UPPSALA
UNIVERSITET

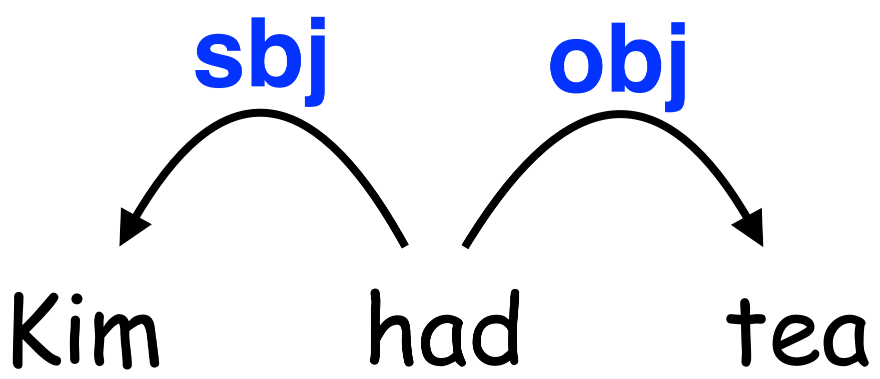
Multilingual Dependency Parsing

From Universal Dependencies to Sesame Street

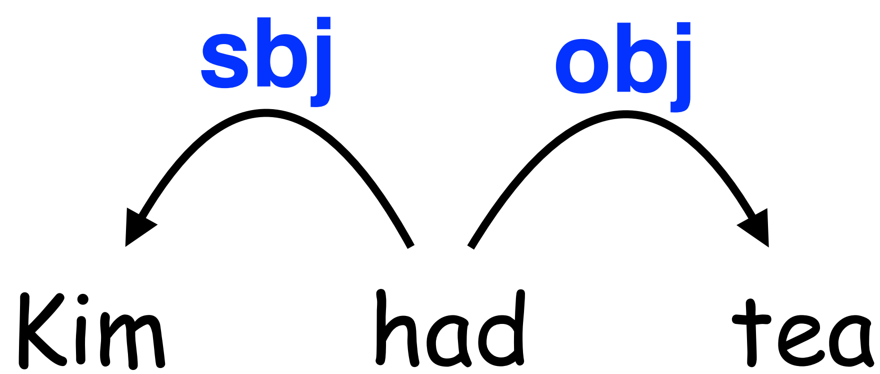
Joakim Nivre

Uppsala University
Department of Linguistics and Philology

Dependency Parsing

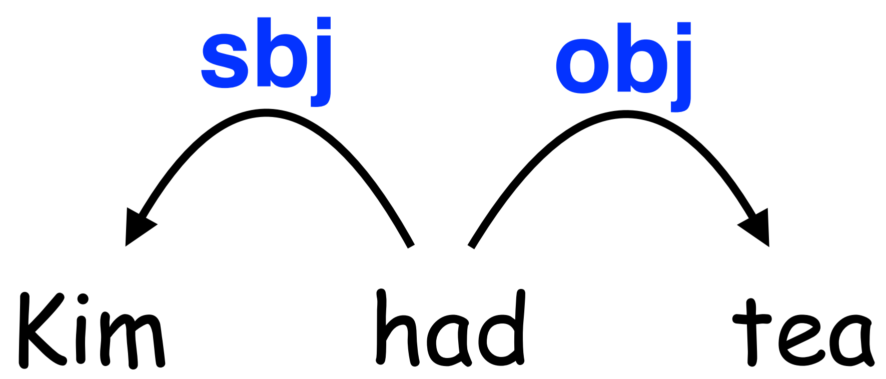


Dependency Parsing



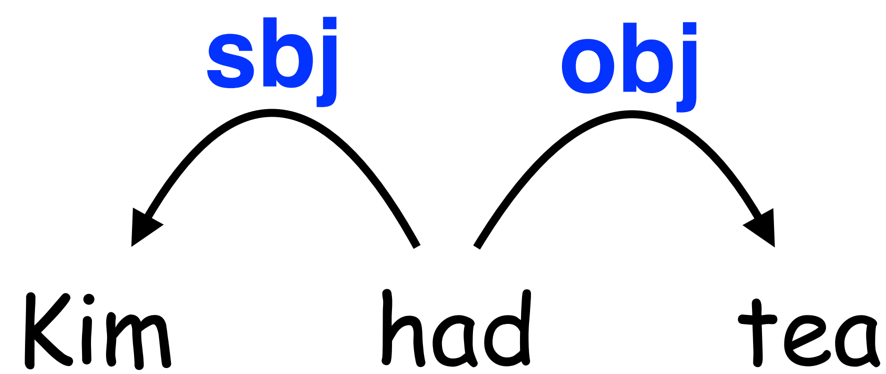
- Transparent encoding of predicate-argument structure

Dependency Parsing



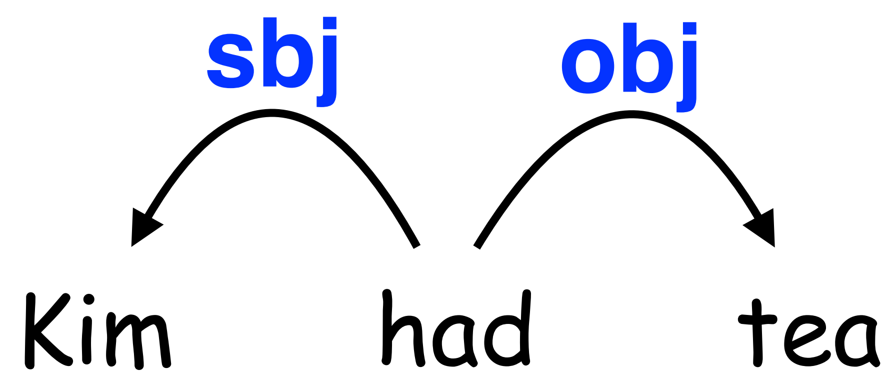
- Transparent encoding of predicate-argument structure
- Simple and efficient computational models

Dependency Parsing



- Transparent encoding of predicate-argument structure
- Simple and efficient computational models
- Compatible with linguistic traditions around the world

Dependency Parsing



- Transparent encoding of predicate-argument structure
- Simple and efficient computational models
- Compatible with linguistic traditions around the world
- Multilingual research tradition from CoNLL 2006–2007

CoNLL-X Shared Task

Sabine Buchholz

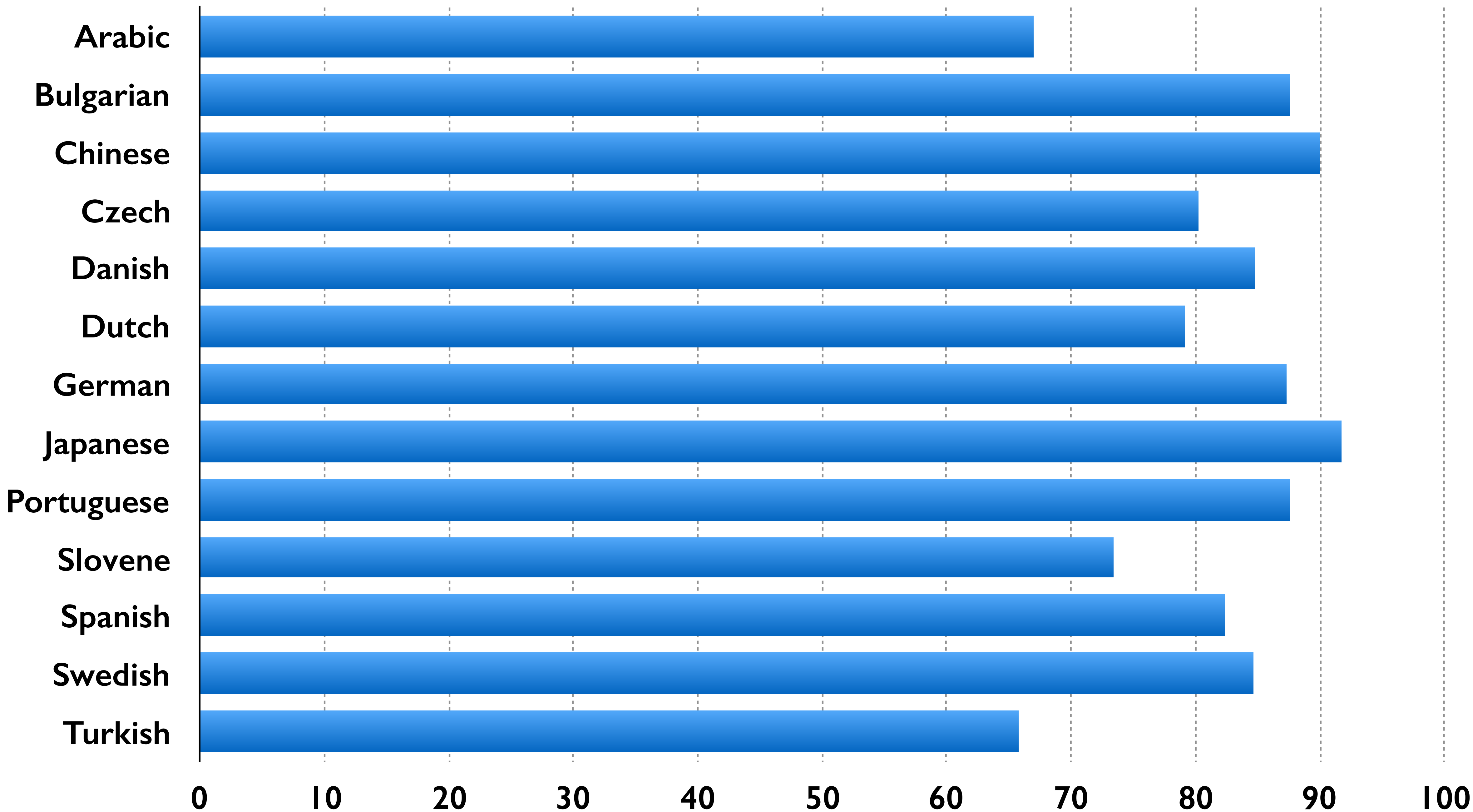
Amit Dubey

Yuval Krymolowski

Erwin Marsi

- First shared task on multilingual dependency parsing
- Data from heterogeneous treebanks in 13 languages
- Standardized into a single unified format (CoNLL-X)
- Enabled a new line of multilingual research

CoNLL-X Results



Why the Differences?

Why the Differences?

- Amount of data – weak predictor overall

Why the Differences?

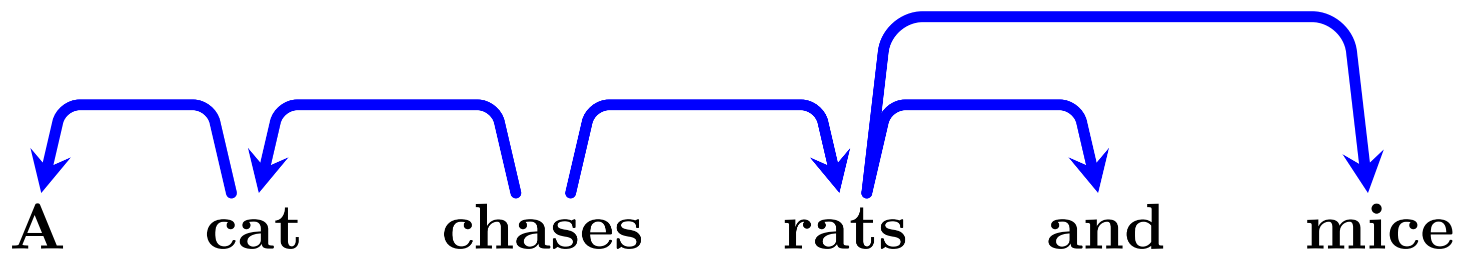
- Amount of data – weak predictor overall
- Text types – important but hard to measure

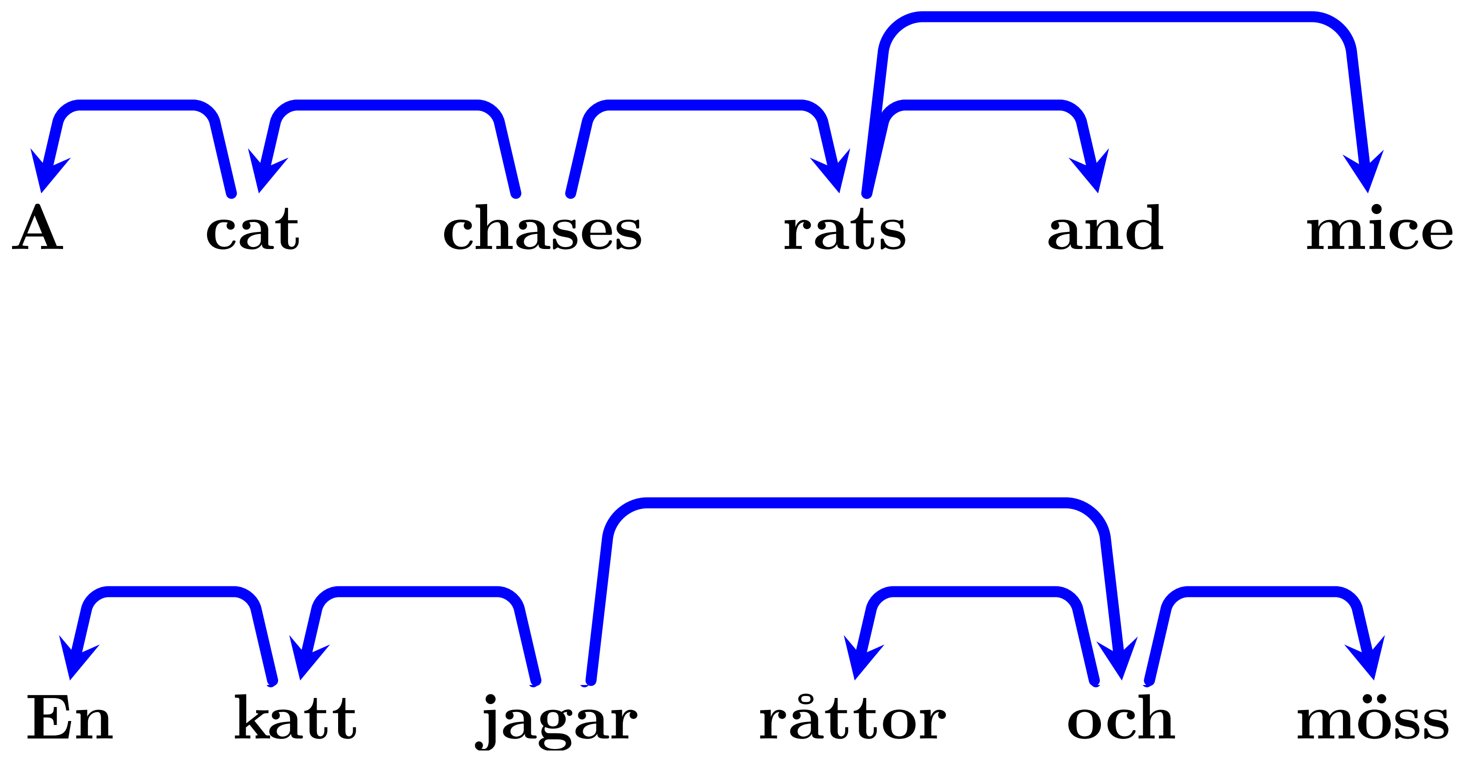
Why the Differences?

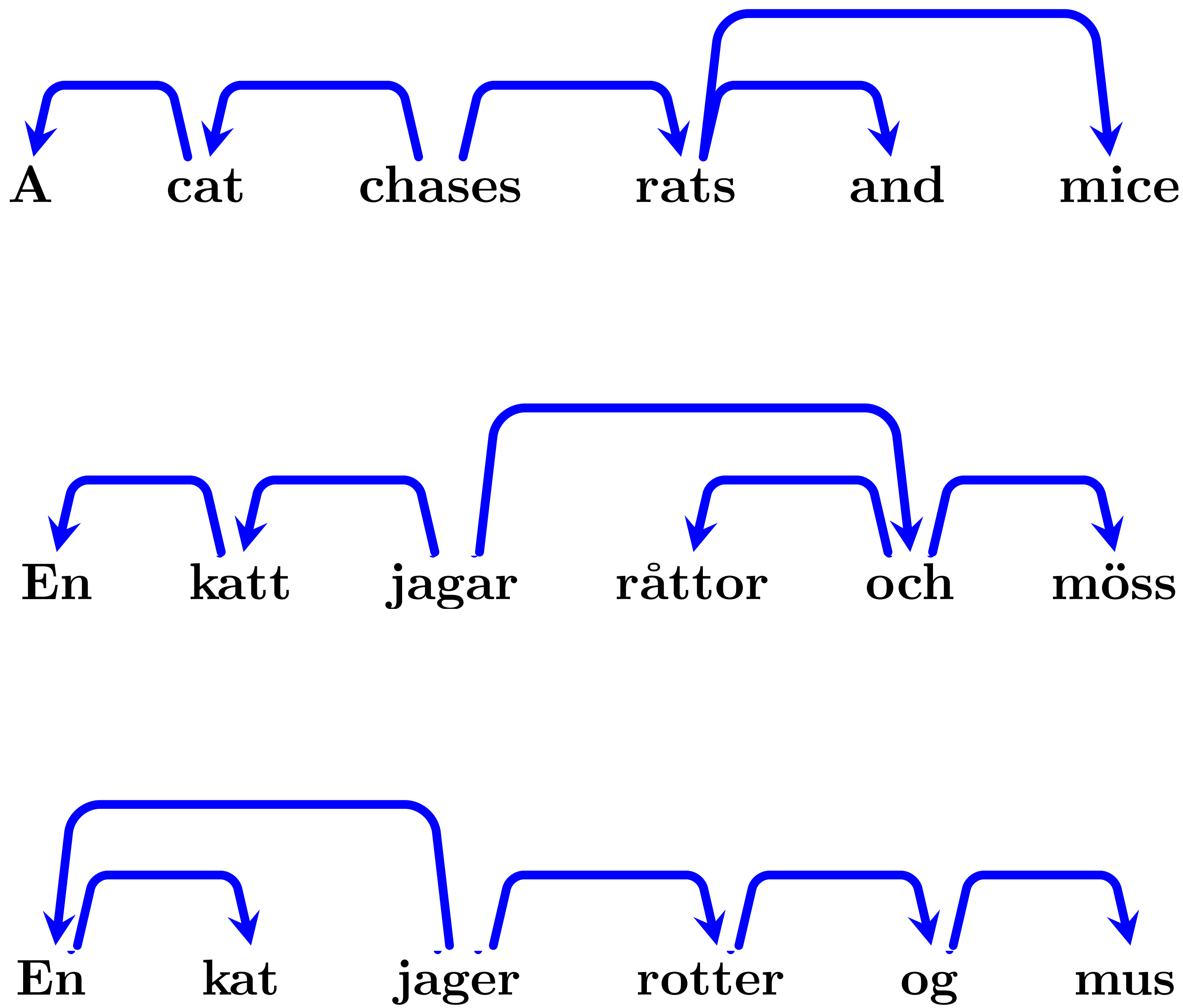
- Amount of data – weak predictor overall
- Text types – important but hard to measure
- Language types – analytical versus synthetic

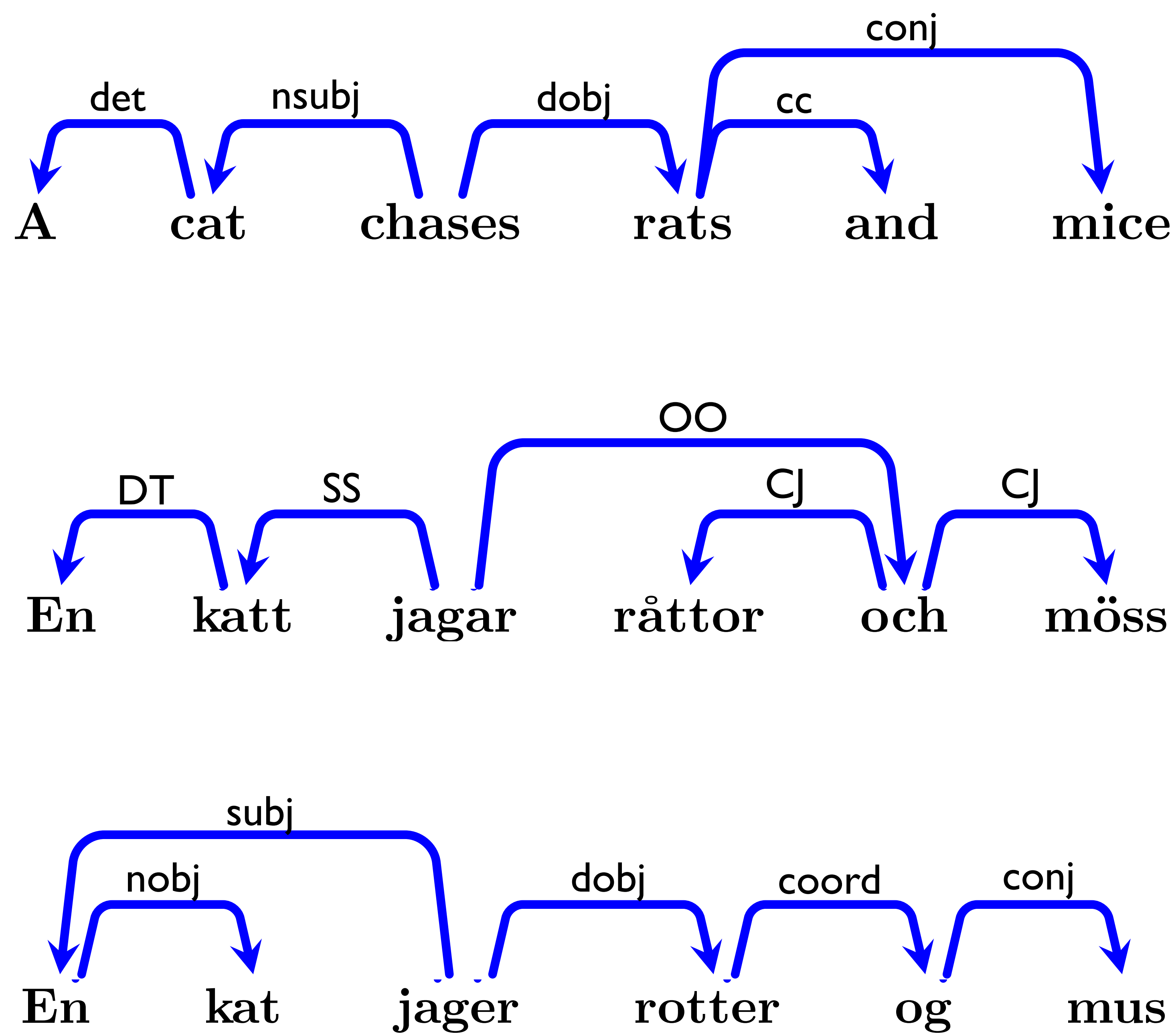
Why the Differences?

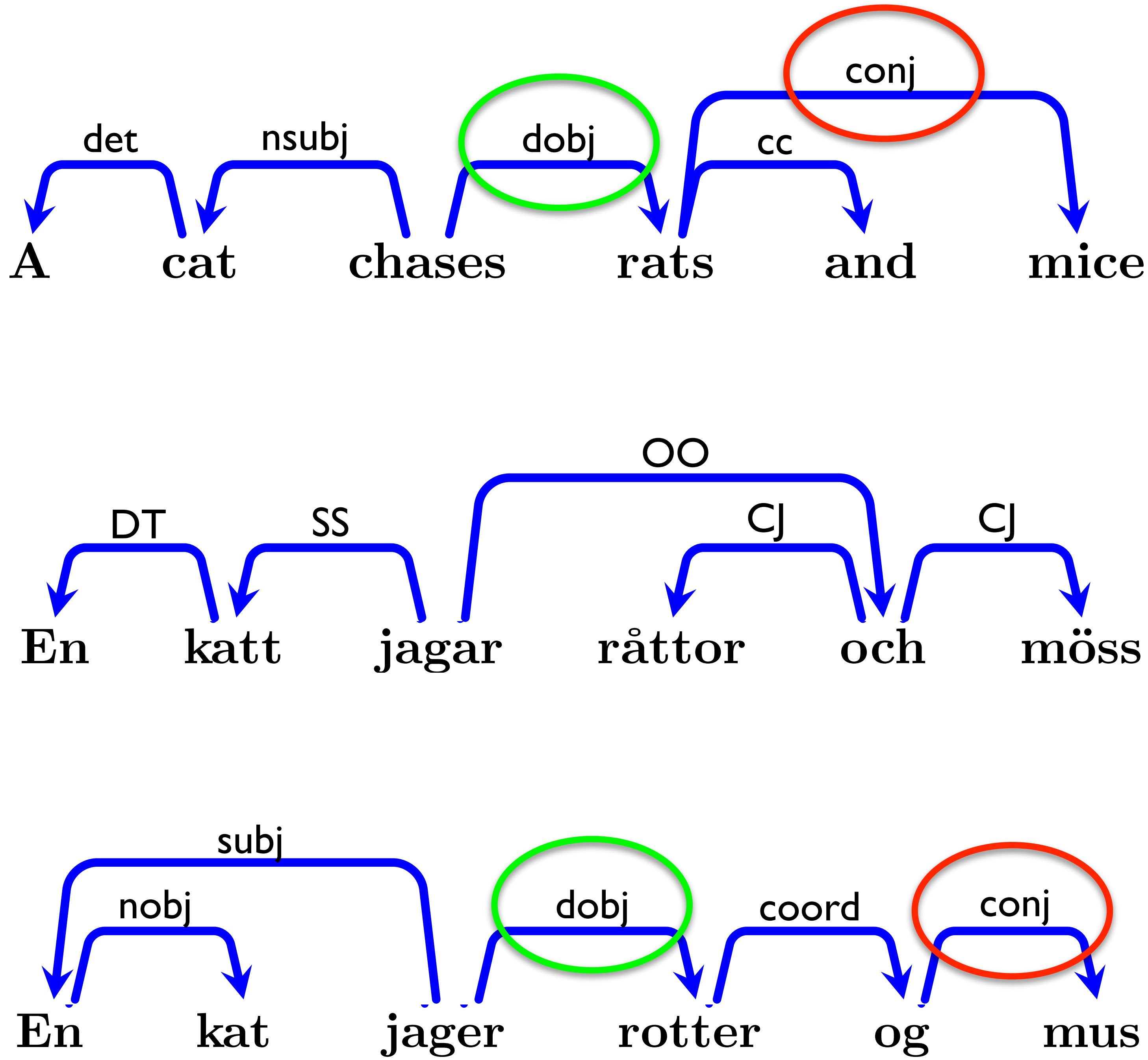
- Amount of data – weak predictor overall
- Text types – important but hard to measure
- Language types – analytical versus synthetic
- Annotation – different descriptive traditions


A
cat
chases
rats
and
mice


A cat chases rats and mice
En katt jagar råttor och möss

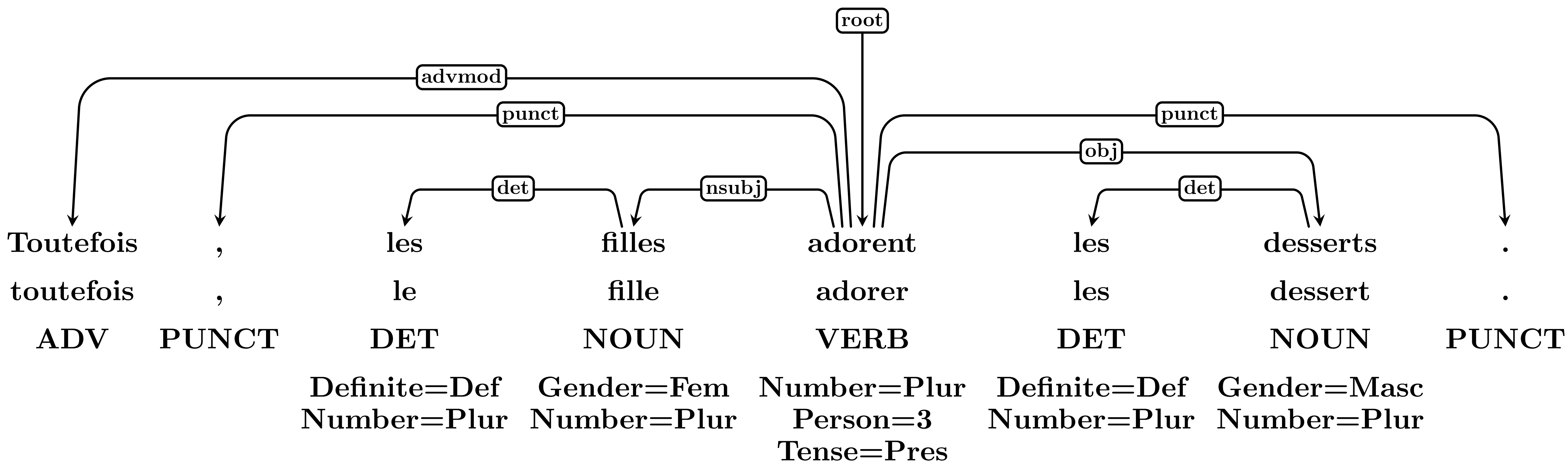

A cat chases rats and mice
En katt jagar råttor och möss
En kat jager rotter og mus


conj
det
nsubj
dobj
cc
A
cat
chases
rats
and
mice
OO
DT
SS
CJ
CJ
En
katt
jagar
råttor
och
möss
subj
nobj
dobj
coord
conj
En
kat
jager
rotter
og
mus


conj
det
nsubj
dobj
cc
A
cat
chases
rats
and
mice
OO
DT
SS
CJ
CJ
En
katt
jagar
råttor
och
möss
subj
nobj
dobj
coord
conj
En
kat
jager
rotter
og
mus

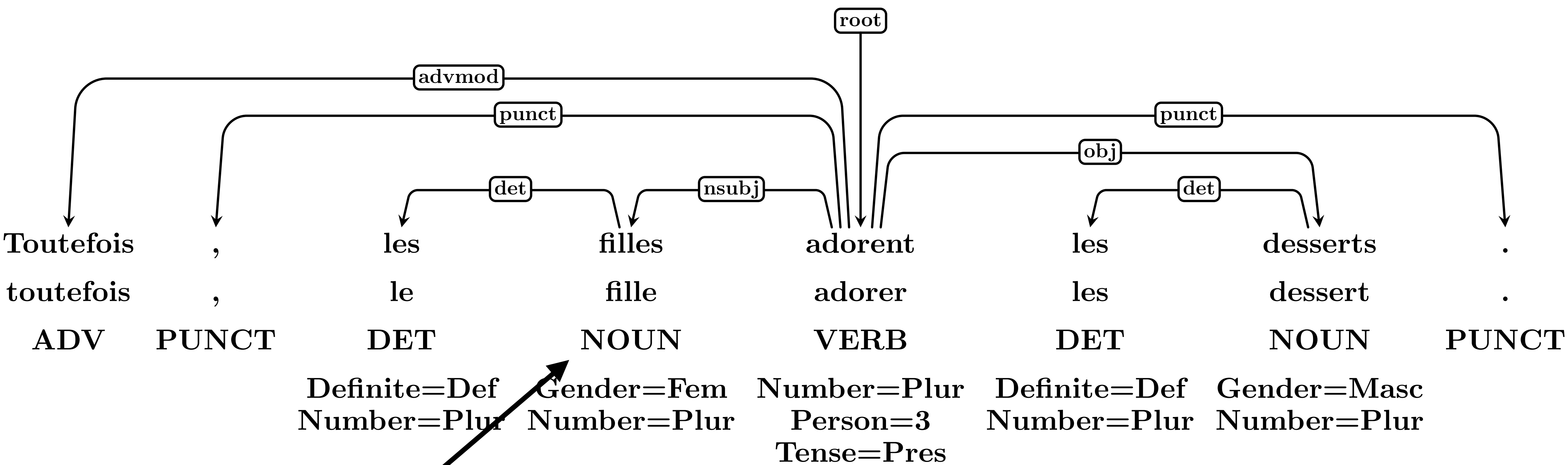
Universal Dependencies

http://universaldependencies.org



Universal Dependencies

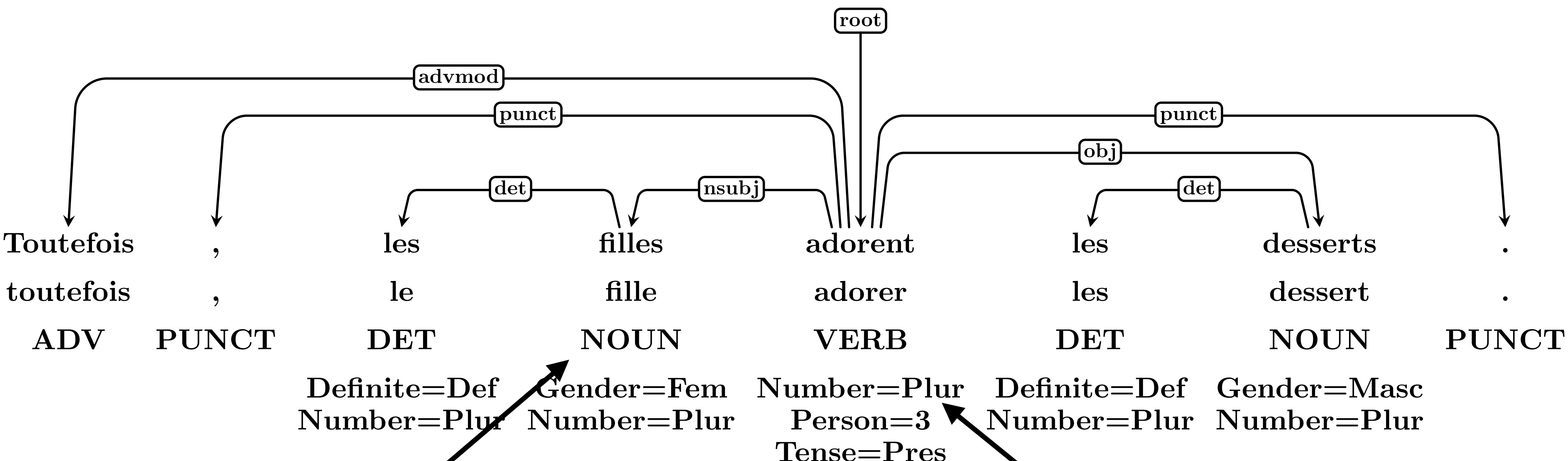
http://universaldependencies.org



Part-of-speech tags Google

Universal Dependencies

http://universaldependencies.org

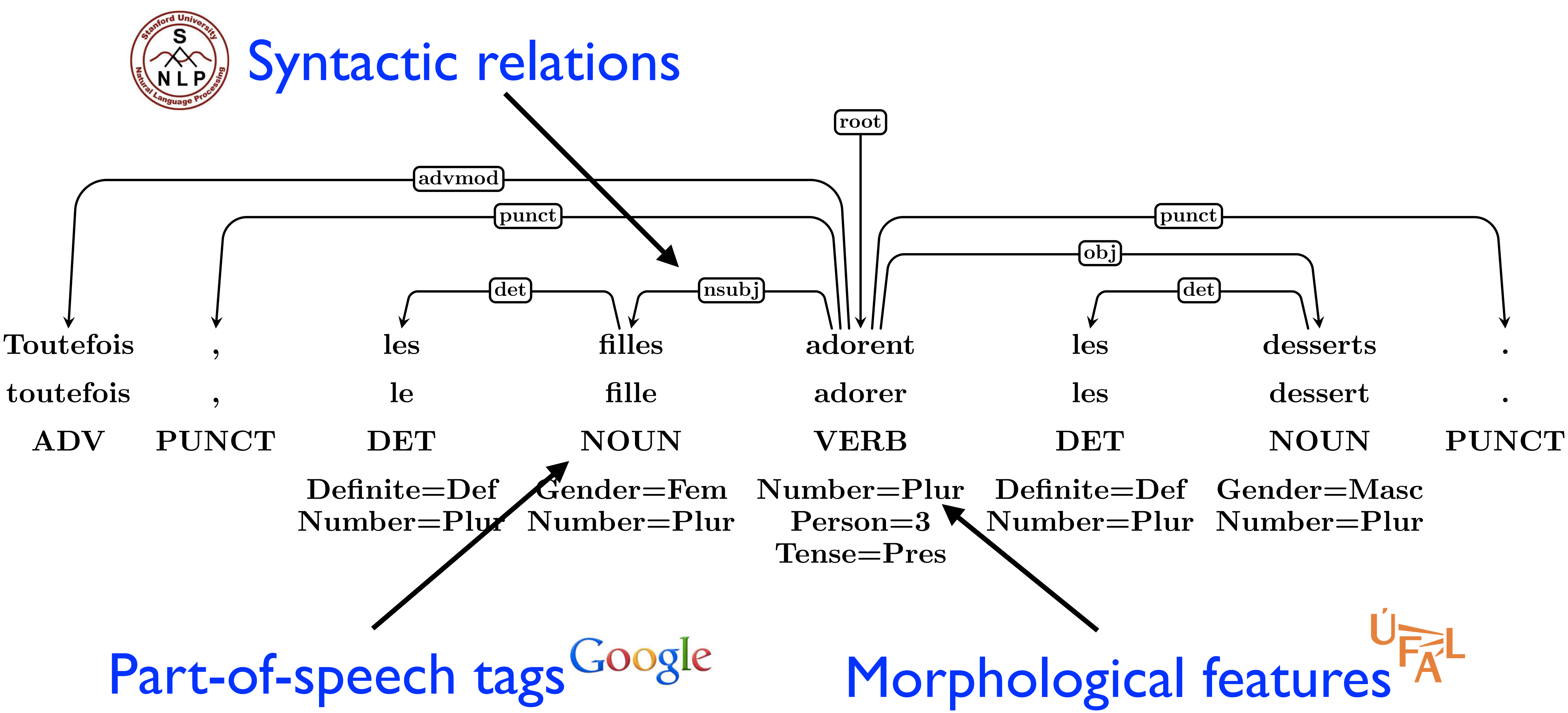


Part-of-speech tags Google

Morphological features ÚFAL

Universal Dependencies

http://universaldependencies.org



Who?

Open community effort – a big tent

UD v2.5: 90 languages, 157 treebanks, 345 contributors

Come join us at http://universaldependencies.org

Marie
de Marneffe

Filip
Ginter

Yoav
Goldberg

Jan
Hajič

Chris
Manning

Ryan
McDonald

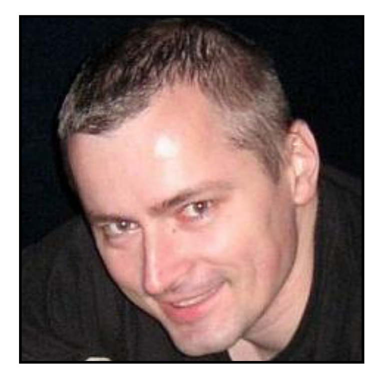
Slav
Petrov

Sampo
Pyysalo

Sebastian
Schuster

Reut
Tsarfaty

Francis
Tyers



Dan
Zeman

Why?

Why?

Cross-linguistically consistent morphosyntactic annotation

Why?

Cross-linguistically consistent morphosyntactic annotation

Facilitate multilingual research in NLP and linguistics

- Meaningful linguistic analysis across languages
- Syntactic parsing in multilingual settings
- NLP systems for multiple languages
- Facilitate resource-building for new languages

Why?

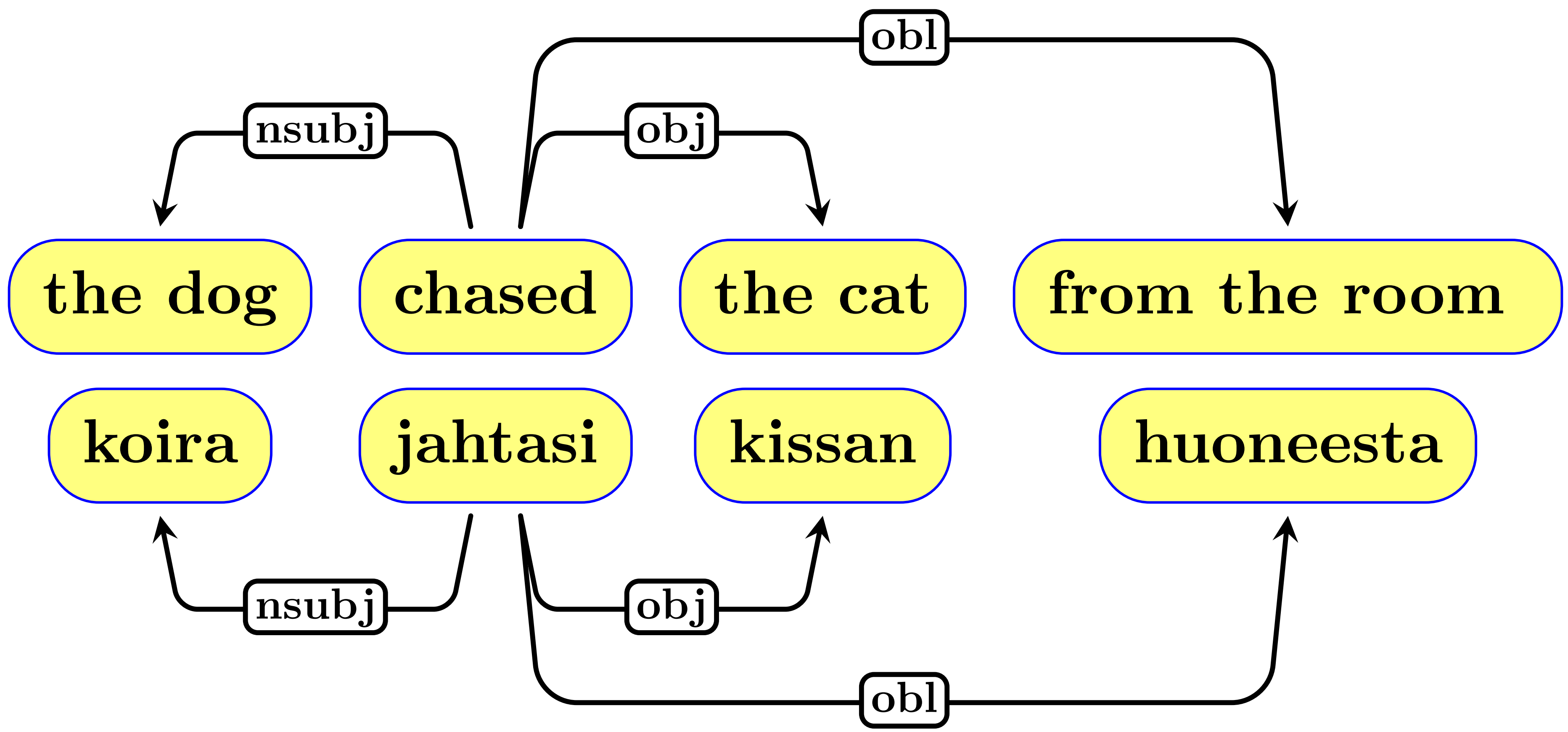
Cross-linguistically consistent morphosyntactic annotation

Facilitate multilingual research in NLP and linguistics

- Meaningful linguistic analysis across languages
- Syntactic parsing in multilingual settings
- NLP systems for multiple languages
- Facilitate resource-building for new languages

Complement – not replace – language-specific schemes

How?



Focus on grammatical relations between (content) words

Morphology

Le	**chat**	**chasse**	**les**	**chiens**	**.**

Morphology

Le	**chat**	**chasse**	**les**	**chiens**	**.**
le	**chat**	**chasser**	**le**	**chien**	**.**

- Lemma representing the semantic content of the word

Morphology

Le	**chat**	**chasse**	**les**	**chiens**	**.**
le	**chat**	**chasser**	**le**	**chien**	**.**
DET	**NOUN**	**VERB**	**DET**	**NOUN**	**PUNCT**

- Lemma representing the semantic content of the word
- Part-of-speech tag representing its grammatical class

Morphology

Le			iens	.
le			hien	.
DET	N		OUN	**PUNCT**

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

- Lemma rep of the word
- Part-of-speech tag representing its grammatical class

Morphology

Le	chat	chasse	les	chiens	.
le	chat	chasser	le	chien	.
DET	**NOUN**	**VERB**	**DET**	**NOUN**	**PUNCT**
Definite=Def **Gender=Masc** **Number=Sing**	**Gender=Masc** **Number=Sing**	**Mood=Ind** **Number=Sing** **Person=3** **Tense=Pres** **VerbForm=Fin**	**Definite=Def** **Gender=Masc** **Number=Plur**	**Gender=Masc** **Number=Plur**	

- Lemma representing the semantic content of the word
- Part-of-speech tag representing its grammatical class
- Features representing lexical and grammatical properties of the lemma or the particular word form

Morphology

Le	...	...iens	.
le	...	...hien	.
DET	**N...**	**...OUN**	**PUNCT**
Definite=Def Gender=Masc Number=Sing	Gend... Num...	...er=Masc ...er=Plur	

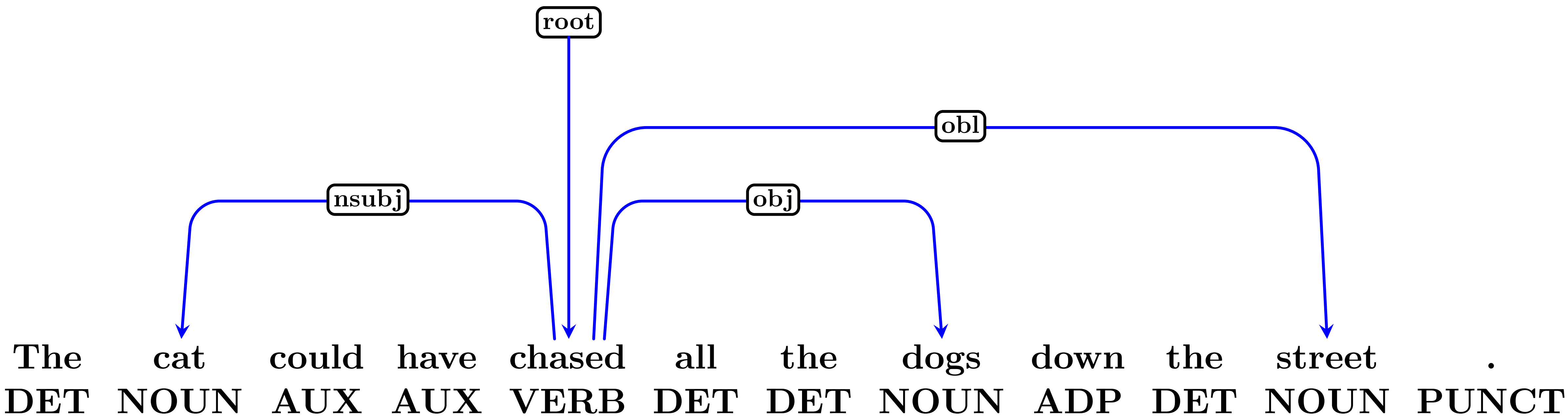
Lexical	Inflectional Nominal	Inflectional Verbal
PronType	Gender	VerbForm
NumType	Animacy	Mood
Poss	Number	Tense
Reflex	Case	Aspect
Foreign	Definite	Voice
Abbr	Degree	Evident
		Polarity
		Person
		Polite

- Lemma rep... of the word
- Part-of-spe... tical class
- Features representing lexical and grammatical properties of the lemma or the particular word form

Syntax

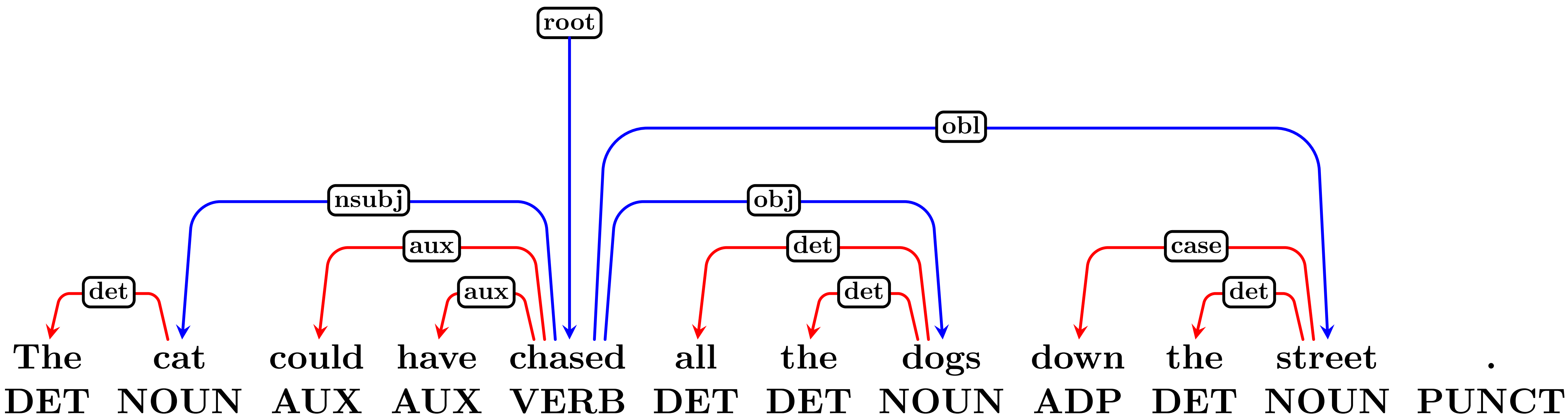
The	cat	could	have	chased	all	the	dogs	down	the	street	.
DET	**NOUN**	**AUX**	**AUX**	**VERB**	**DET**	**DET**	**NOUN**	**ADP**	**DET**	**NOUN**	**PUNCT**

Syntax



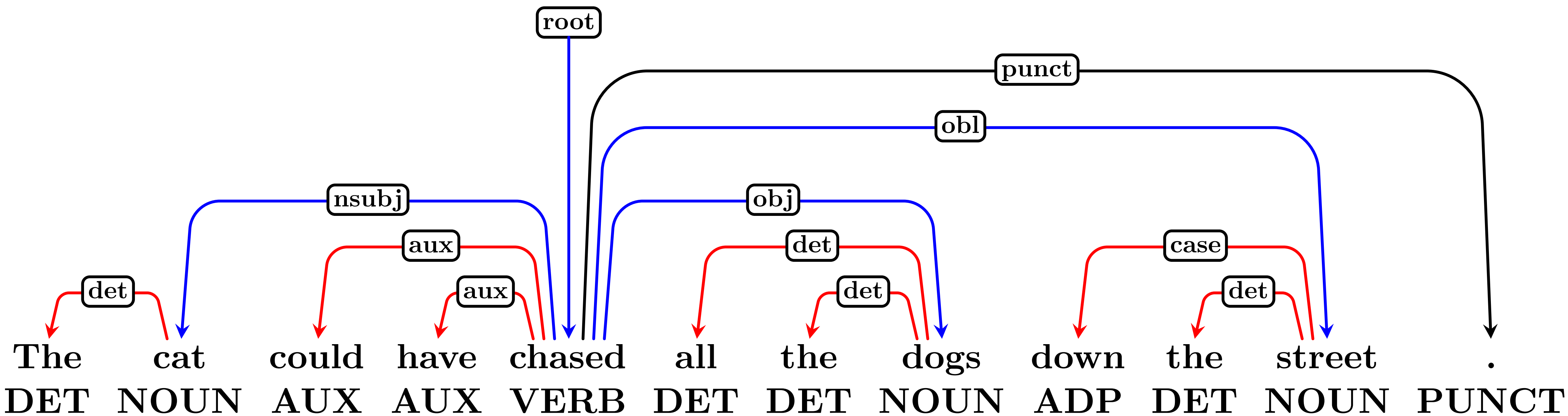
- Content words are linked by grammatical relations

Syntax



- Content words are linked by grammatical relations
- Function words attach to the content word they modify

Syntax

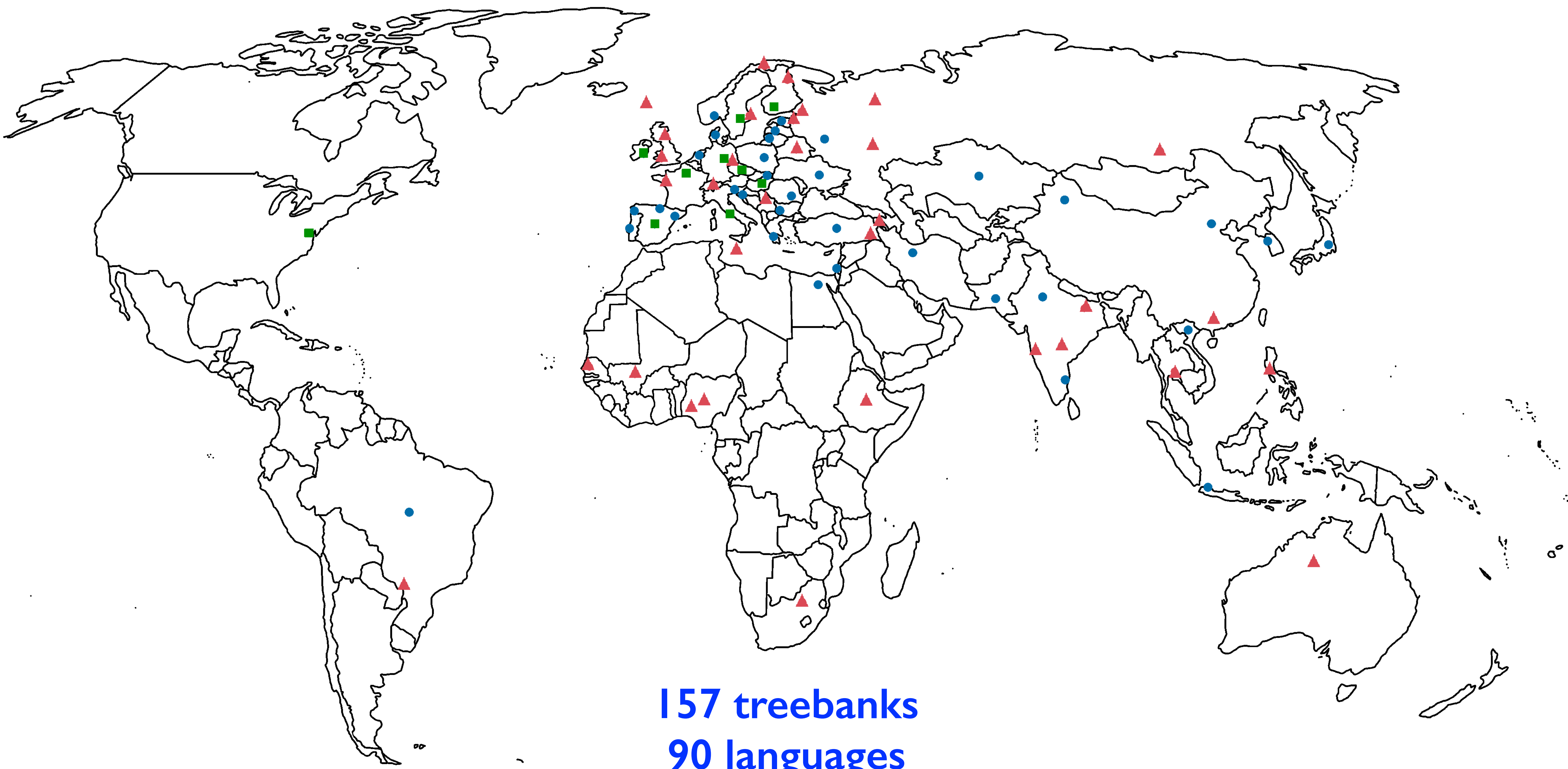


- Content words are linked by grammatical relations
- Function words attach to the content word they modify
- Punctuation attach to head of phrase or clause

Syntax

	Nominal	Clause	Modifier Word	Function Word
Core Predicate Dep	nsubj obj iobj	csubj ccomp xcomp		
Non-Core Predicate Dep	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal Dep	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	parataxis list	orphan goeswith reparandum	punct root dep

UD v2.5



157 treebanks
90 languages
20 language families

Figure by Francis Tyers

CoNLL Shared Tasks 2017–18

Multilingual Parsing from Raw Text to Universal Dependencies

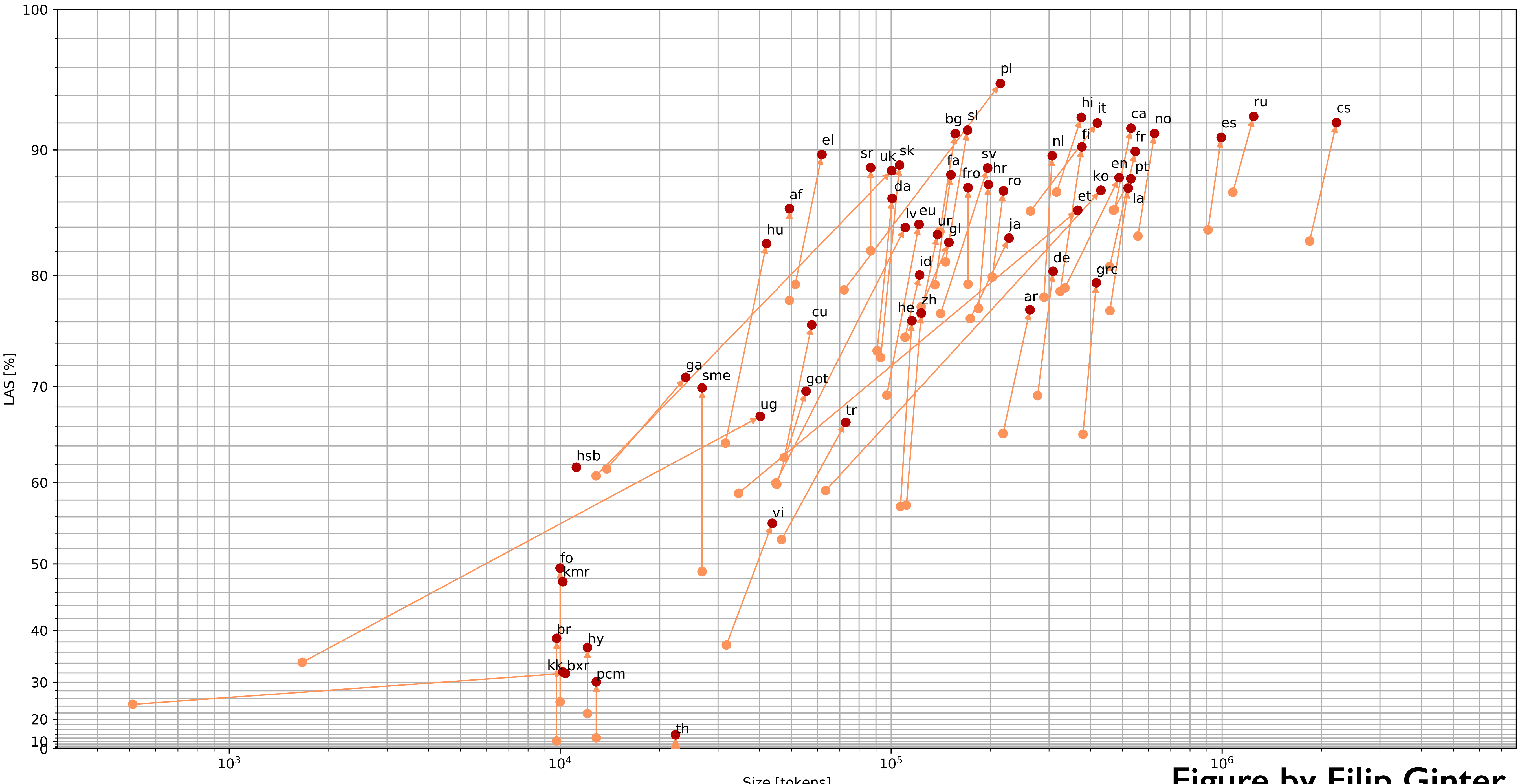


Figure by Filip Ginter

Three Case Studies

- Representing words – characters, words, parts of speech
- Adding deep contextualized word representations
- Probing deep contextualized word representations

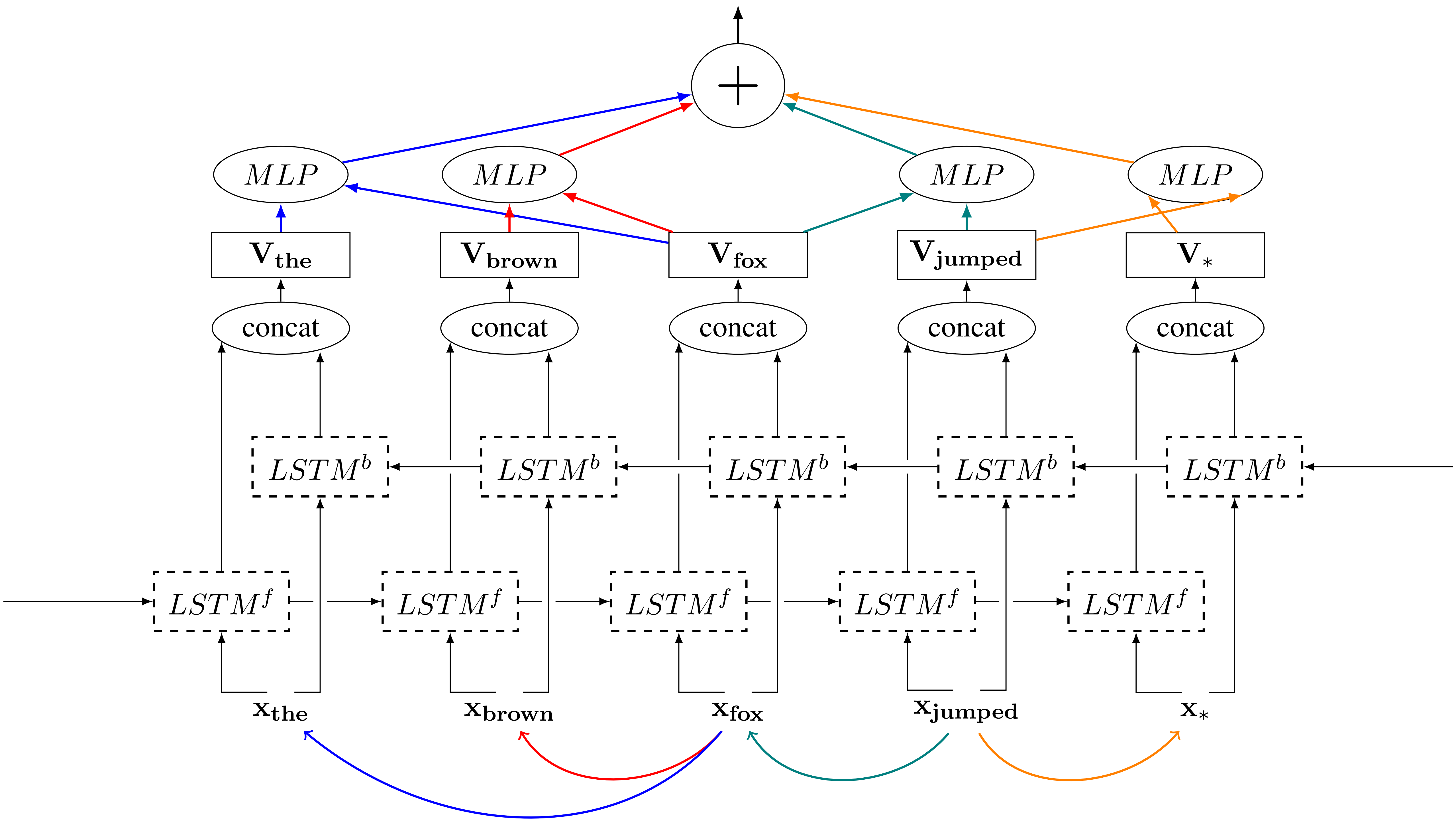
Uppsala Parsing Group


Sara Stymne

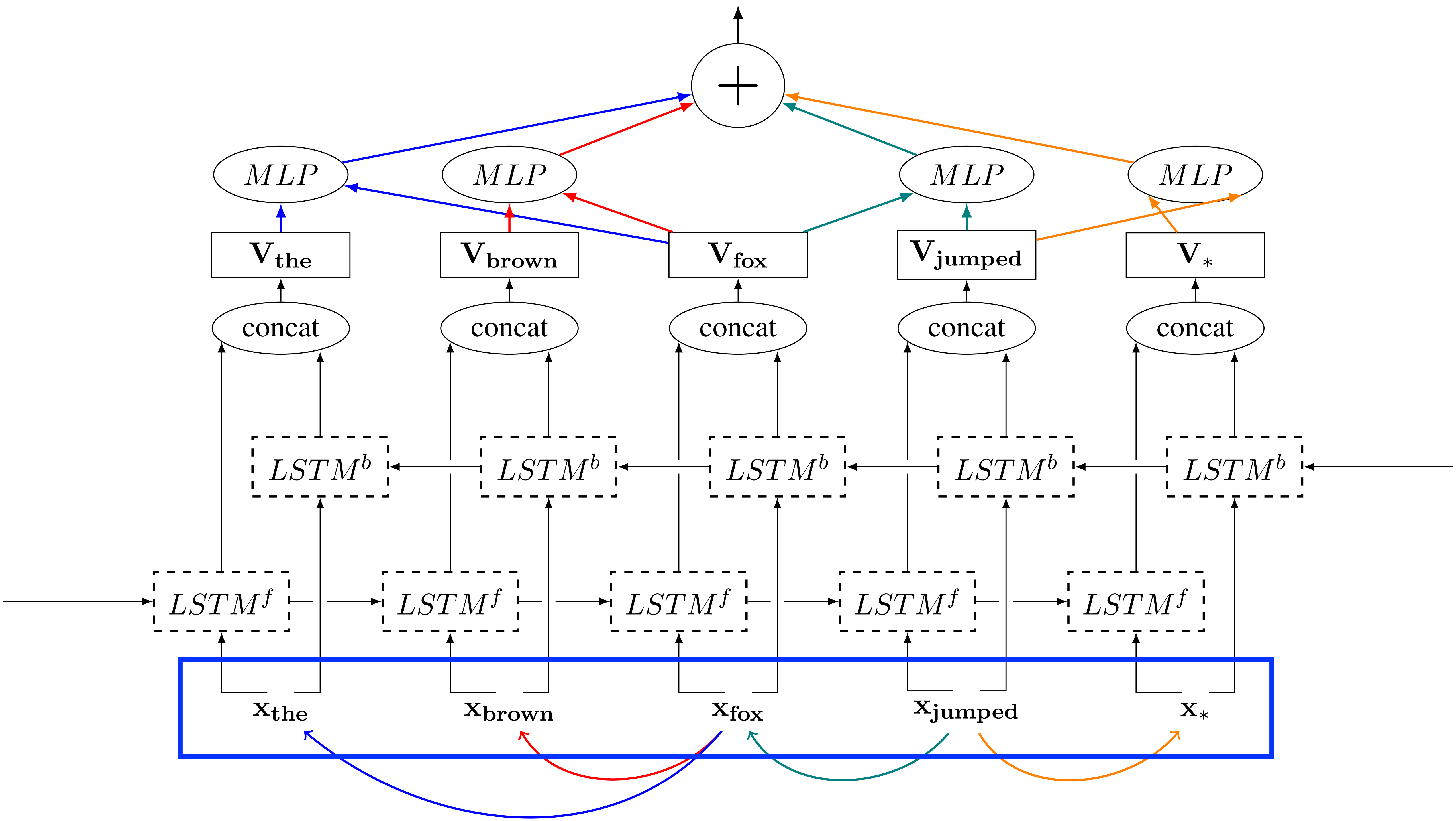

Aaron Smith


Miryam de Lhoneux

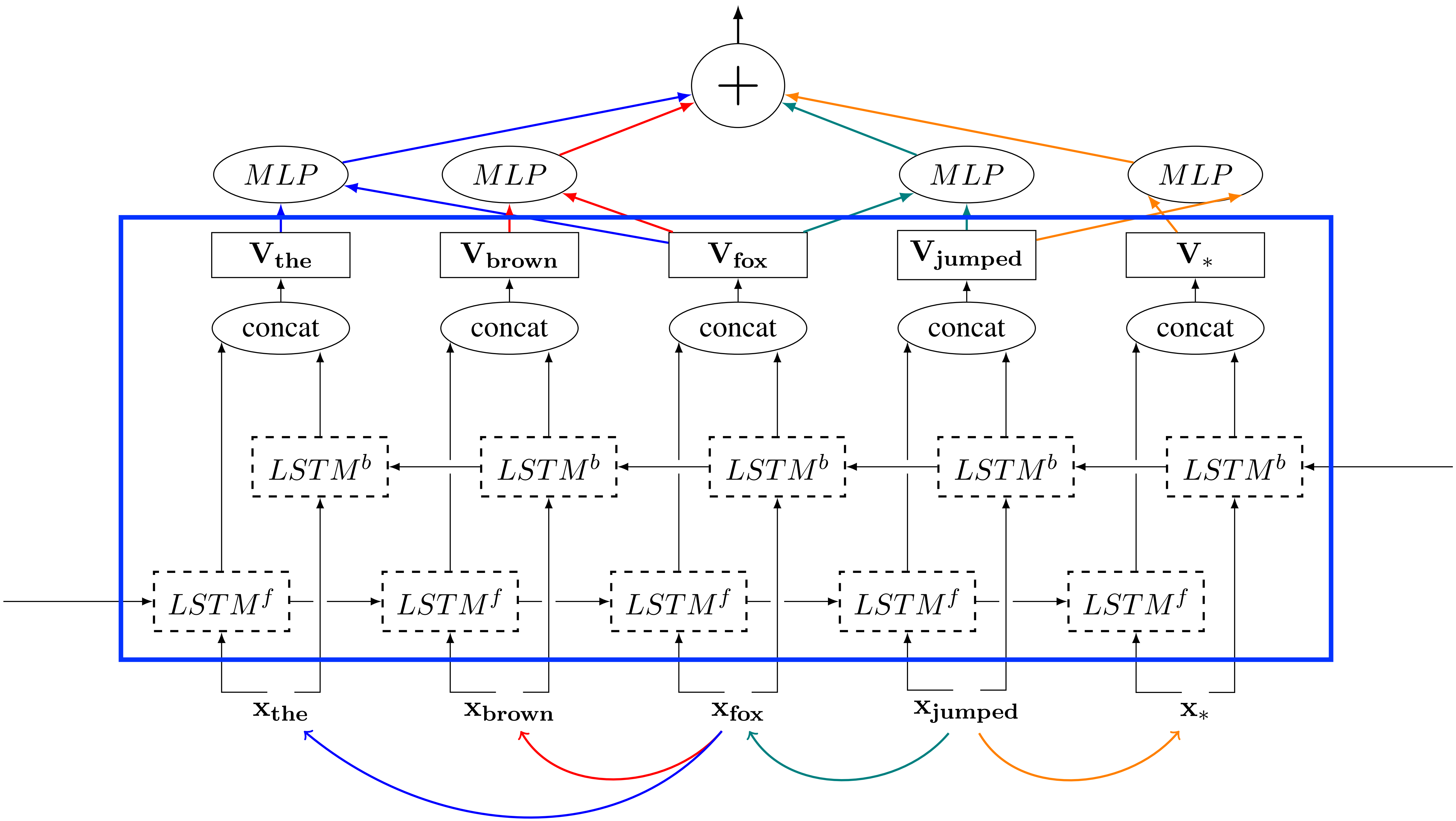

Artur Kulmizev



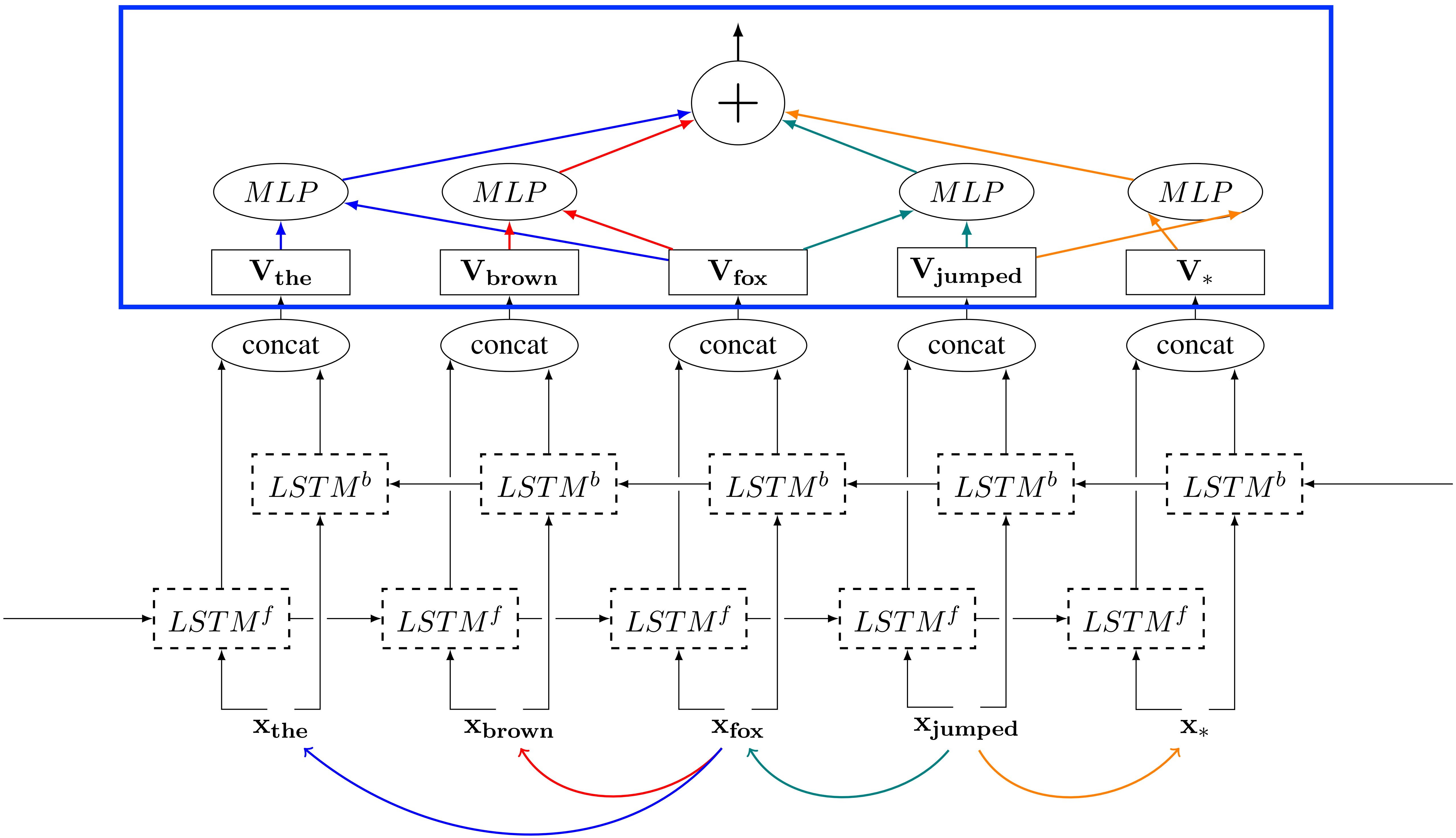
Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representation Networks. *TACL* 4: 313–327.



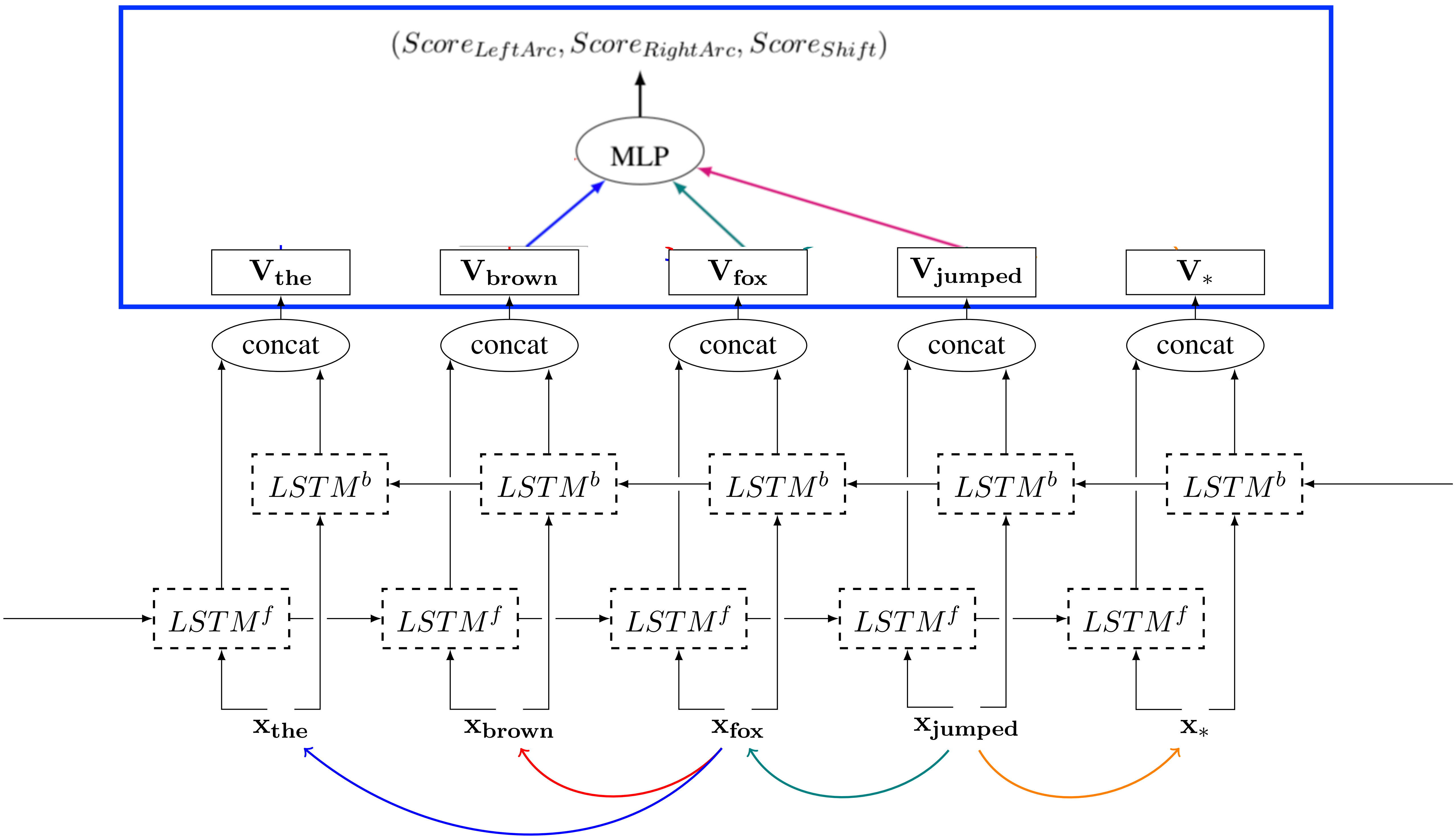
Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representation Networks. *TACL* 4: 313–327.



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



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



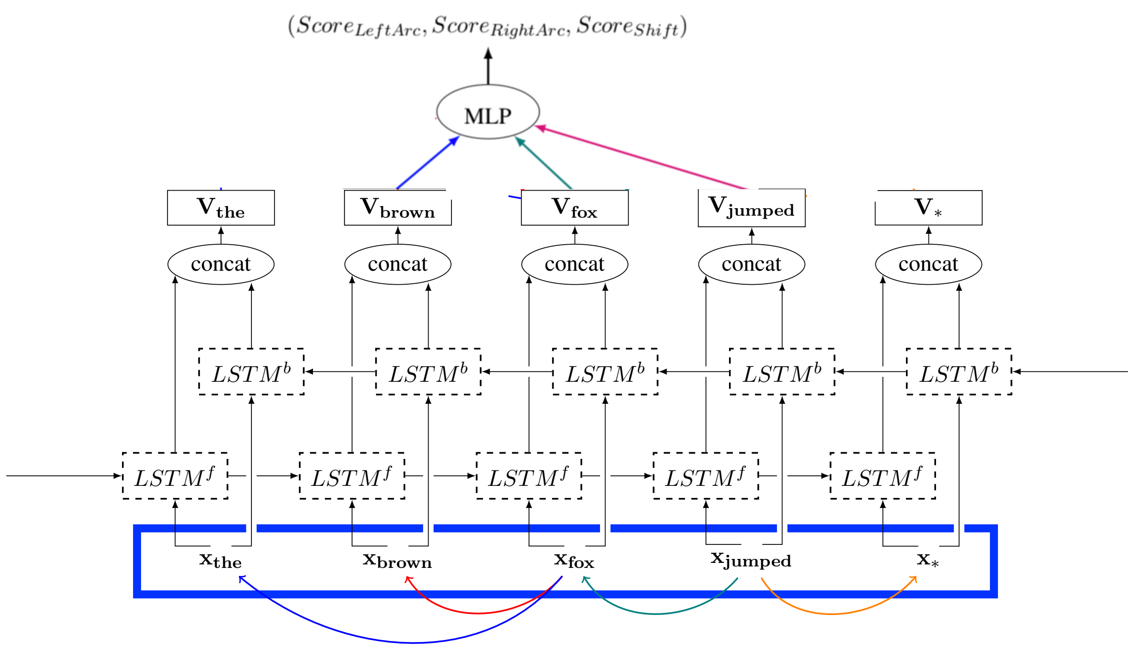
Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representation Networks. *TACL* 4: 313–327.

Representing Words

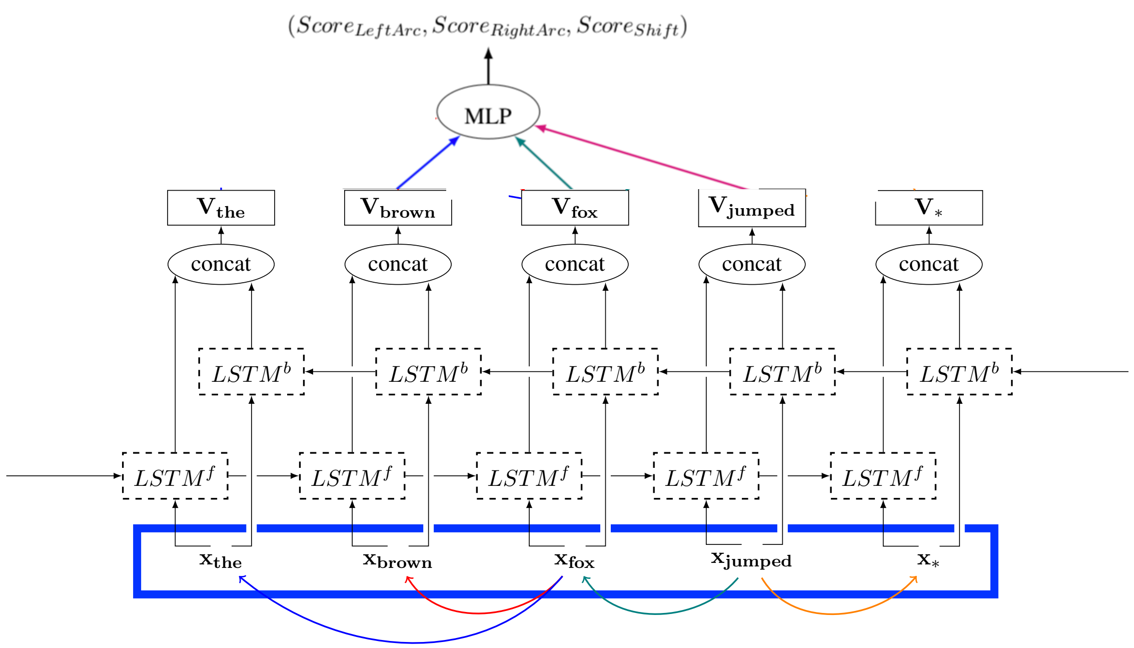
- How do parsers benefit from pre-trained word embeddings, character models and part-of-speech tags?
- Are the techniques complementary or redundant?
- How do results vary across word frequencies, word categories and languages?

Aaron Smith, Miryam de Lhoneux, Sara Stymne, and Joakim Nivre. 2018. An Investigation of the Interactions between Pre-Trained Word Embeddings, Character Models and PoS Tags in Dependency Parsing. In *Proceedings of EMNLP*.

Experimental Setup

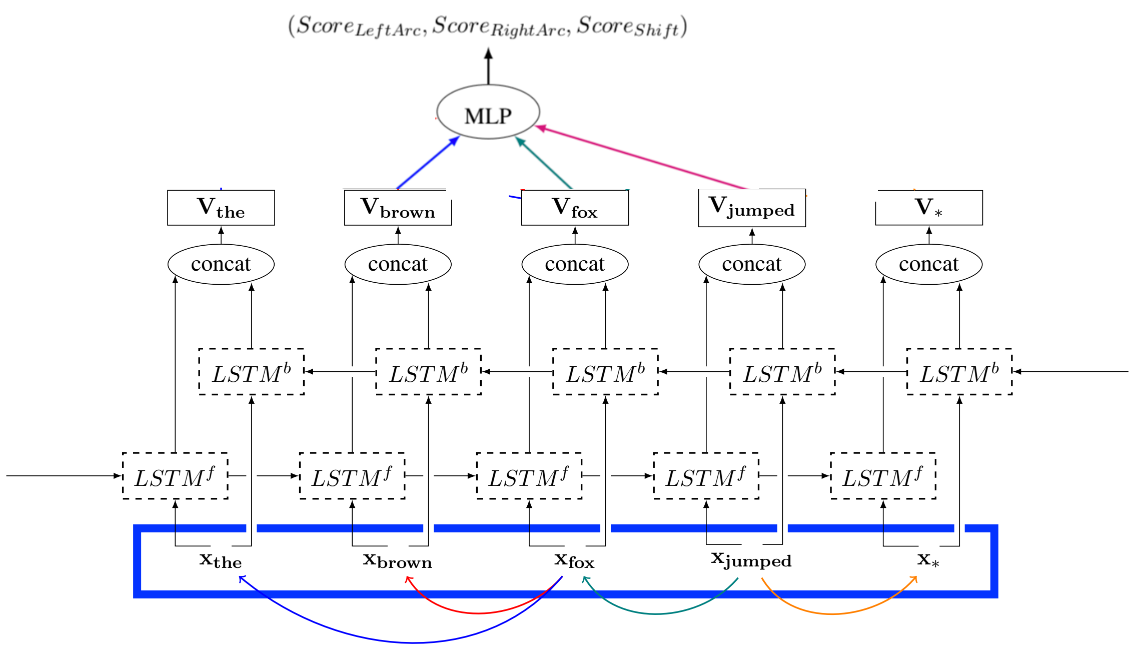


Experimental Setup



$$\boxed{x_i = e^r(w_i)}$$

Experimental Setup



$$\boxed{x_i = e^r(w_i)}$$

$$\boxed{x_i = e^t(w_i) \circ \text{BiLSTM}(ch_{1:m}) \circ e(p_i)}$$

Results

baseline	67.7	combined	81.0
+EXT	76.1	−EXT	79.9
+CHAR	78.3	−CHAR	79.2
+POS	75.9	−POS	80.3

Treebank	Sentences		TTR	Chars
Ancient Greek	14864	1019	0.15	179
Arabic	6075	909	0.10	105
Chinese	3997	500	0.16	3571
English	12534	2002	0.07	108
Finnish	12217	1364	0.26	244
Hebrew	5241	484	0.11	53
Korean	4400	950	0.46	1730
Russian	3850	579	0.30	189
Swedish	4303	504	0.16	86

Results

baseline	67.7	combined	81.0
+EXT	76.1	−EXT	79.9
+CHAR	78.3	−CHAR	79.2
+POS	75.9	−POS	80.3

Treebank	Sentences		TTR	Chars
Ancient Greek	14864	1019	0.15	179
Arabic	6075	909	0.10	105
Chinese	3997	500	0.16	3571
English	12534	2002	0.07	108
Finnish	12217	1364	0.26	244
Hebrew	5241	484	0.11	53
Korean	4400	950	0.46	1730
Russian	3850	579	0.30	189
Swedish	4303	504	0.16	86

Results

baseline	67.7	combined	81.0
+EXT	76.1	−EXT	79.9
+CHAR	78.3	−CHAR	79.2
+POS	75.9	−POS	80.3

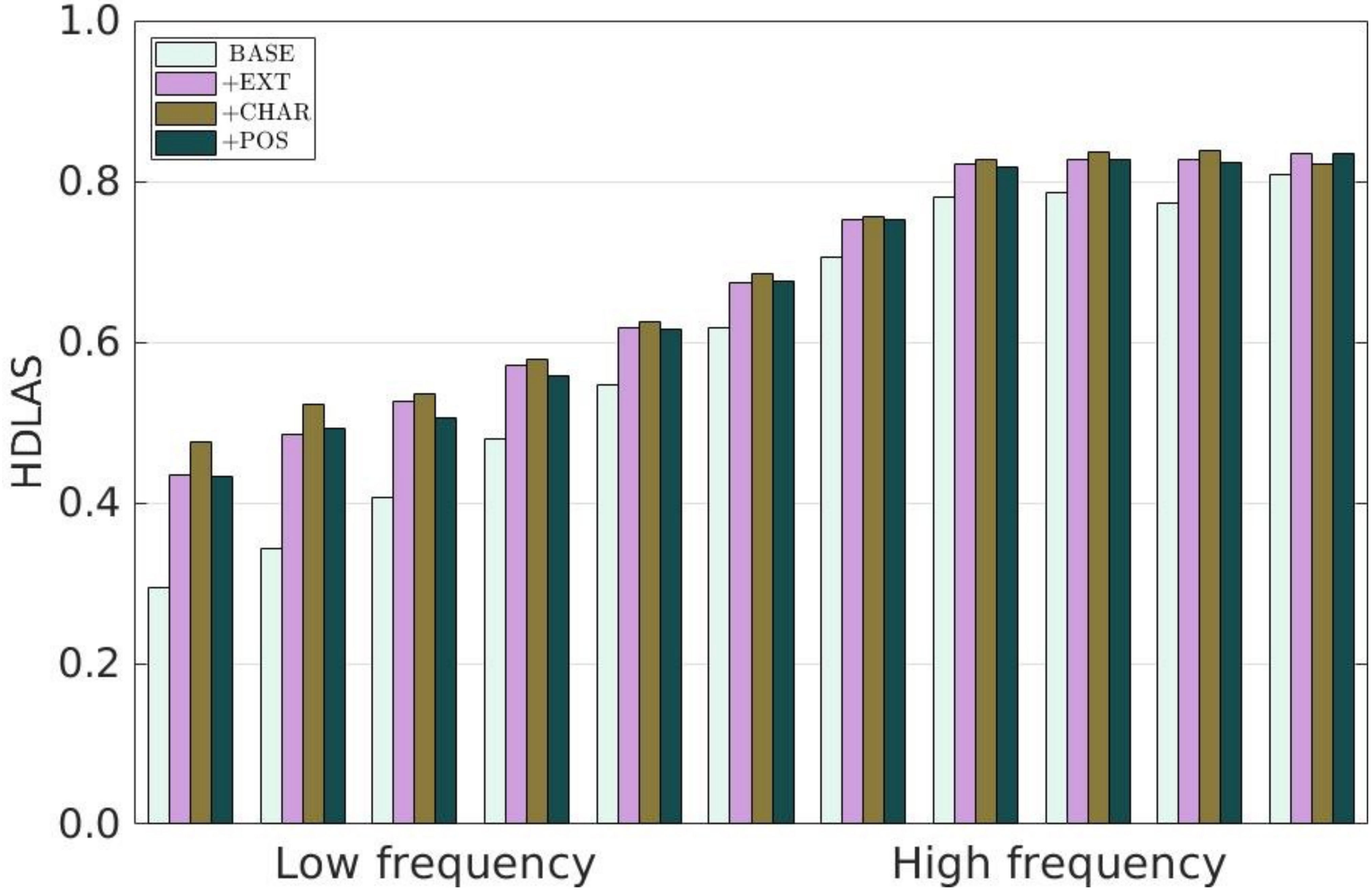
Treebank	Sentences		TTR	Chars
Ancient Greek	14864	1019	0.15	179
Arabic	6075	909	0.10	105
Chinese	3997	500	0.16	3571
English	12534	2002	0.07	108
Finnish	12217	1364	0.26	244
Hebrew	5241	484	0.11	53
Korean	4400	950	0.46	1730
Russian	3850	579	0.30	189
Swedish	4303	504	0.16	86

Results

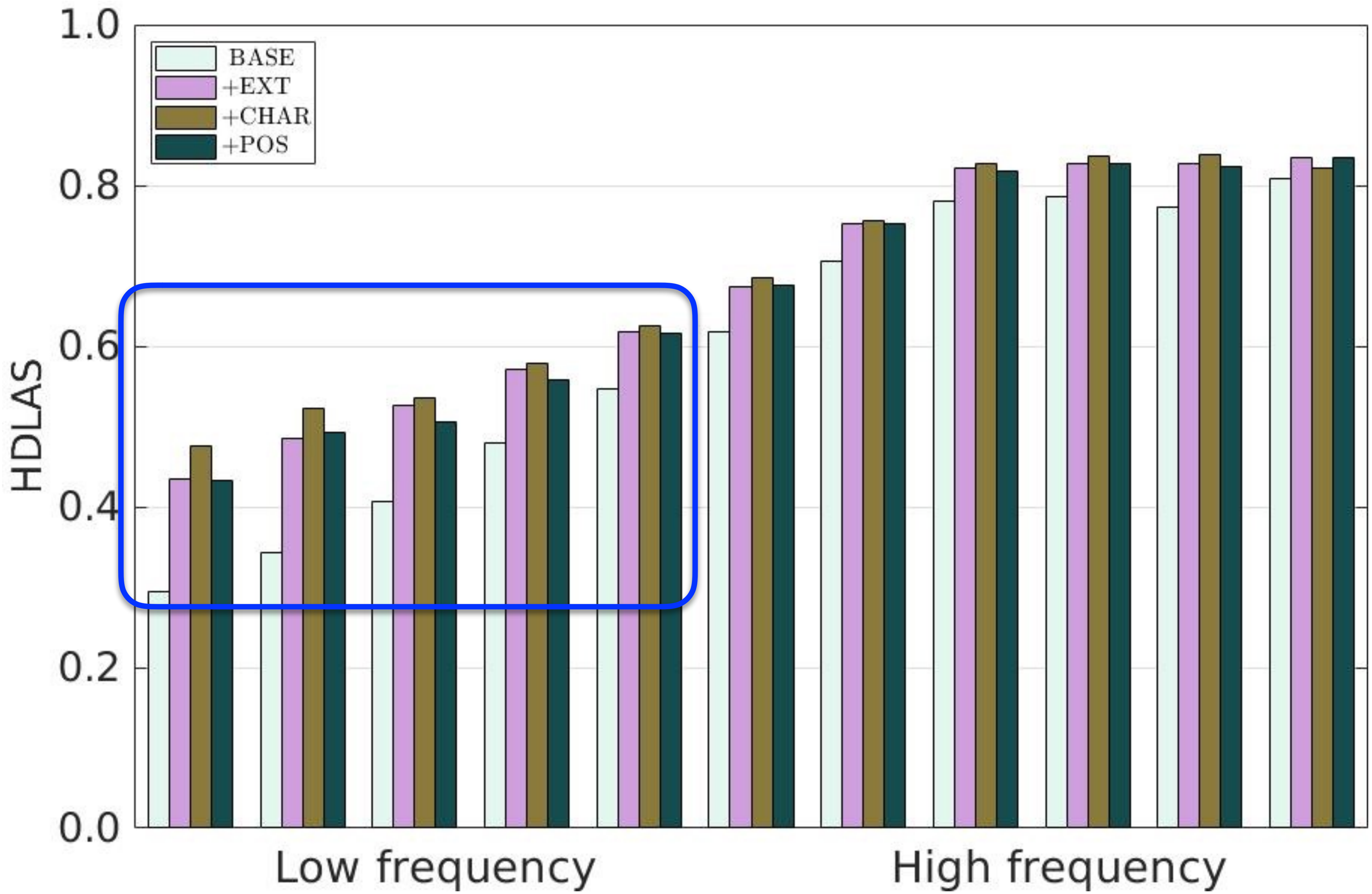
baseline	67.7	combined	81.0
+EXT	76.1	−EXT	79.9
+CHAR	78.3	−CHAR	79.2
+POS	75.9	−POS	80.3

Treebank	Sentences		TTR	Chars
Ancient Greek	14864	1019	0.15	179
Arabic	6075	909	0.10	105
Chinese	3997	500	0.16	3571
English	12534	2002	0.07	108
Finnish	12217	1364	0.26	244
Hebrew	5241	484	0.11	53
Korean	4400	950	0.46	1730
Russian	3850	579	0.30	189
Swedish	4303	504	0.16	86

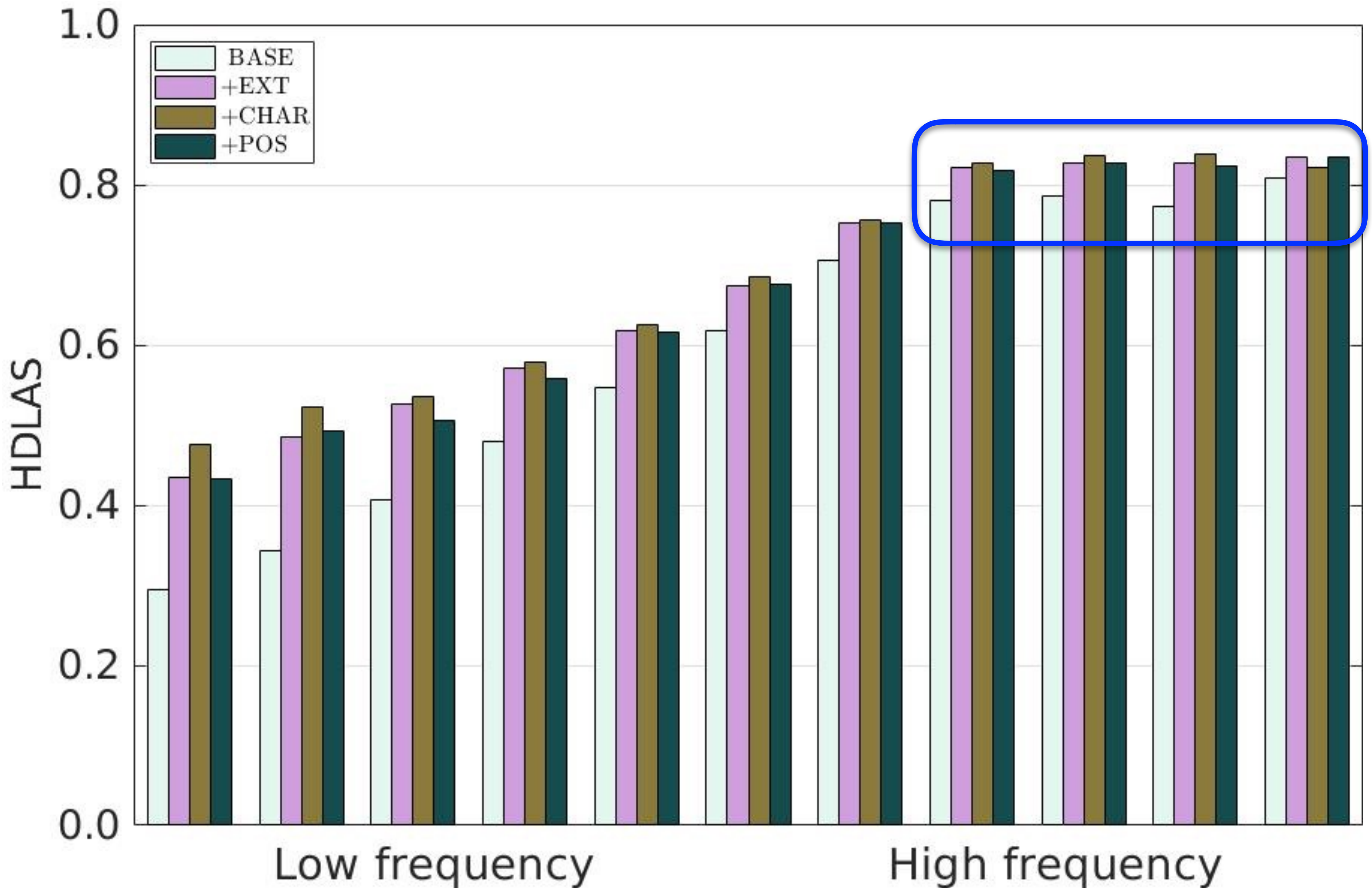
Results by Frequency



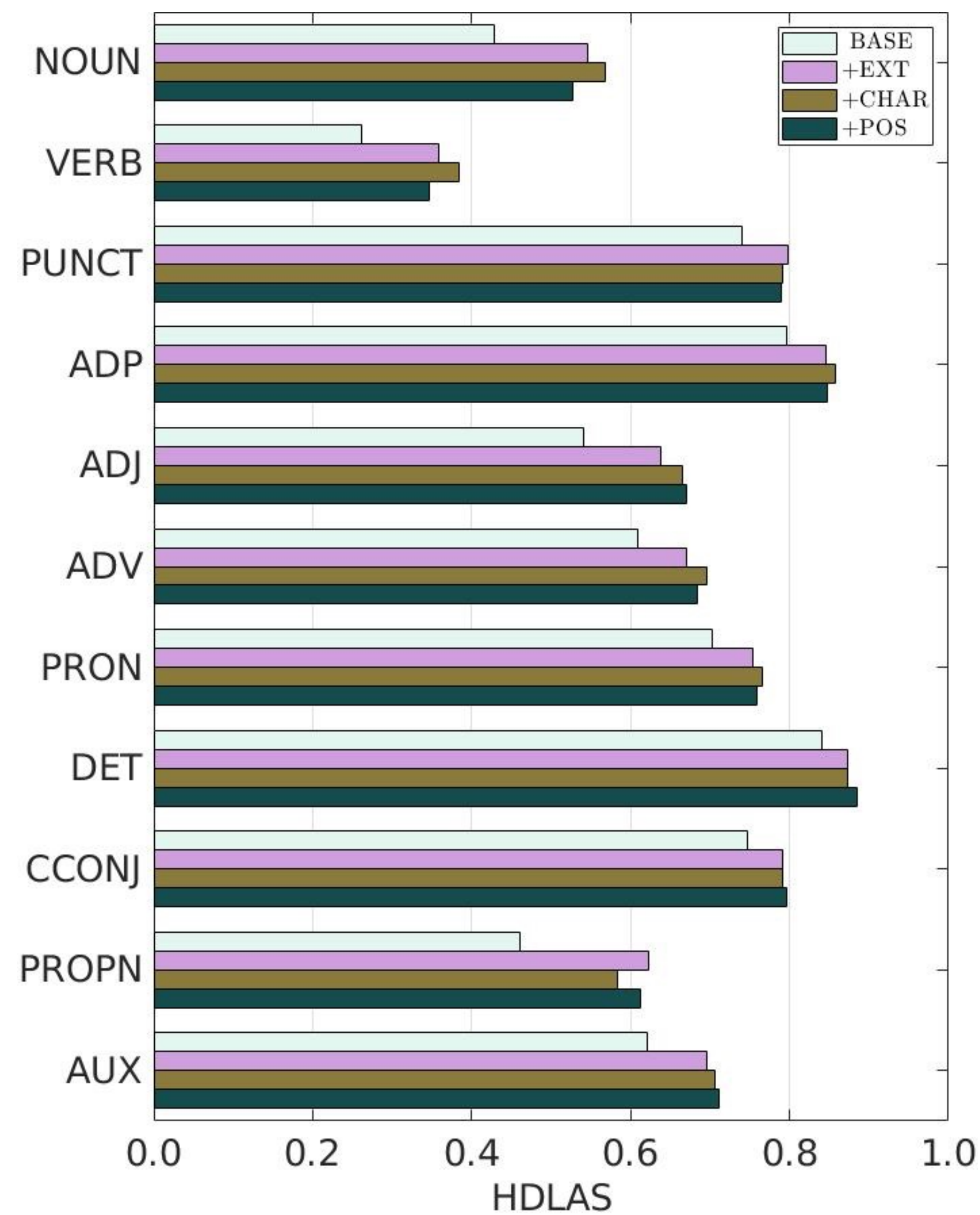
Results by Frequency



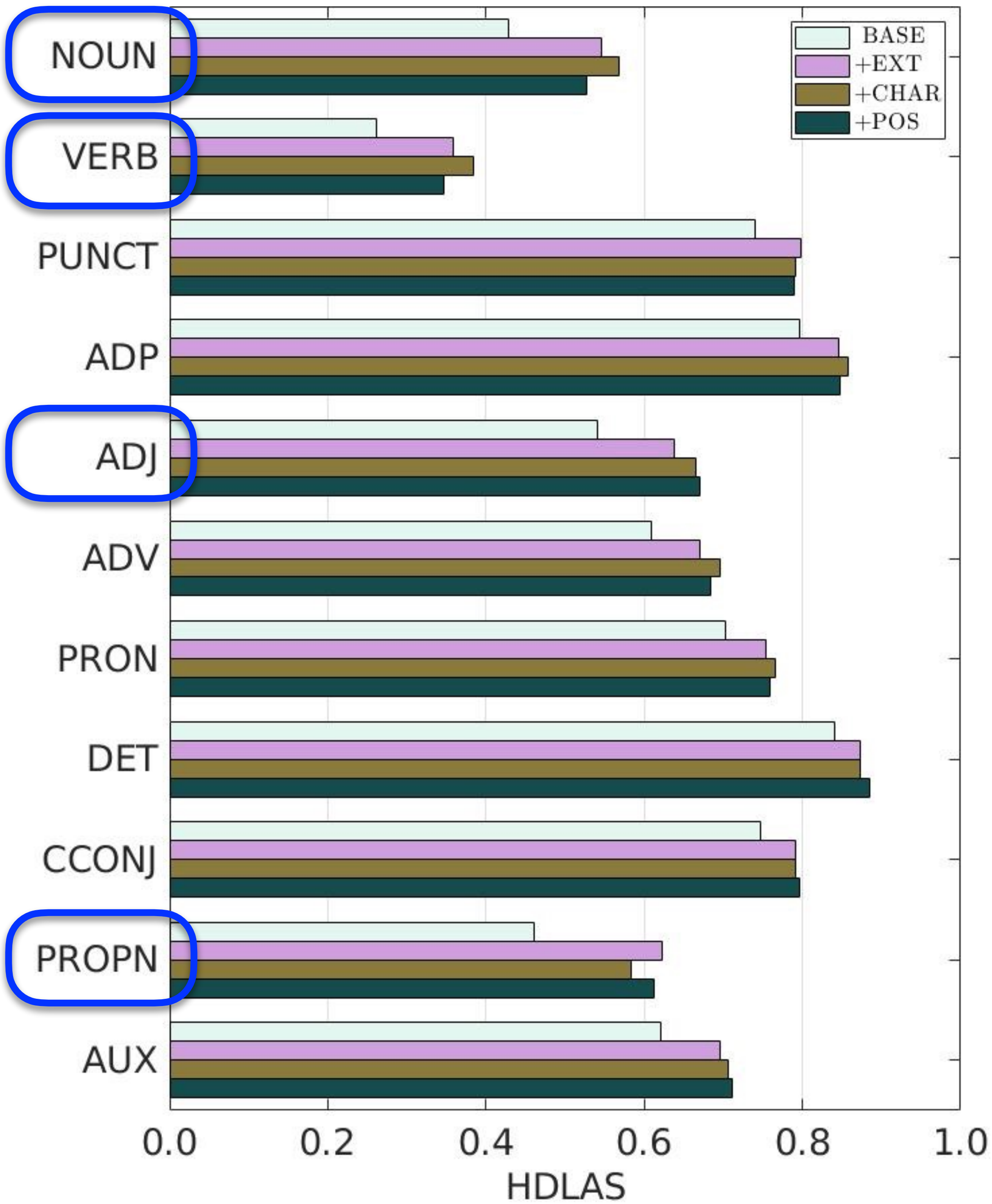
Results by Frequency



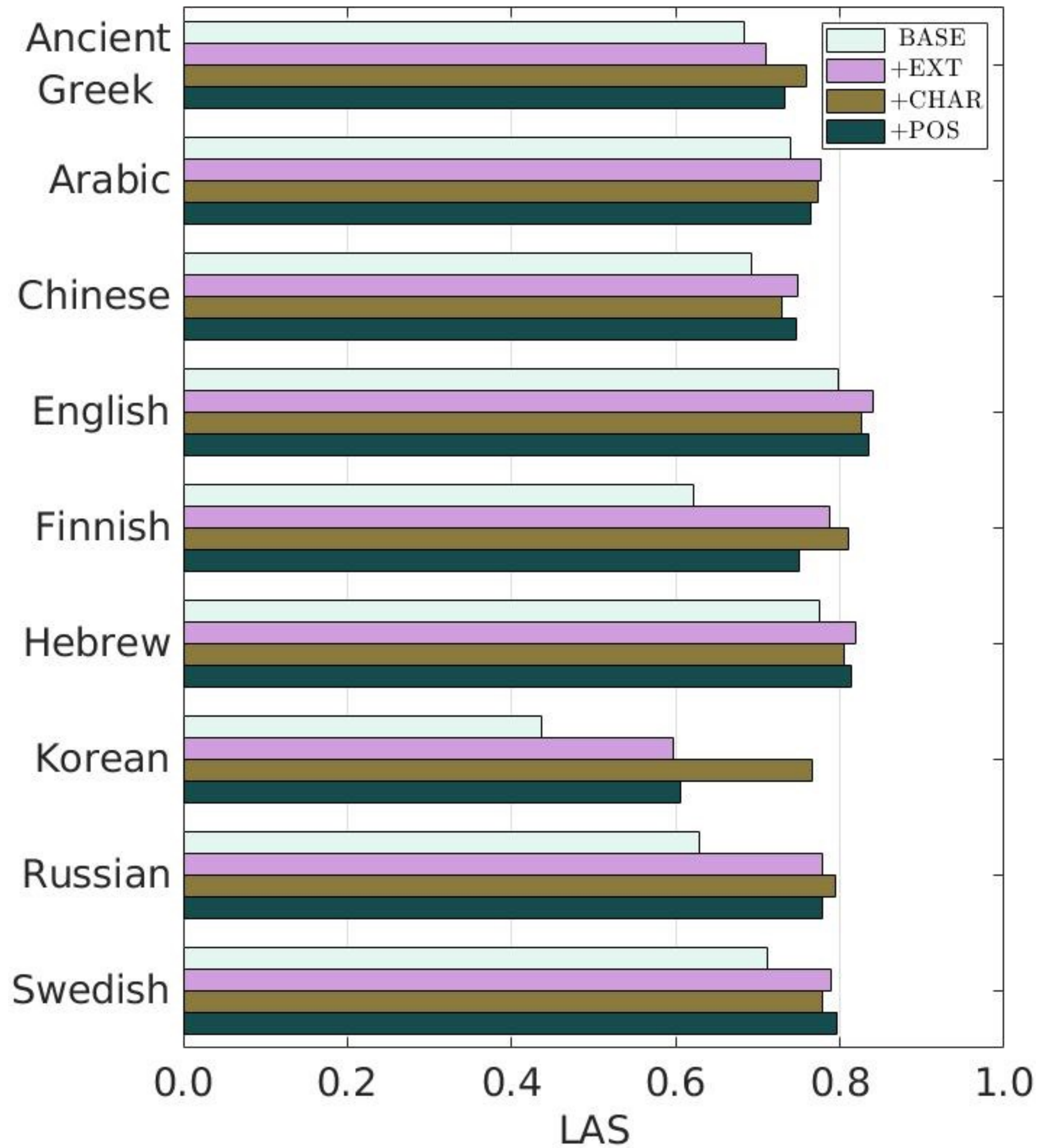
Results by PoS Tags



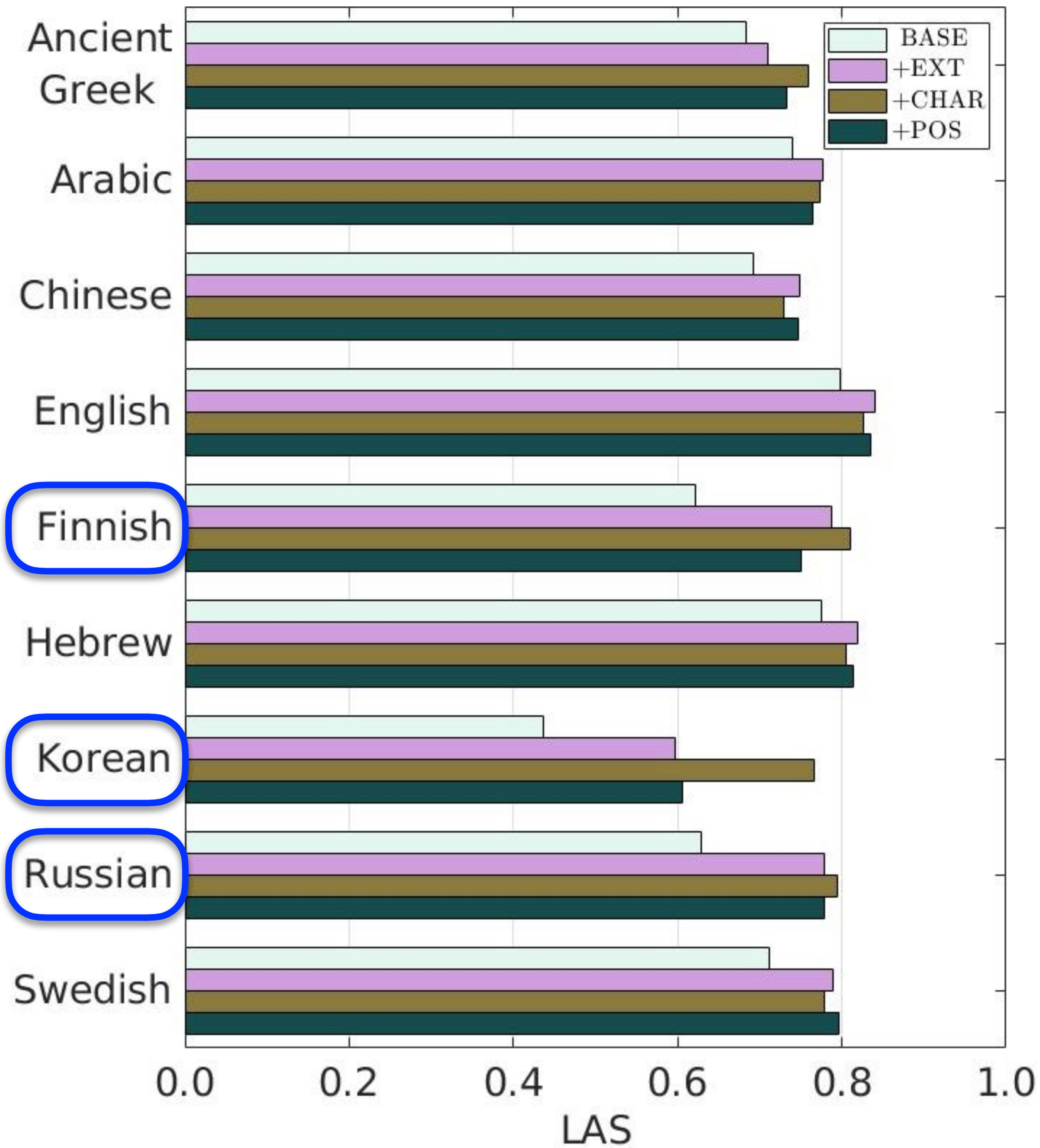
Results by PoS Tags



Results by Language



Results by Language



Main Findings

- We see the largest improvements for low-frequency and open-class words and for morphologically rich languages
- Techniques are mutually redundant, but character models are the most effective for low-frequency words
- Part-of-speech tags are potentially effective for high-frequency function words, but current taggers are not accurate enough to realize this potential

A Tale of Two Parsers

- Transition-based and graph-based dependency parsers are known to have distinctive error profiles
- Do these patterns persist in the presence of neural network techniques?
- Do deep contextualized word representations benefit transition-based parsers more than graph-based parsers?

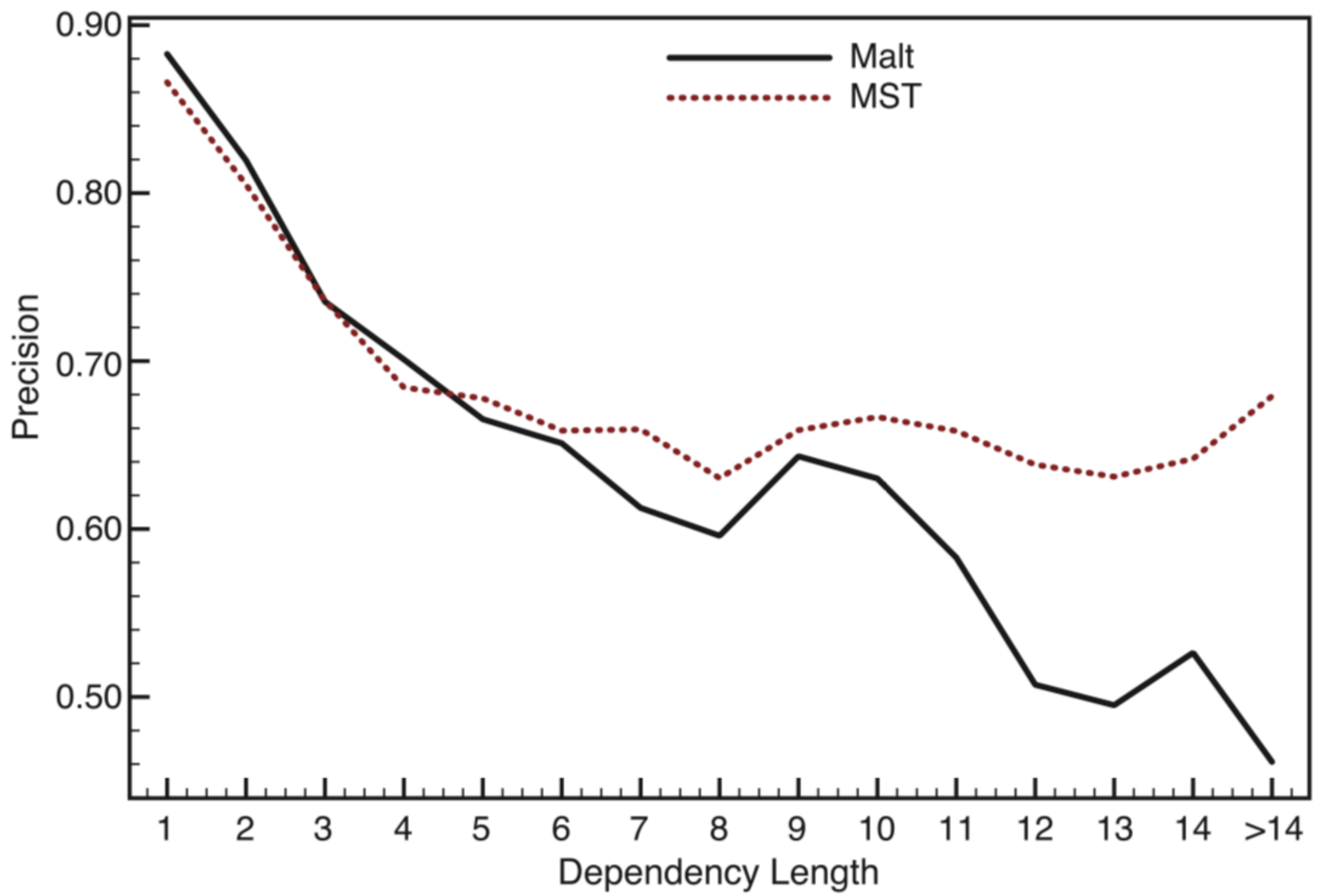
Artur Kulmizev, Miryam de Lhoneux, Johannes Gontrum, Elena Fano and Joakim Nivre 2019. Deep Contextualized Word Embeddings in Transition-Based and Graph-Based Dependency Parsing – A Tale of Two Parsers Revisited. In *Proceedings of EMNLP*.

Historical Background

Transition-Based		Graph-Based
short sentences short dependencies nouns core arguments	**better on**	long sentences long dependencies verbs main predicates
rich features greedy decoding	**due to**	limited features exact decoding

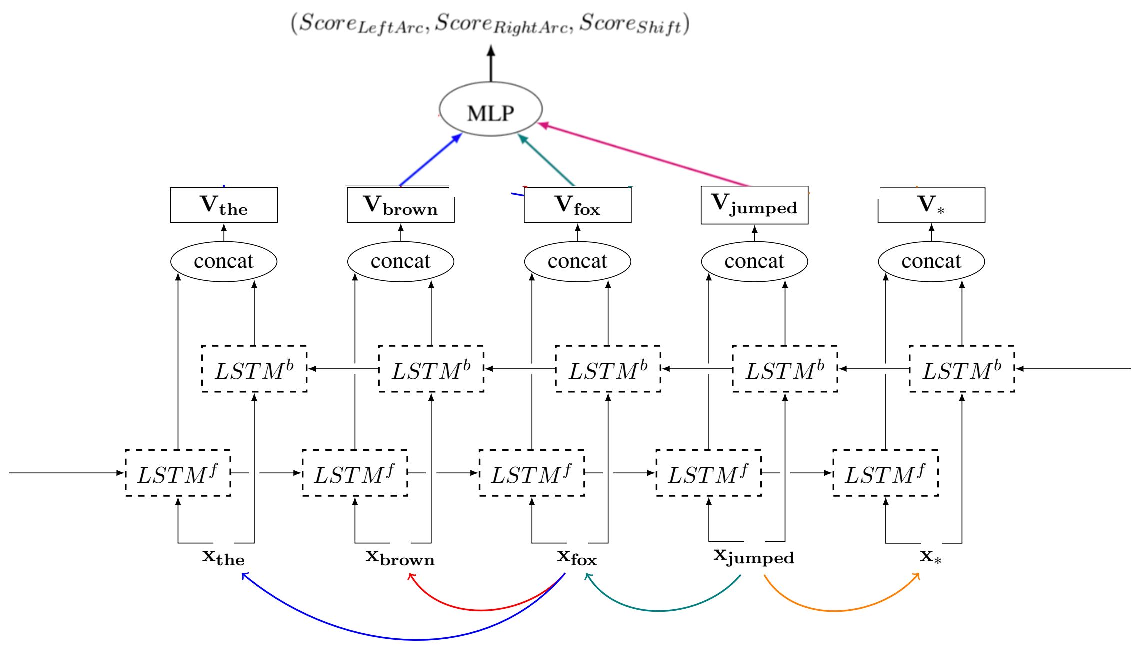
Ryan McDonald and Joakim Nivre. 2007. Characterizing the Errors of Data-Driven Dependency Parsing Models. In Proceedings of EMNLP, pages 122–131.

Historical Background

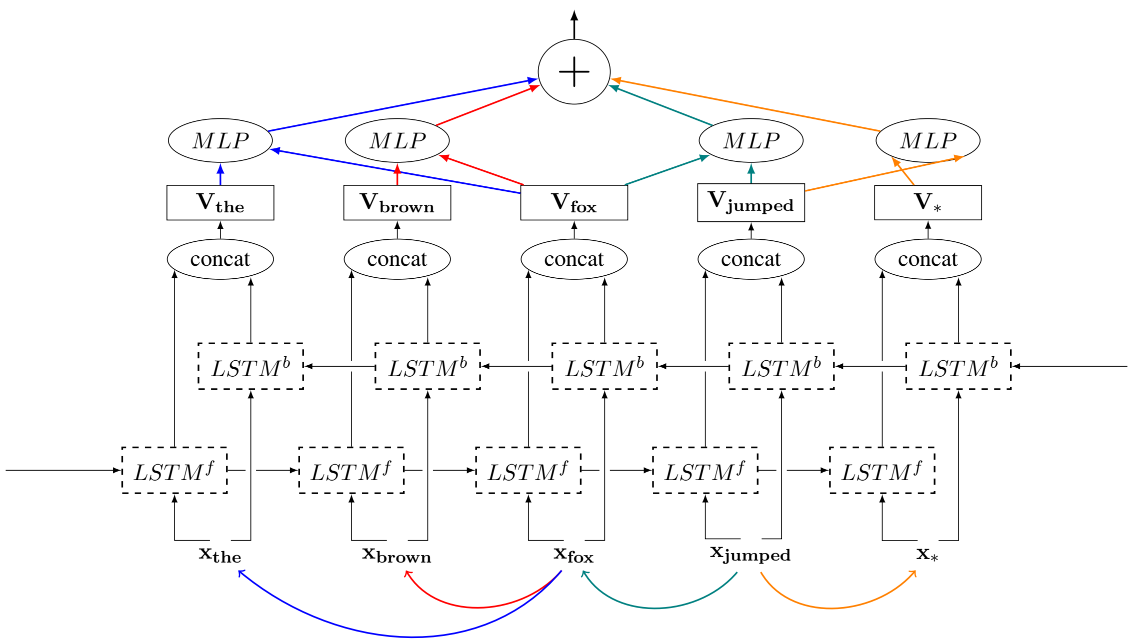


Ryan McDonald and Joakim Nivre. 2007. Characterizing the Errors of Data-Driven Dependency Parsing Models. In Proceedings of EMNLP, pages 122–131.

Experimental Setup

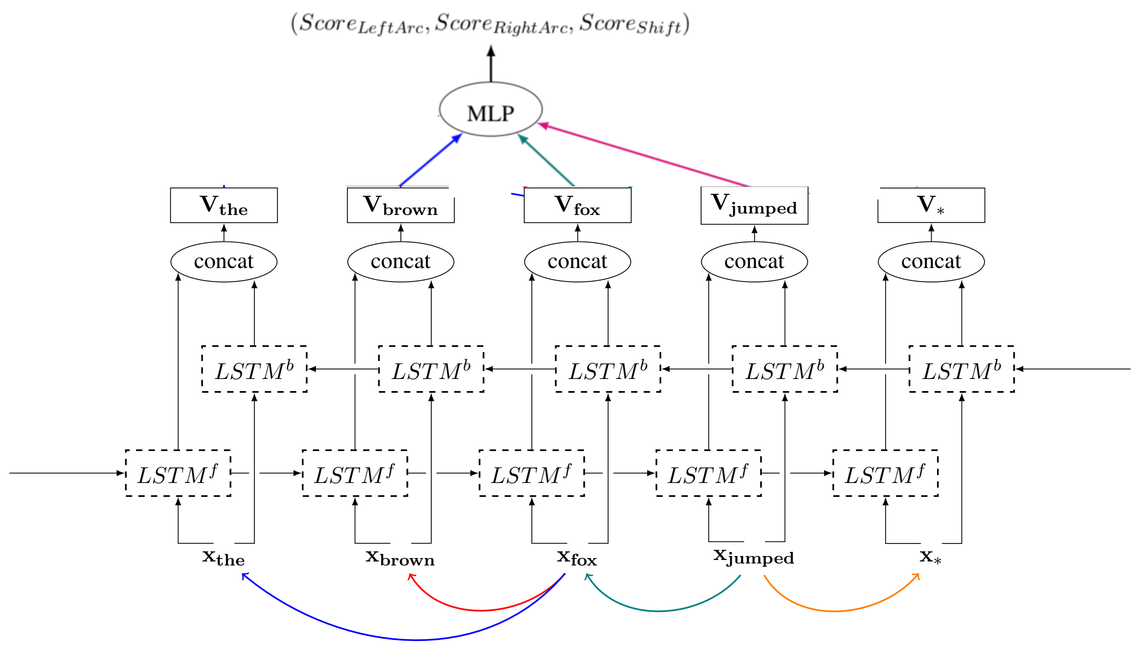


TR

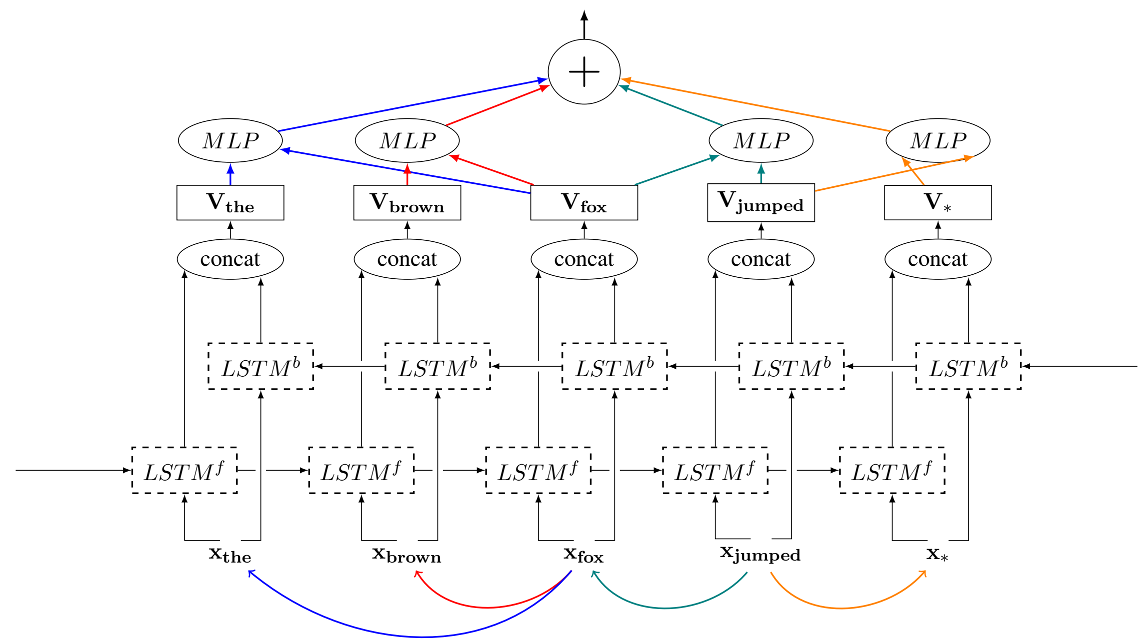


GR

Experimental Setup



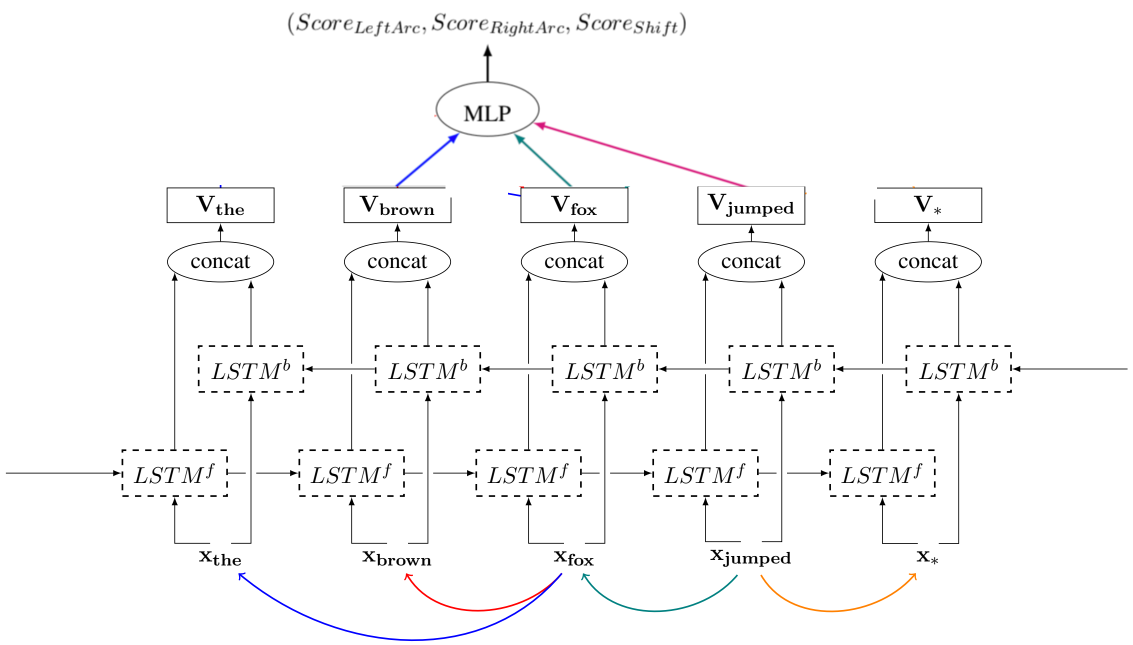
TR



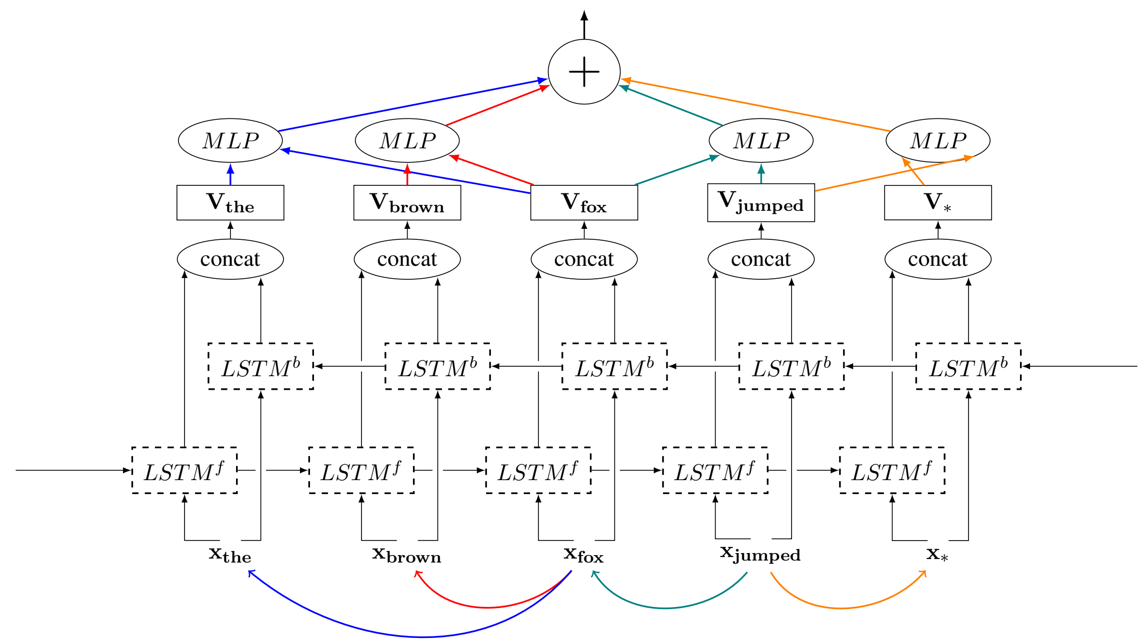
GR

$$x_i = e^t(w_i) \circ \text{BiLSTM}(ch_{1:m})$$

Experimental Setup



TR



GR

$$x_i = e^t(w_i) \circ \mathrm{BiLSTM}(ch_{1:m})$$

+E(LMo)

+B(ERT)

Results

Language	TR	GR	TR+E	GR+E	TR+B	GR+B
Arabic	79.1	79.9	82.0	81.7	81.9	81.8
Basque	73.6	77.6	80.1	81.4	77.9	79.8
Chinese	75.3	76.7	79.8	80.4	83.7	83.4
English	82.7	83.3	87.0	86.5	87.8	87.6
Finnish	80.0	81.4	87.0	86.6	85.1	83.9
Hebrew	81.1	82.4	85.2	85.9	85.5	85.9
Hindi	88.4	89.6	91.0	91.2	89.5	90.8
Italian	88.0	88.2	90.9	90.6	92.0	91.7
Japanese	92.1	92.2	93.1	93.0	92.9	92.1
Korean	79.6	81.2	82.3	82.3	83.7	84.2
Russian	88.3	88.0	90.7	90.6	91.5	91.0
Swedish	80.5	81.6	86.9	86.2	87.6	86.9
Turkish	57.8	61.2	62.6	63.8	64.2	64.9
Average	**80.5**	**81.8**	**84.5**	**84.6**	**84.9**	**84.9**

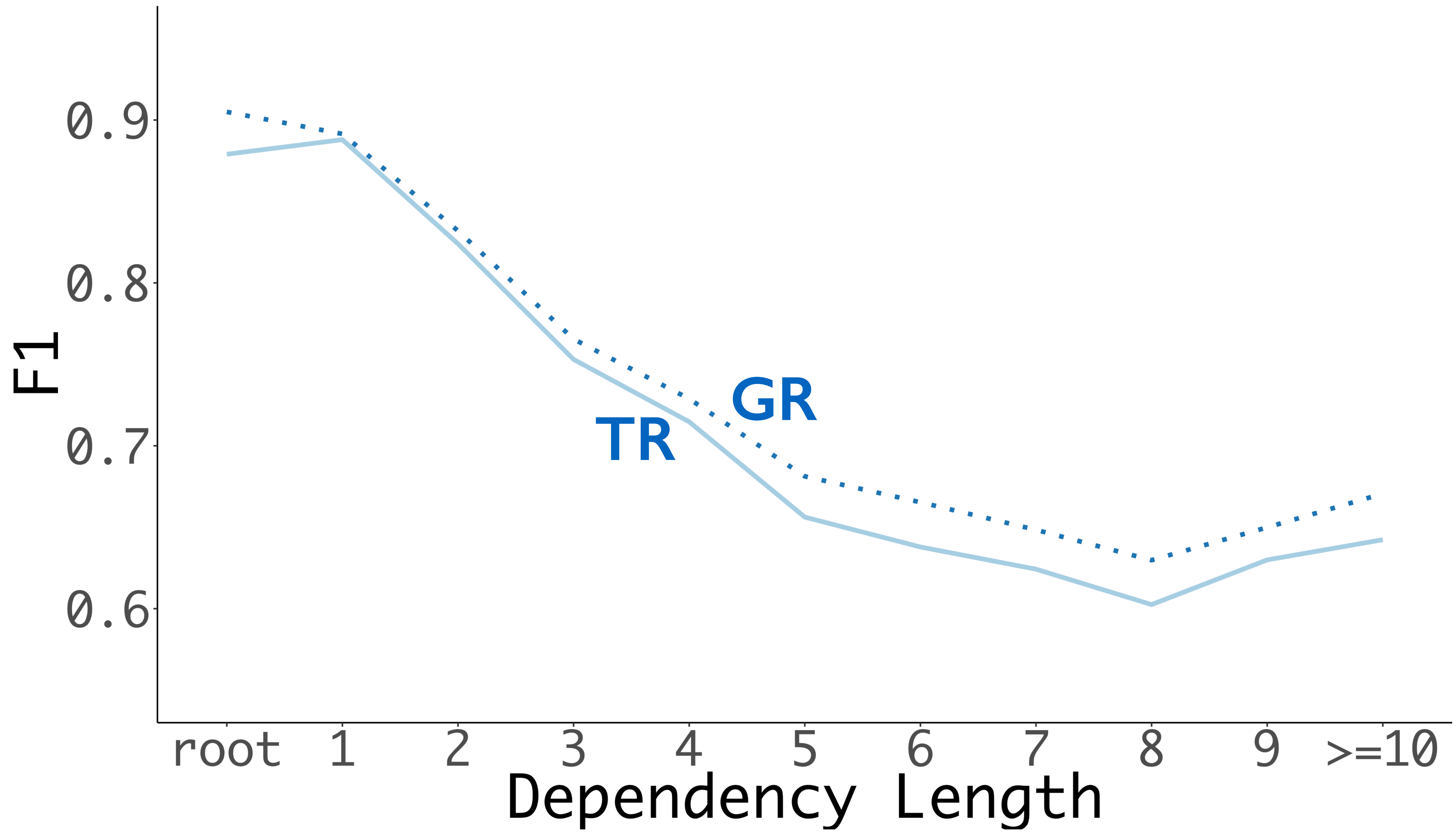
Results

Language	TR	GR	TR+E	GR+E	TR+B	GR+B
Arabic	79.1	79.9	82.0	81.7	81.9	81.8
Basque	73.6	77.6	80.1	81.4	77.9	79.8
Chinese	75.3	76.7	79.8	80.4	83.7	83.4
English	82.7	83.3	87.0	86.5	87.8	87.6
Finnish	80.0	81.4	87.0	86.6	85.1	83.9
Hebrew	81.1	82.4	85.2	85.9	85.5	85.9
Hindi	88.4	89.6	91.0	91.2	89.5	90.8
Italian	88.0	88.2	90.9	90.6	92.0	91.7
Japanese	92.1	92.2	93.1	93.0	92.9	92.1
Korean	79.6	81.2	82.3	82.3	83.7	84.2
Russian	88.3	88.0	90.7	90.6	91.5	91.0
Swedish	80.5	81.6	86.9	86.2	87.6	86.9
Turkish	57.8	61.2	62.6	63.8	64.2	64.9
Average	**80.5**	**81.8**	**84.5**	**84.6**	**84.9**	**84.9**
			+3.99	**+2.85**		

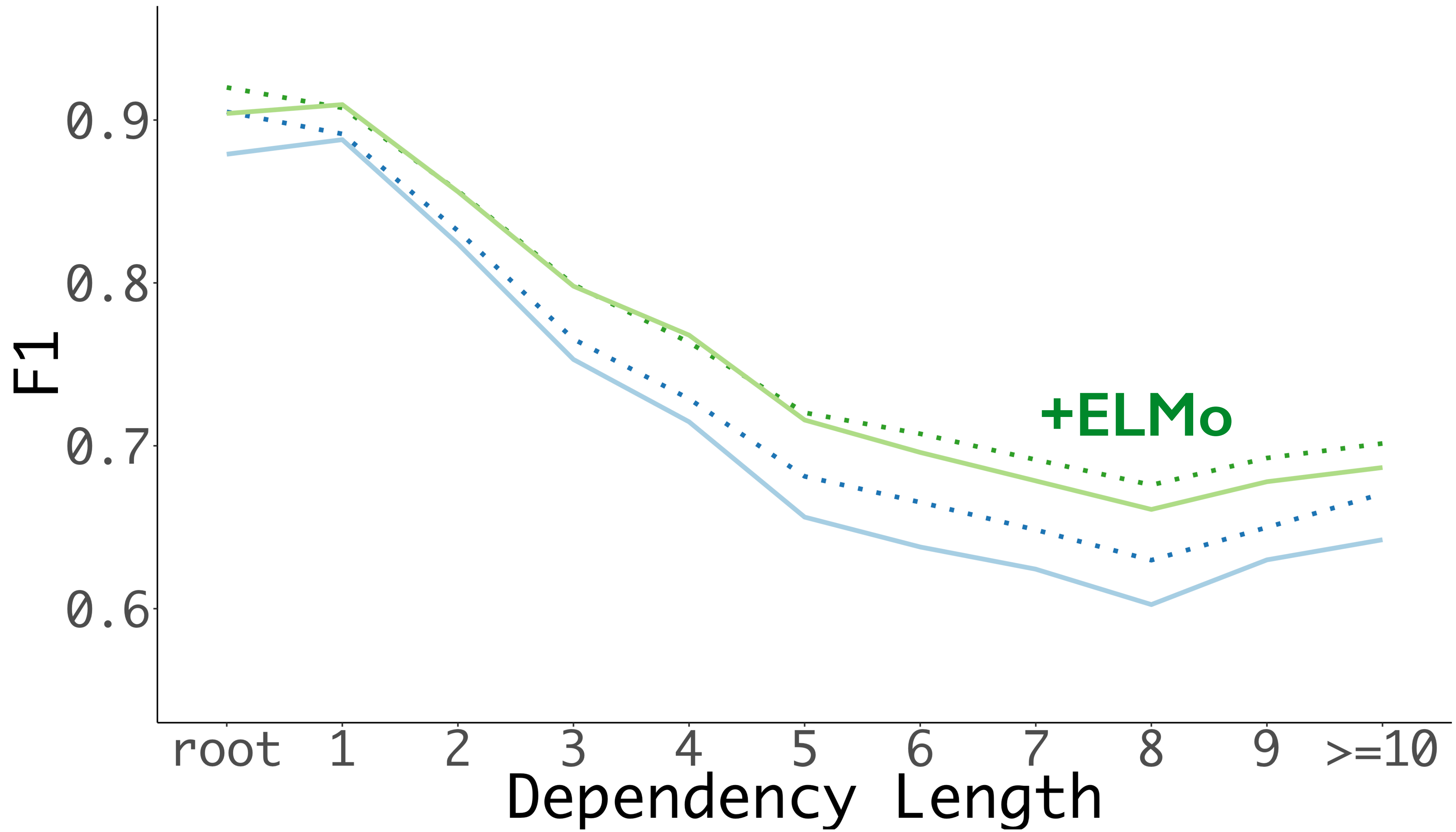
Results

Language	TR	GR	TR+E	GR+E	TR+B	GR+B
Arabic	79.1	79.9	82.0	81.7	81.9	81.8
Basque	73.6	77.6	80.1	81.4	77.9	79.8
Chinese	75.3	76.7	79.8	80.4	83.7	83.4
English	82.7	83.3	87.0	86.5	87.8	87.6
Finnish	80.0	81.4	87.0	86.6	85.1	83.9
Hebrew	81.1	82.4	85.2	85.9	85.5	85.9
Hindi	88.4	89.6	91.0	91.2	89.5	90.8
Italian	88.0	88.2	90.9	90.6	92.0	91.7
Japanese	92.1	92.2	93.1	93.0	92.9	92.1
Korean	79.6	81.2	82.3	82.3	83.7	84.2
Russian	88.3	88.0	90.7	90.6	91.5	91.0
Swedish	80.5	81.6	86.9	86.2	87.6	86.9
Turkish	57.8	61.2	62.6	63.8	64.2	64.9
Average	**80.5**	**81.8**	**84.5**	**84.6**	**84.9**	**84.9**
					+4.47	**+3.13**

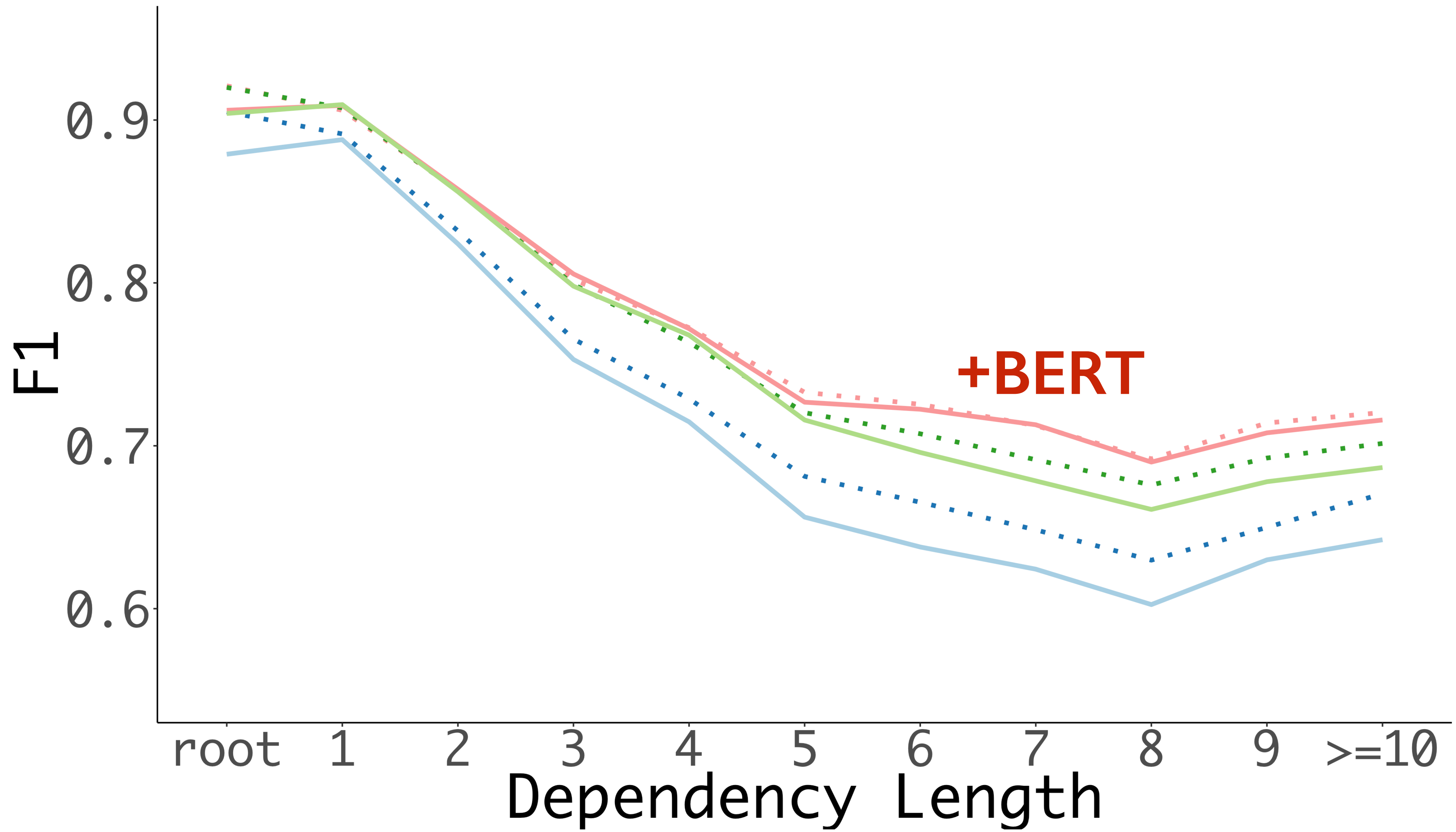
Error Analysis



Error Analysis



Error Analysis

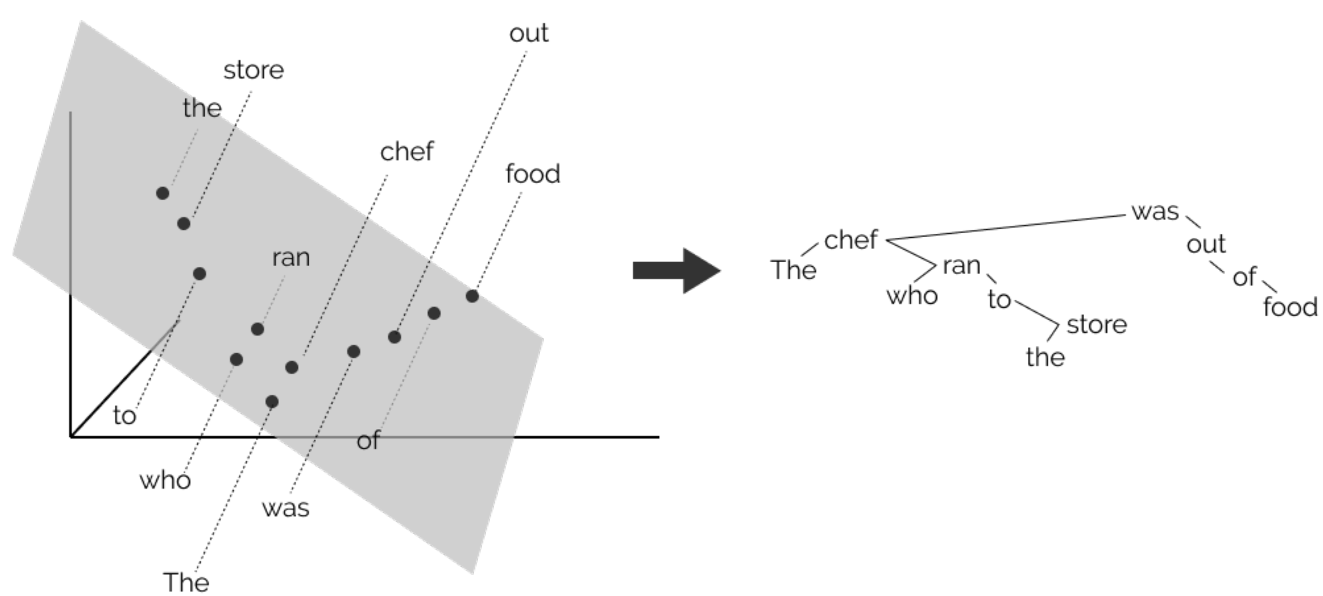


Main Findings

- The distinctive error profiles of transition-based and graph-based parsers are still visible but less pronounced
- Deep contextualized word representations improve transition-based parsers more than graph-based parsers and eliminate most of the differences
- These patterns are remarkably consistent across languages in a typologically diverse sample

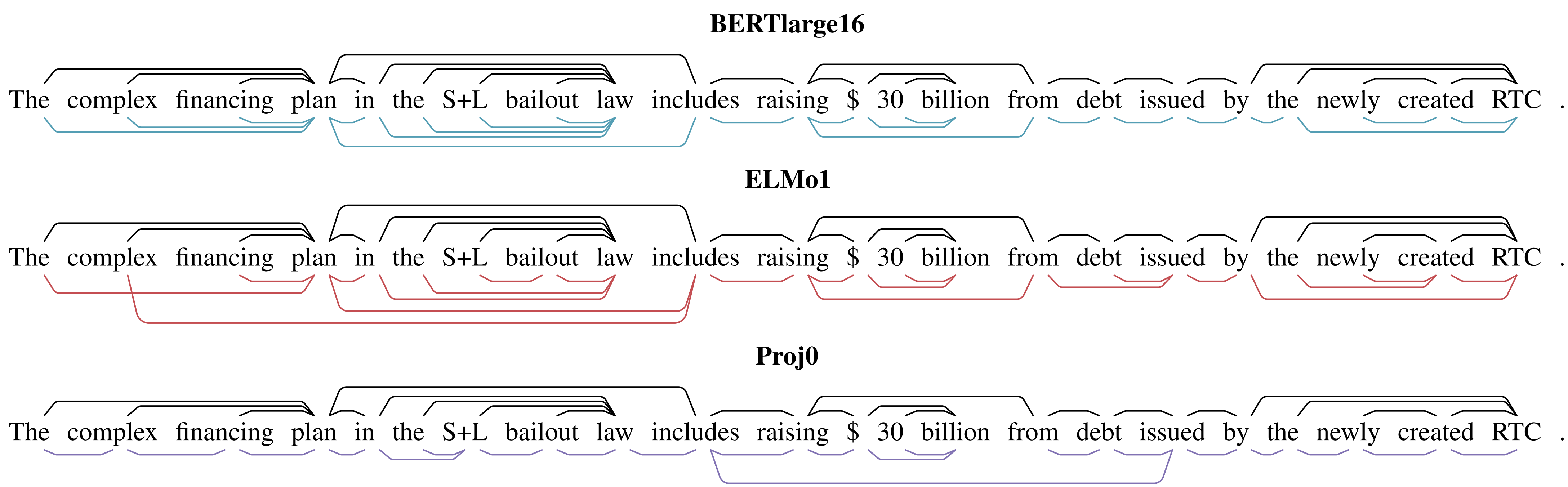
Do we need parsers at all?

- Do the vector spaces of deep contextualized word representations encode parse trees implicitly?
- Learn linear transform such that distance encodes tree distance and norm encodes tree depth

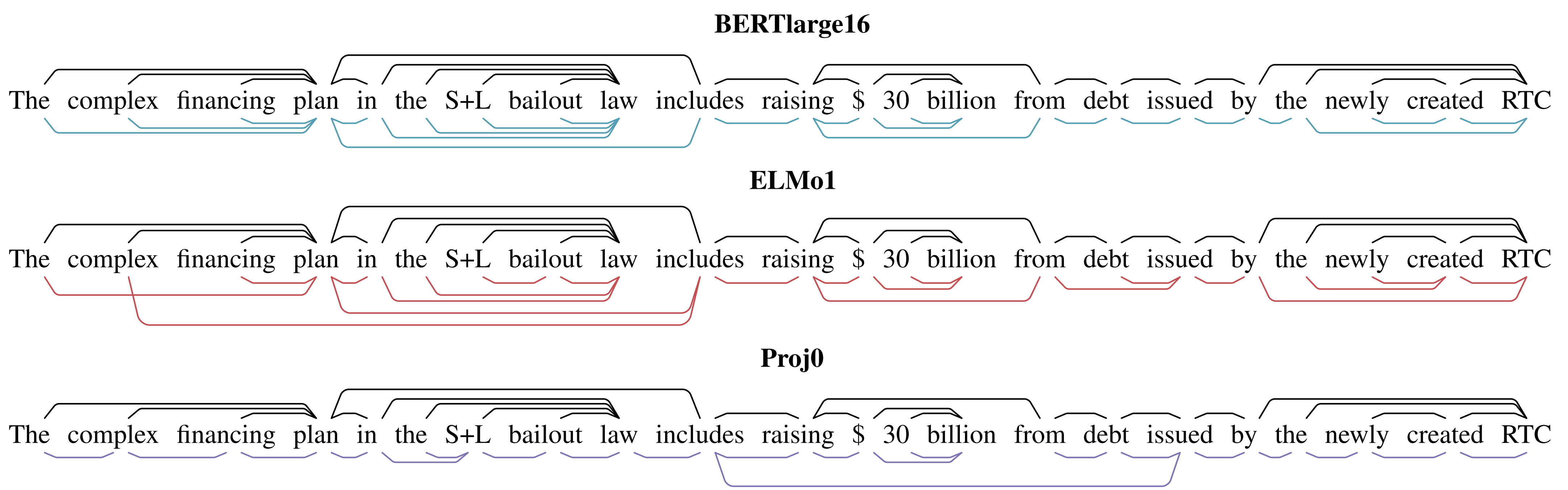


John Hewitt and Christopher D. Manning 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of NAACL*, pages 4129–4138.

Almost Dependency Parsing



Almost Dependency Parsing



- How can we extract rooted directed dependency trees?
- How do results vary across different languages?

Directed Dependency Trees

- Derive (directed) arc scores from distances and depths
- Extract maximum spanning tree using the CLE algorithm

Directed Dependency Trees

- Derive (directed) arc scores from distances and depths
- Extract maximum spanning tree using the CLE algorithm

$$\boxed{score(w_i, w_j) = \begin{cases} -dist(w_i, w_j) & \text{if } depth(w_i) < depth(w_j) \\ -\infty & \text{otherwise} \end{cases}}$$

Directed Dependency Trees

- Derive (directed) arc scores from distances and depths
- Extract maximum spanning tree using the CLE algorithm

$$\boxed{score(w_i, w_j) = \begin{cases} -dist(w_i, w_j) & \text{if } depth(w_i) < depth(w_j) \\ -\infty & \text{otherwise} \end{cases}}$$

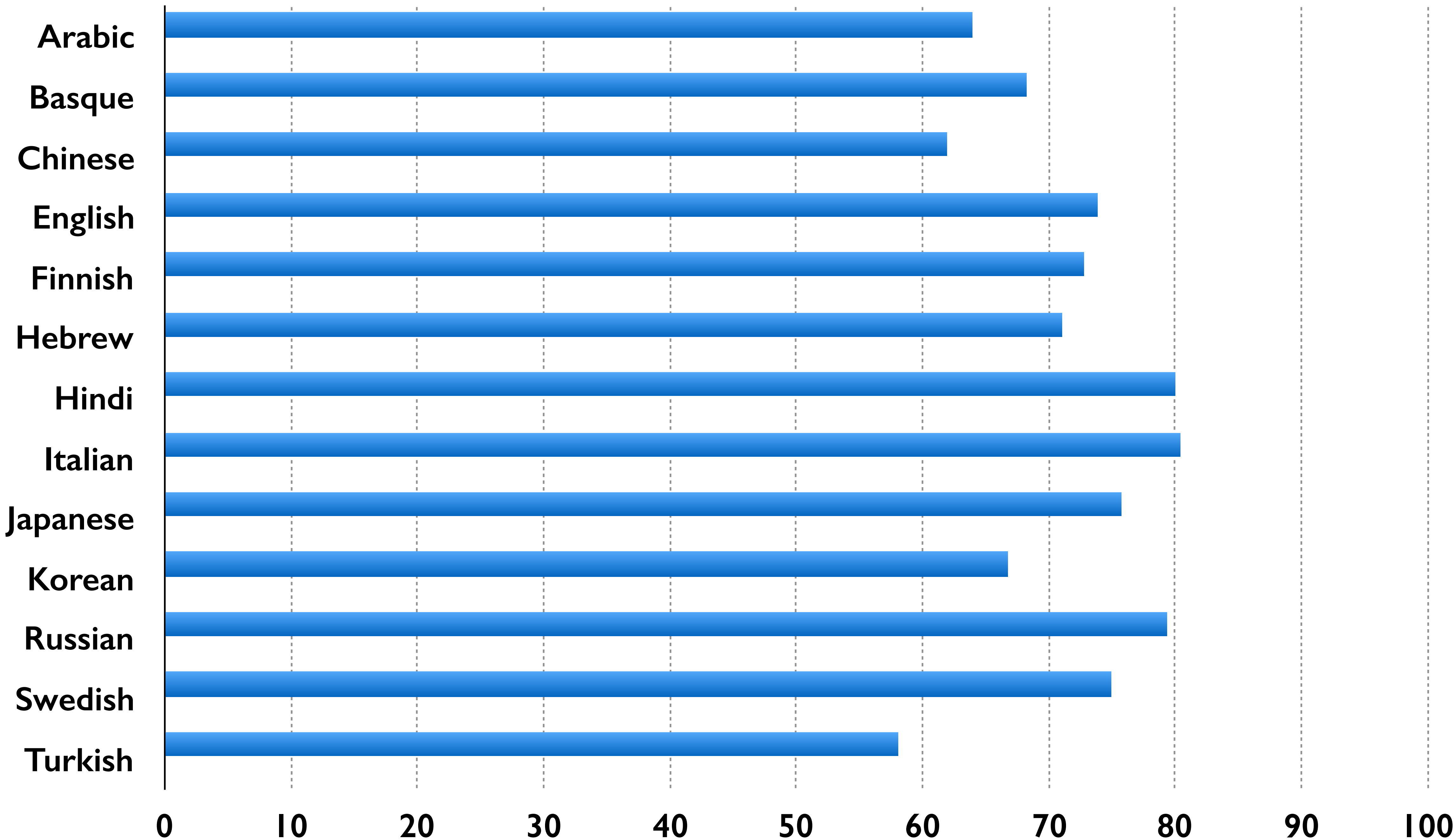
- Shorter distances correspond to higher arc scores
- Arcs from lower to higher nodes are excluded

Experimental Setup

- Multilingual BERT
- Fit probe on each of BERT’s twelve layers
- Learn weighted average across all layers
- Evaluate on same 13 UD languages as in previous studies

Mean UAS = 71.3

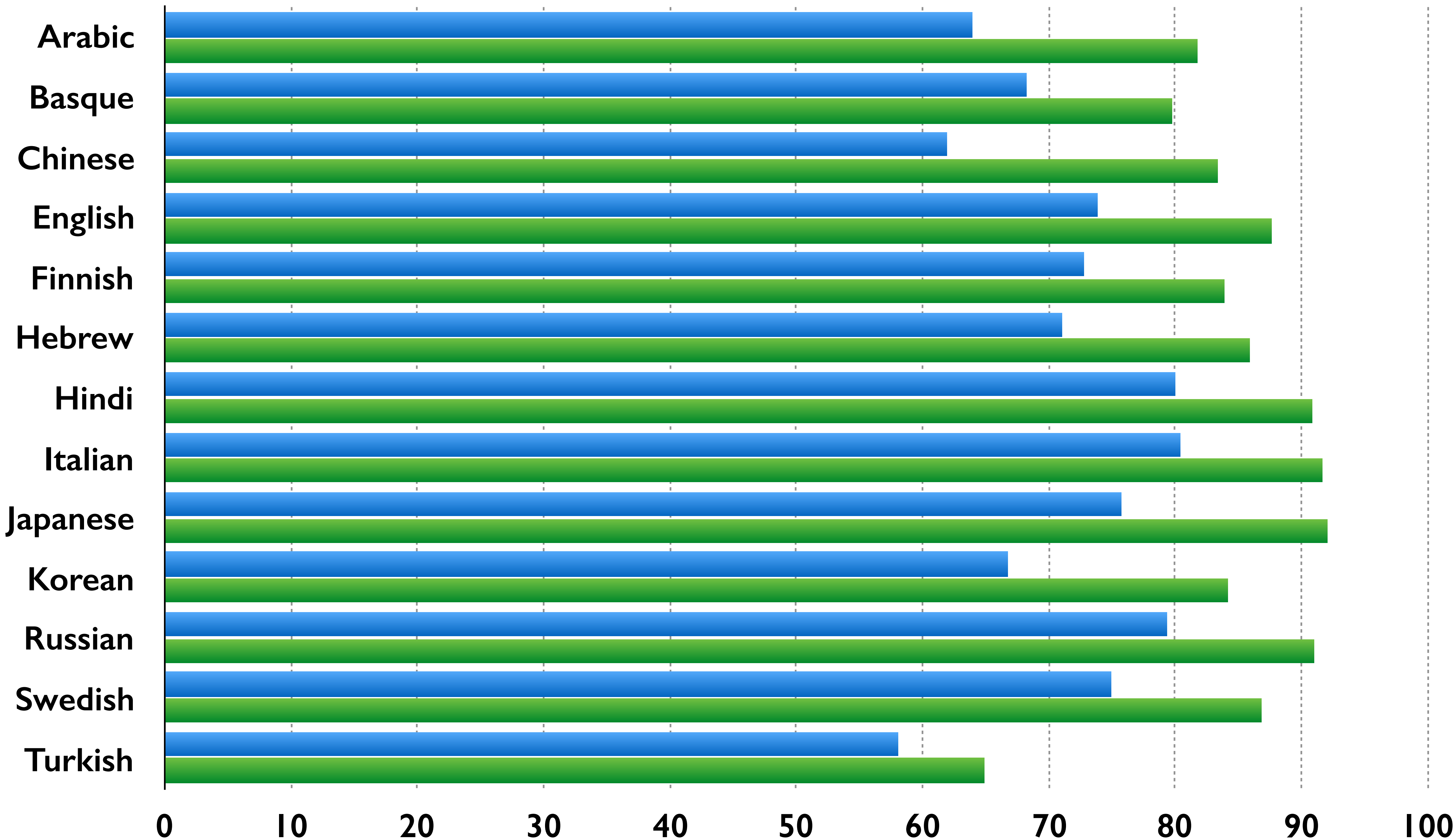
Results



Results

Mean UAS = 71.3

Mean LAS = 84.9

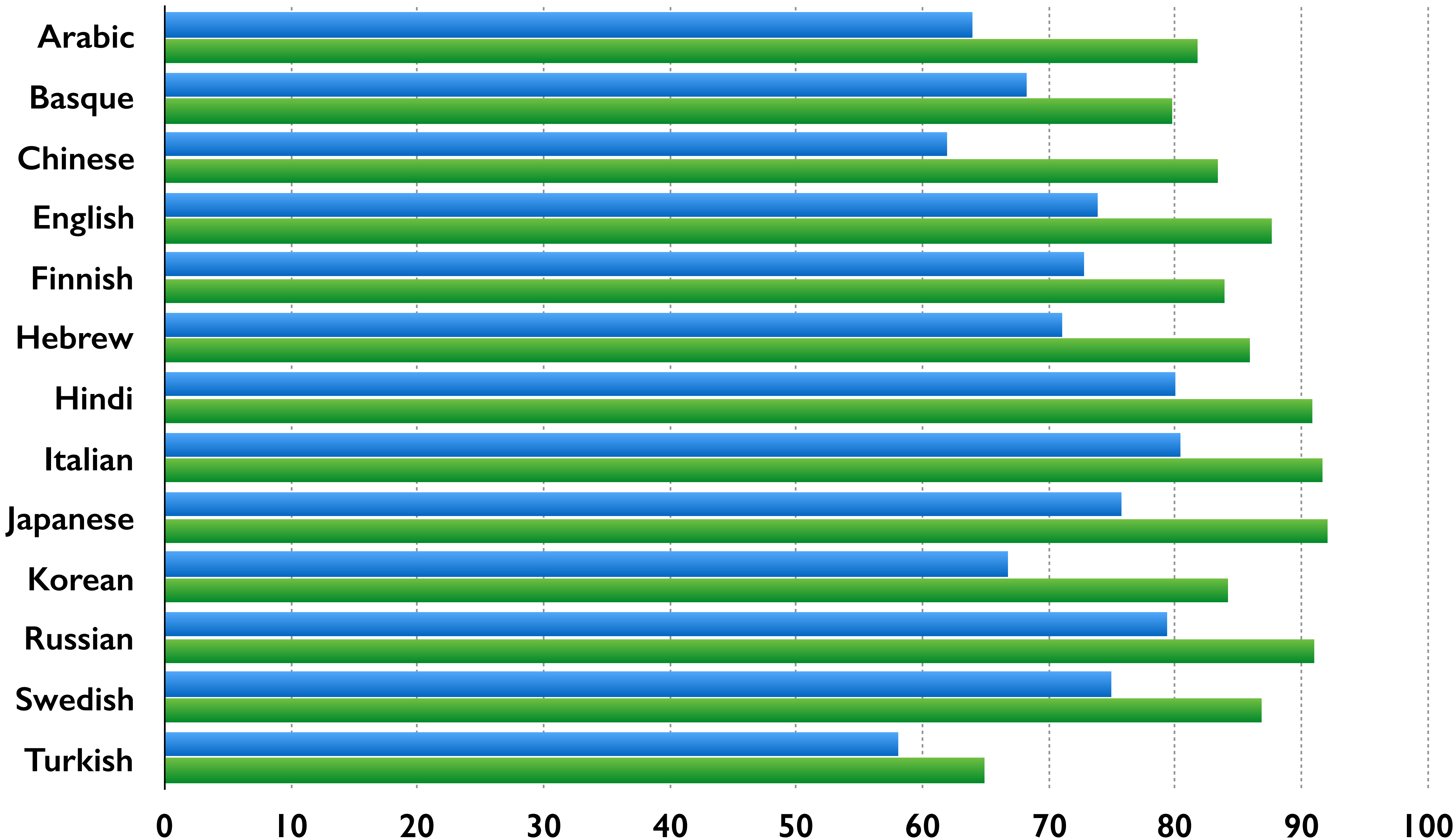


Results

Mean UAS = 71.3

Mean LAS = 84.9

Pearson's *r* = 0.49



Main Findings

- We can extract directed dependency trees from deep contextualized word representations
- Correspondence with treebank trees is substantially lower than for supervised parsers
- Variation across languages correlate with supervised parsing results

Conclusion

Conclusion

- Deep neural language models learn aspects of syntax

Conclusion

- Deep neural language models learn aspects of syntax
- Convergence across parsing models and algorithms

Conclusion

- Deep neural language models learn aspects of syntax
- Convergence across parsing models and algorithms
- No corresponding convergence across languages

Conclusion

- Deep neural language models learn aspects of syntax
- Convergence across parsing models and algorithms
- No corresponding convergence across languages
- A multilingual perspective is still important

Conclusion

- Deep neural language models learn aspects of syntax
- Convergence across parsing models and algorithms
- No corresponding convergence across languages
- A multilingual perspective is still important
- UD as a touchstone for parsing and probing studies