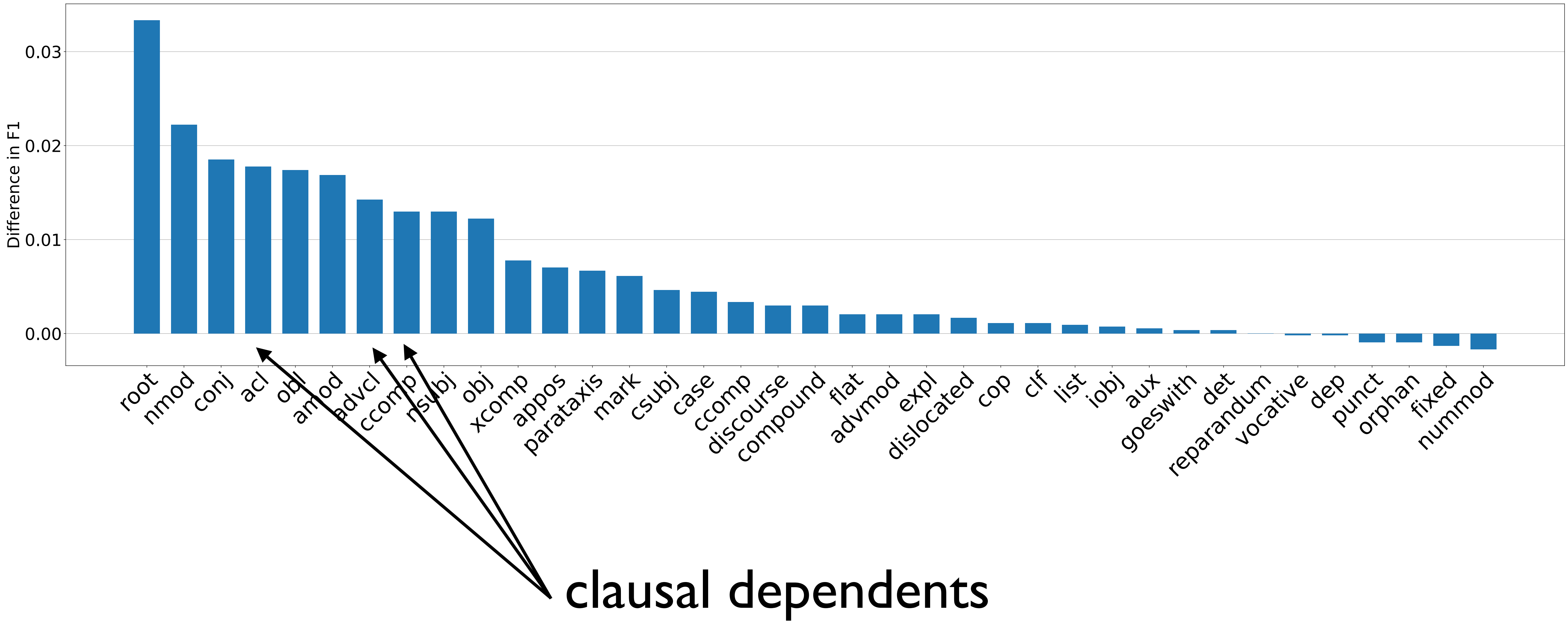
Improvement per Relation



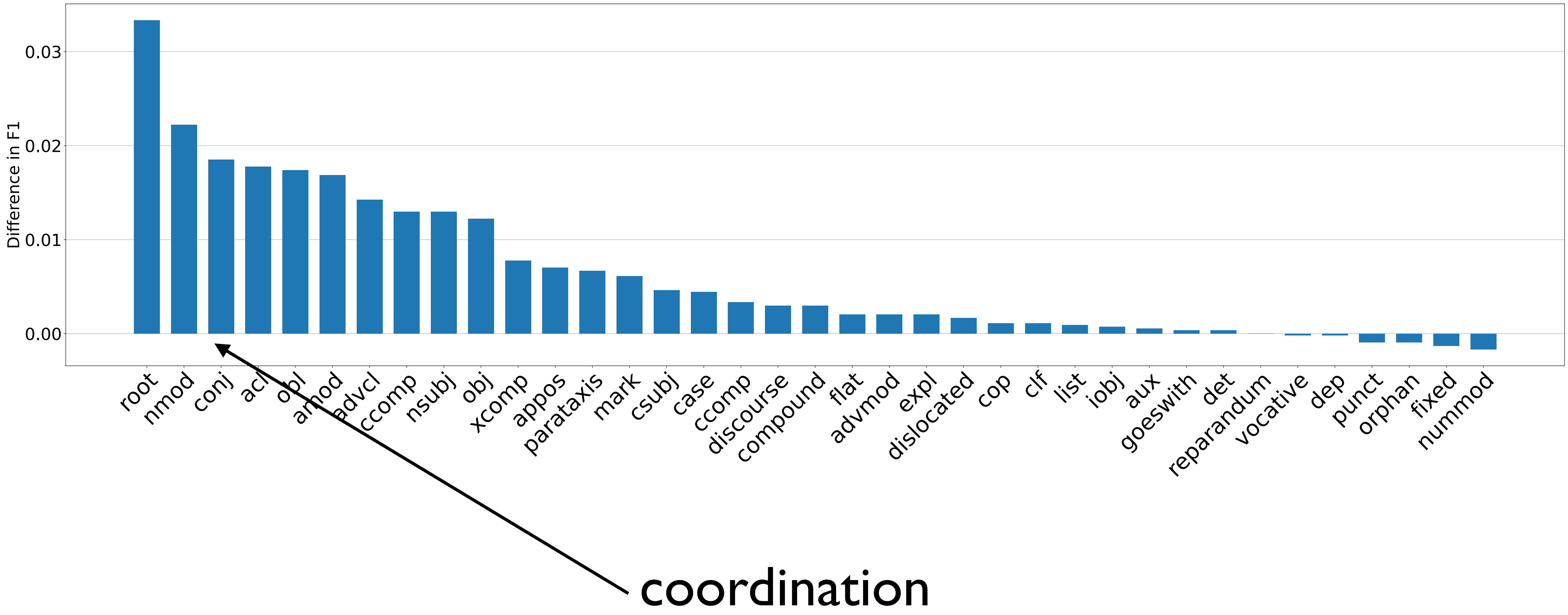
Improvement per Relation



Improvement per Relation



Improvement per Relation



Cross-Linguistic Variation

- Can we predict improvement rates across languages?
- Linear-mixed effects models of CLAS improvement

Cross-Linguistic Variation

- Can we predict improvement rates across languages?
- Linear-mixed effects models of CLAS improvement

Standard Model

Predictors	Estimates	CI	p
(Intercept)	0.65	0.56 - 0.76	$<$**0.001**
det frequency	0.59	0.20 - 0.98	**0.003**
cc rel entropy	0.77	0.27 - 1.26	**0.003**
cc POS entropy	0.79	0.30 - 1.28	**0.002**
Random Effects			
σ^2	0.17		
$\tau_{00\ \text{language}}$	0.01		
ICC	0.07		
N_{language}	20		
Observations	100		
Marginal R^2/Conditional R^2		0.266/0.315	

Cross-Linguistic Variation

- Can we predict improvement rates across languages?
- Linear-mixed effects models of CLAS improvement

Standard Model

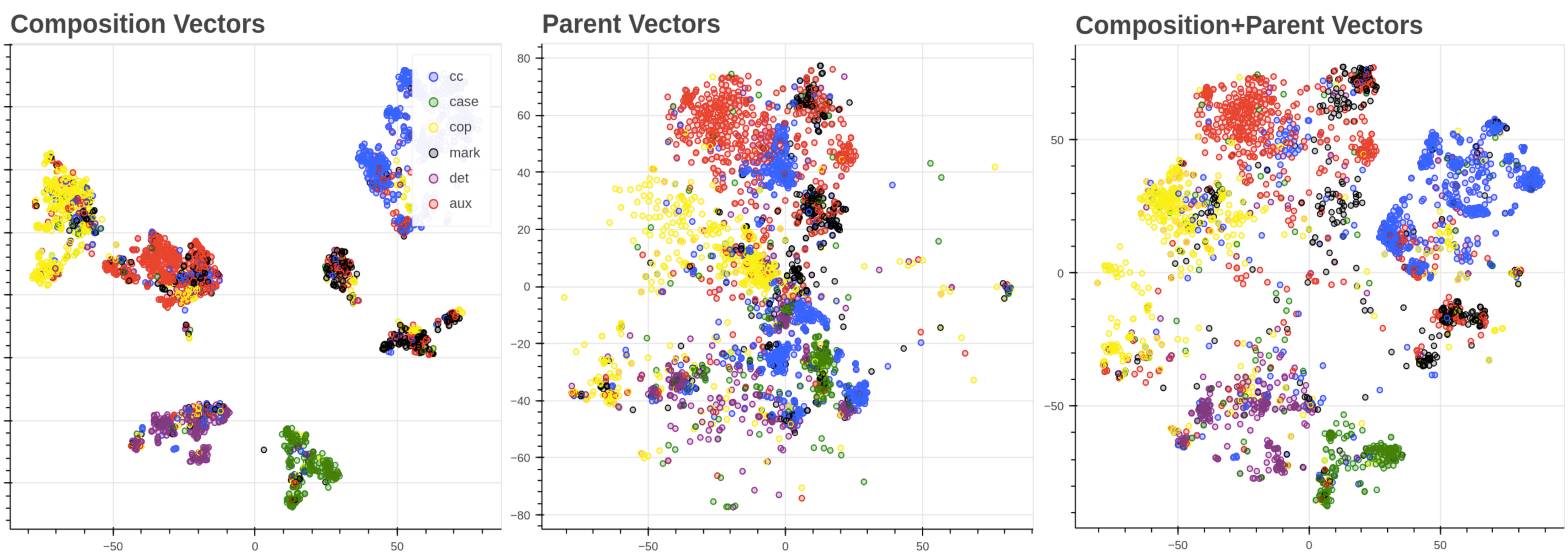
Predictors	Estimates	CI	p
(Intercept)	0.65	0.56 - 0.76	<**0.001**
det frequency	0.59	0.20 - 0.98	**0.003**
cc rel entropy	0.77	0.27 - 1.26	**0.003**
cc POS entropy	0.79	0.30 - 1.28	**0.002**
Random Effects			
σ^2	0.17		
$\tau_{00\ language}$	0.01		
ICC	0.07		
$N_{language}$	20		
Observations	100		
Marginal R^2/Conditional R^2		0.266/0.315	

Model without BILSTM

Predictors	Estimates	CI	p
(Intercept)	9.99	9.31 - 10.66	<**0.001**
det frequency	6.06	3.28 - 8.84	<**0.001**
cop frequency	4.25	1.98 - 6.52	<**0.001**
aux frequency	3.83	1.49 - 6.17	**0.002**
case dep length	1.63	-0.34 - 3.60	0.104
case frequency	14.04	11.66 - 16.42	<**0.001**
Random Effects			
σ^2	0.27		
$\tau_{00\ language}$	2.28		
ICC	0.89		
$N_{language}$	20		
Observations	100		
Marginal R^2/Conditional R^2		0.900/0.989	

Visualising Composition

- Diagnostic classifiers to predict categories and relations
- Dimensionality reduction and visualisation



$g(h, d, l)$ h $h + g(h, d, l)$

Conclusion

- Syntactic nuclei as elementary syntactic units increase cross-language similarity
- Syntactic nuclei can be (roughly) defined in the Universal Dependencies framework
- Syntactic nuclei can be represented in a transition-based parser using nucleus composition

Conclusion

- Small but consistent improvements for most languages – largely redundant together with contextual encoders
- Improved accuracy for main predicates, clausal dependents, nominal dependents, and coordination
- Significant factors explaining rate of improvement are entropy in coordination and frequency of function words
- Nucleus composition appears to increase similarity of vectors representing nuclei of the same syntactic type