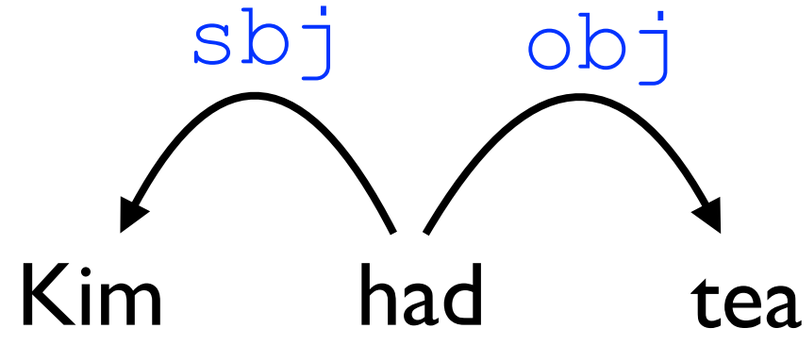
Dependency Parsing: Then and Now

NLPL Winter School '20

Joakim Nivre Artur Kulmizev

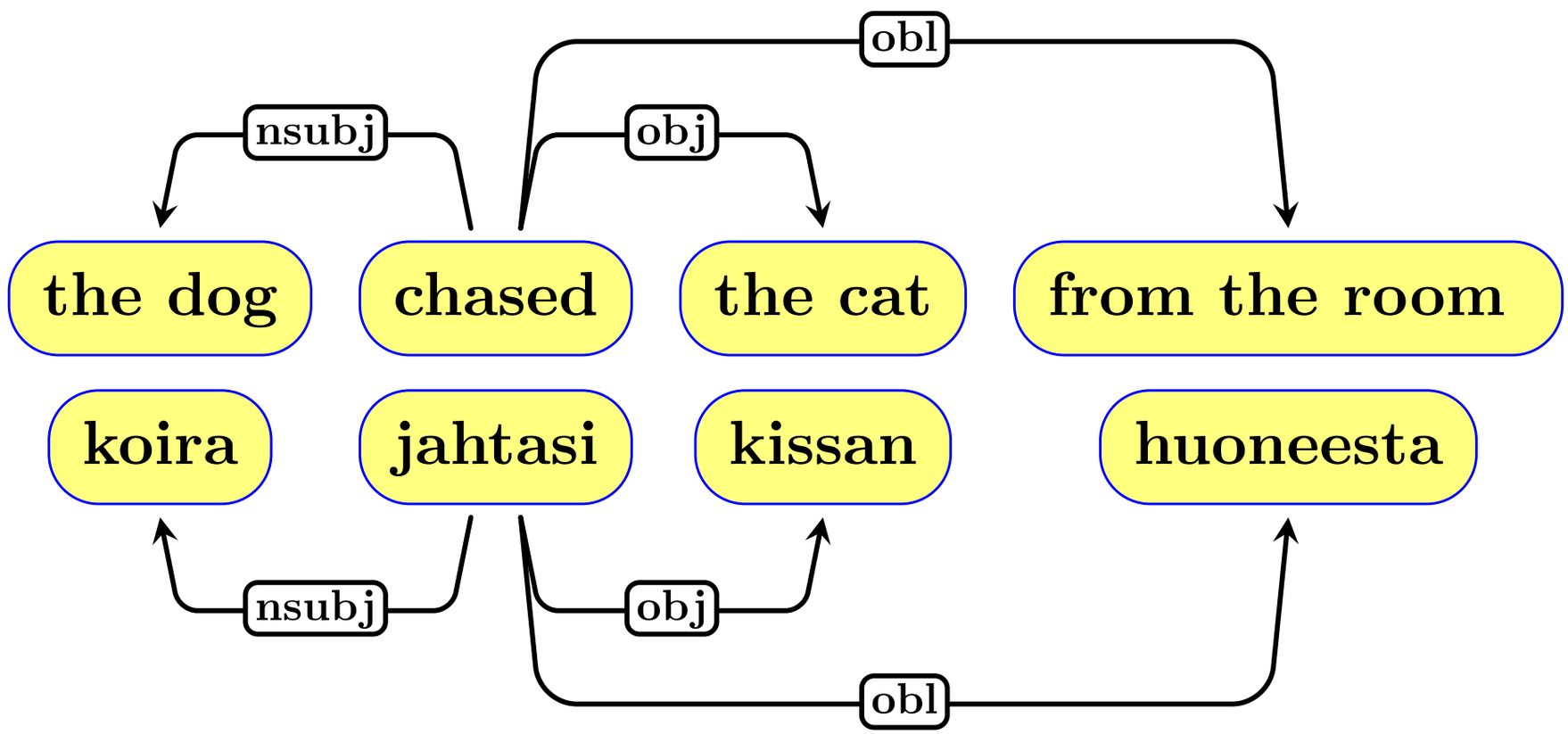
Uppsala University

Dependency Parsing

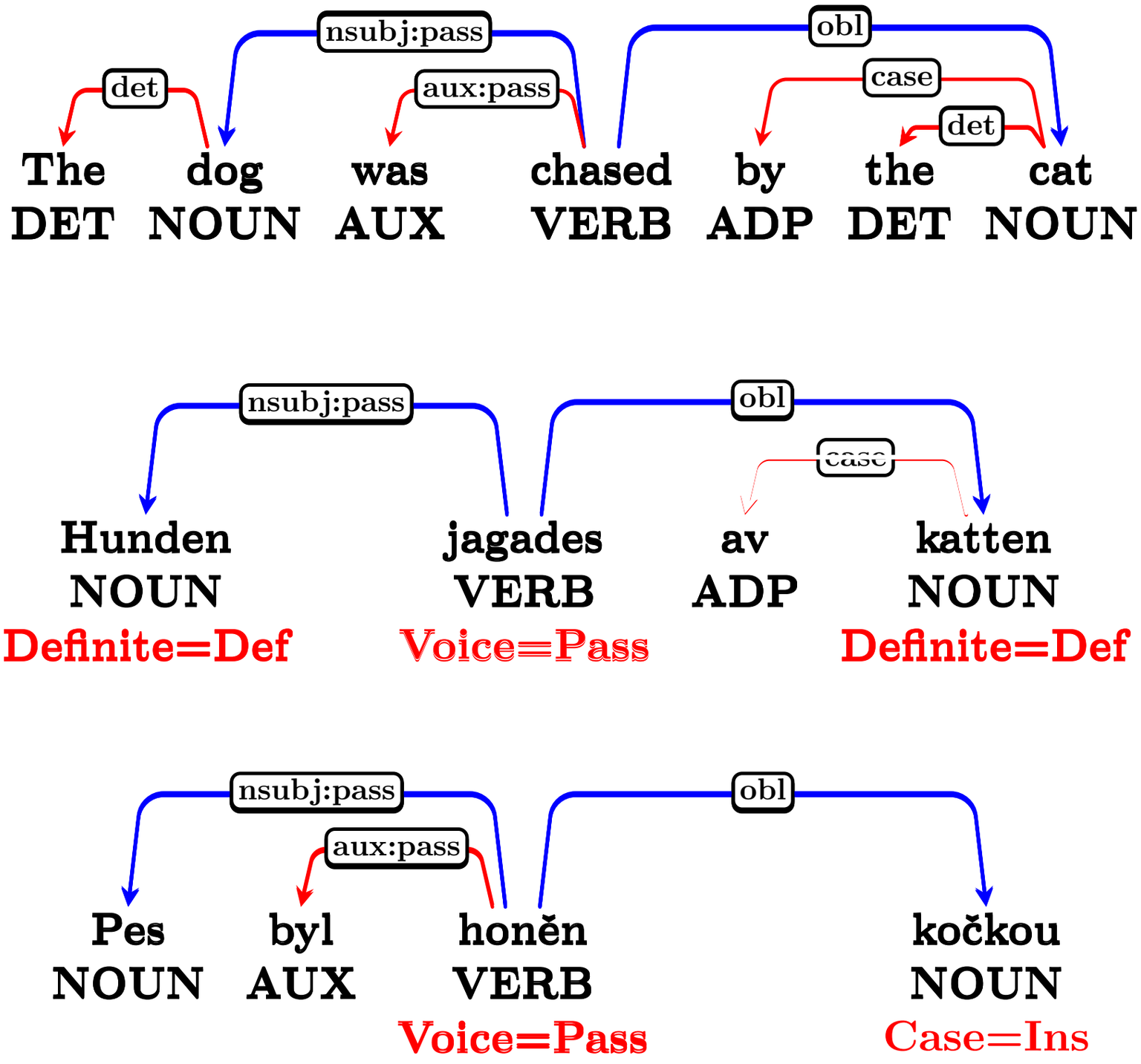


- Transparent encoding of predicate-argument structure
- Simple and efficient computational models
- Compatible with linguistic traditions around the world
- Multilingual research tradition from CoNLL 2006–2007

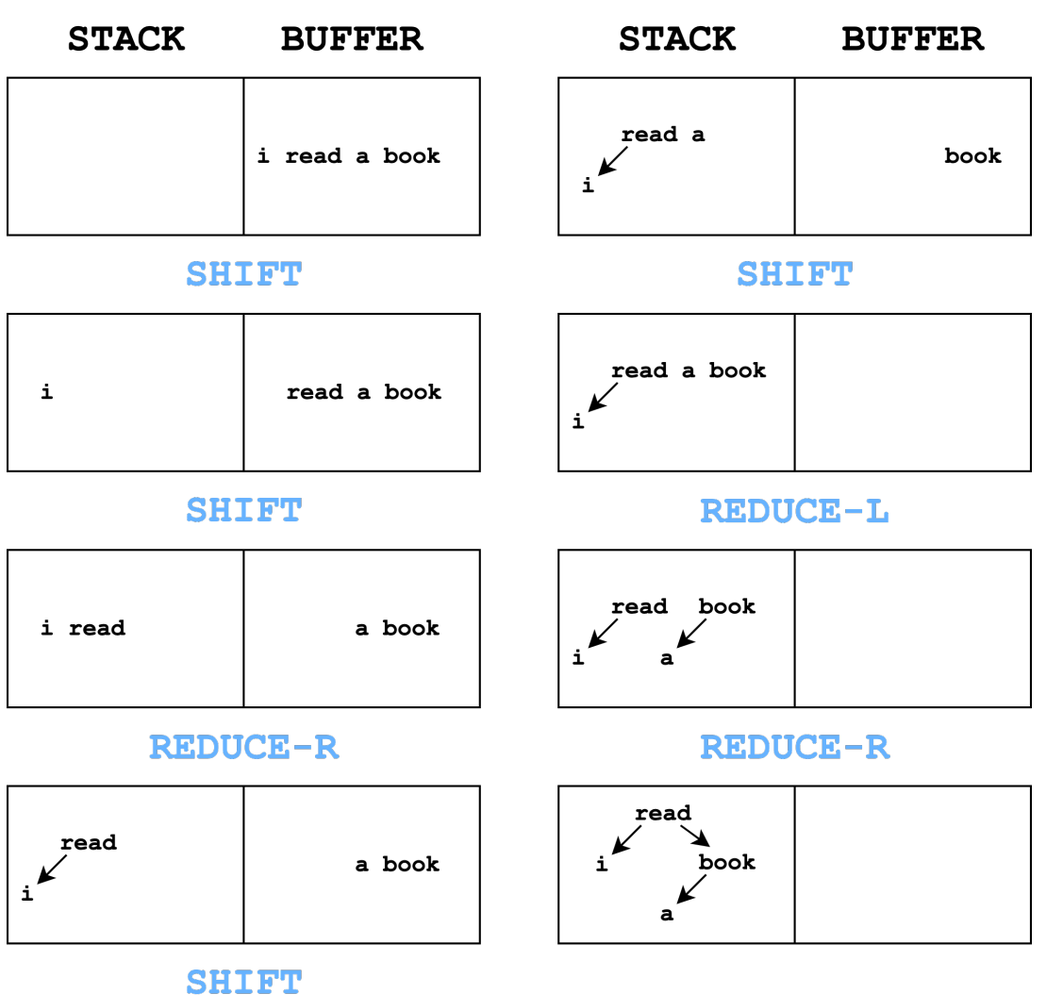
Universal Dependencies



Focus on grammatical relations between (content) words


nsubj:pass
obl
det
aux:pass
case
det
The dog was chased by the cat
DET NOUN AUX VERB ADP DET NOUN
nsubj:pass
obl
case
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 Case=Ins

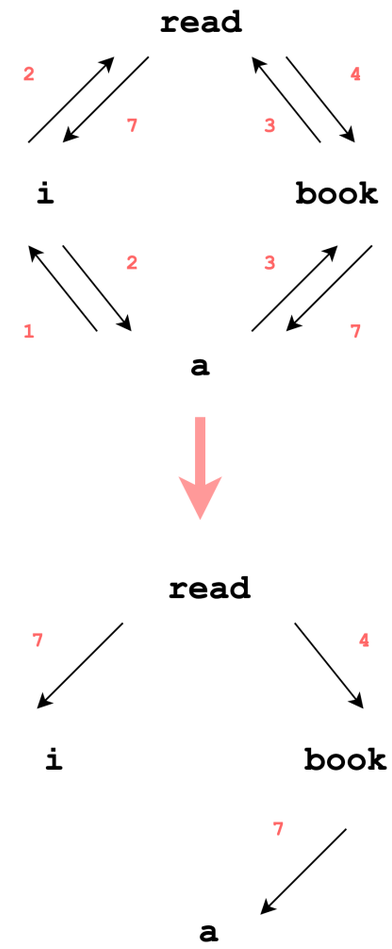
Transition-Based



Greedy search for optimal sequence of actions

Learn to score actions

Graph-Based



Exact search for highest scoring tree

Learn to score subgraphs

TB Trade-Offs

Highly efficient parsing — linear time complexity with constant time oracles and transitions

Rich history-based feature representations — no rigid constraints from inference algorithm

Sensitive to search errors and error propagation due to greedy inference and local learning

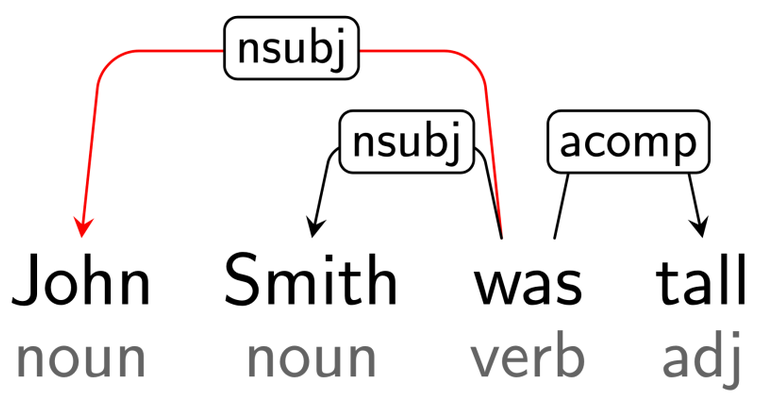
GB Trade-Offs

Learning and inference are global

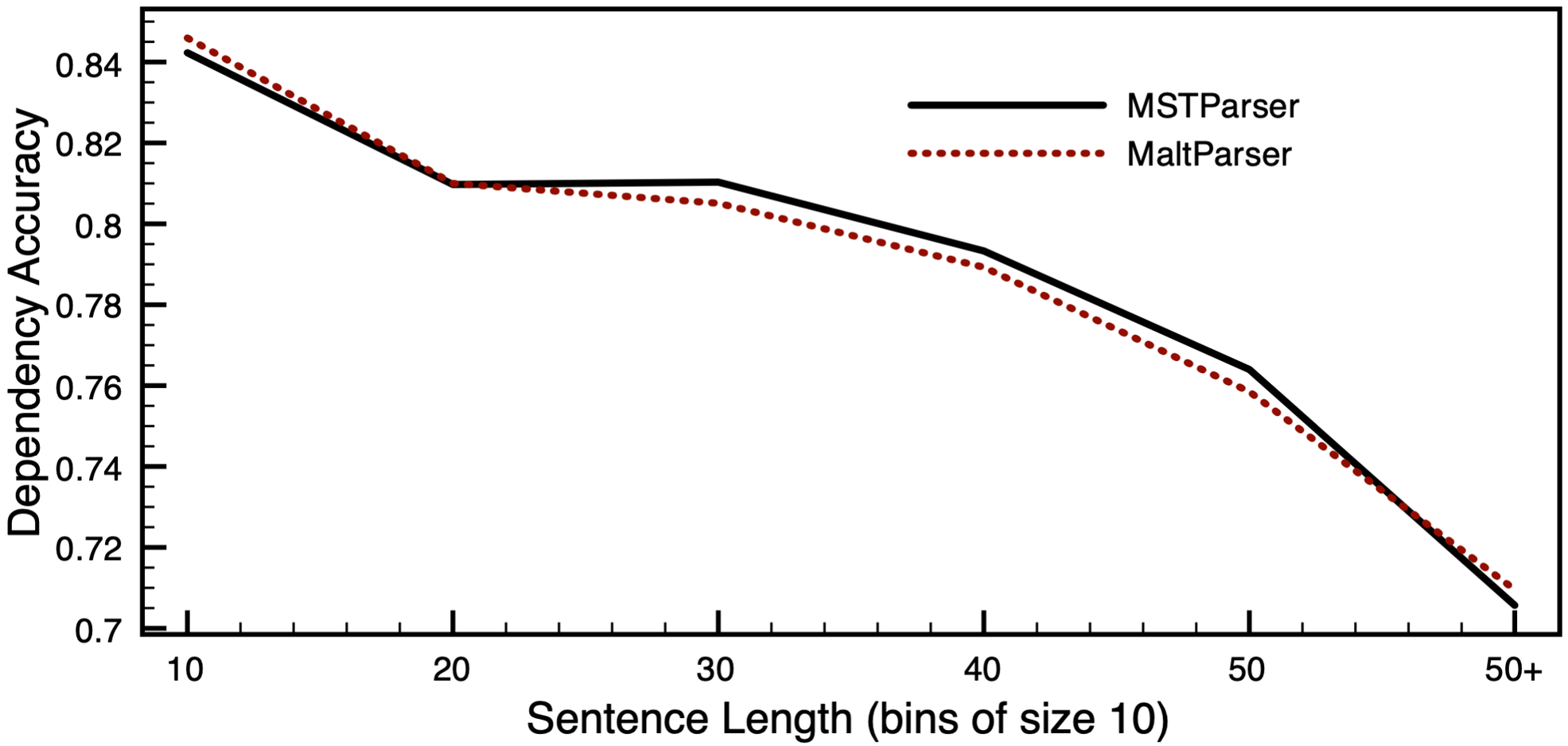
- Decoding guaranteed to find highest scoring tree
- Training algorithms use global structure learning

But this is only possible with local feature factorizations

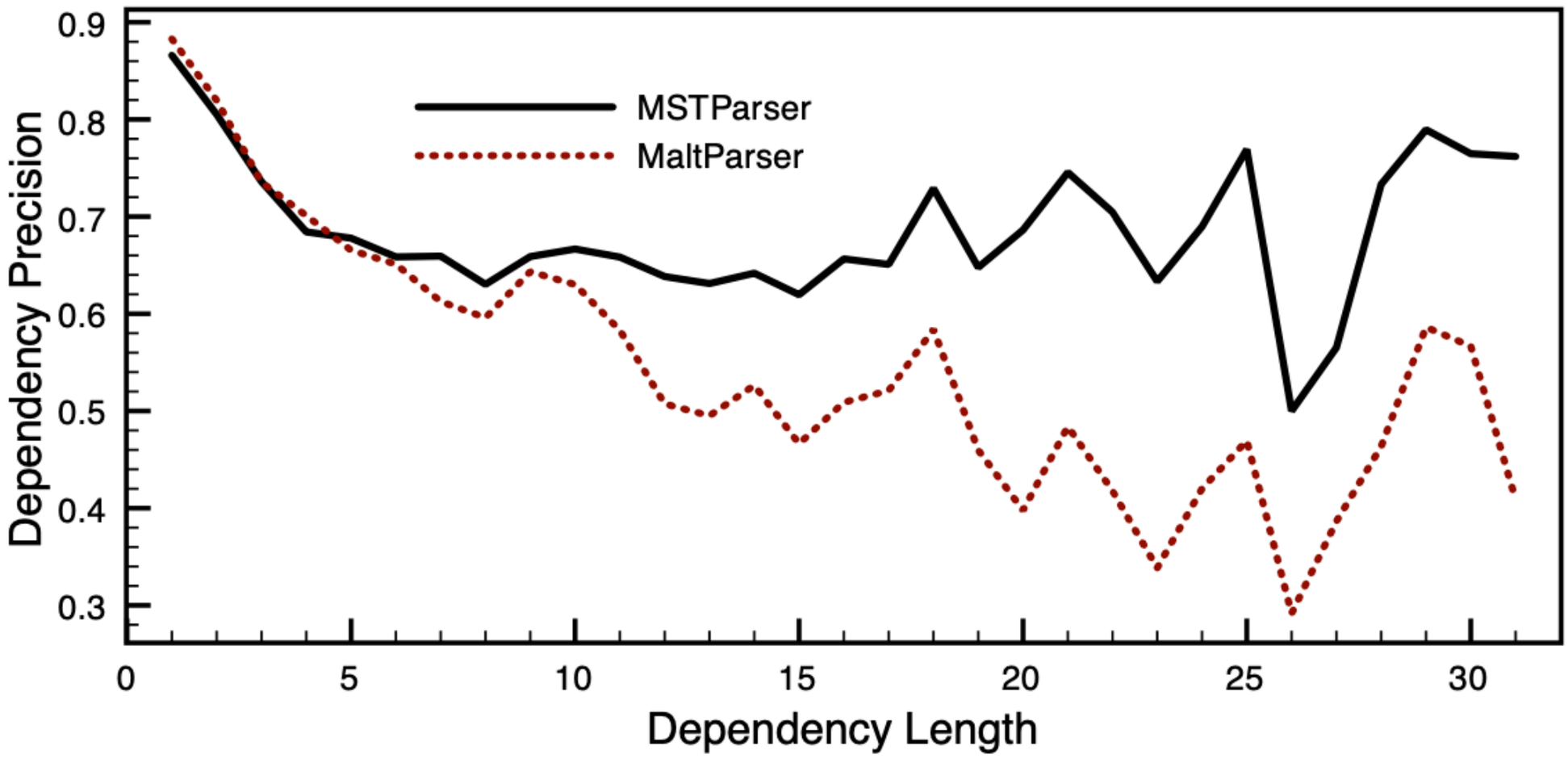
- Must limit context statistical model can look at
- Results in bad 'easy' decisions
  - For example, first-order models often predict two subjects
  - No parameter exists to discourage this



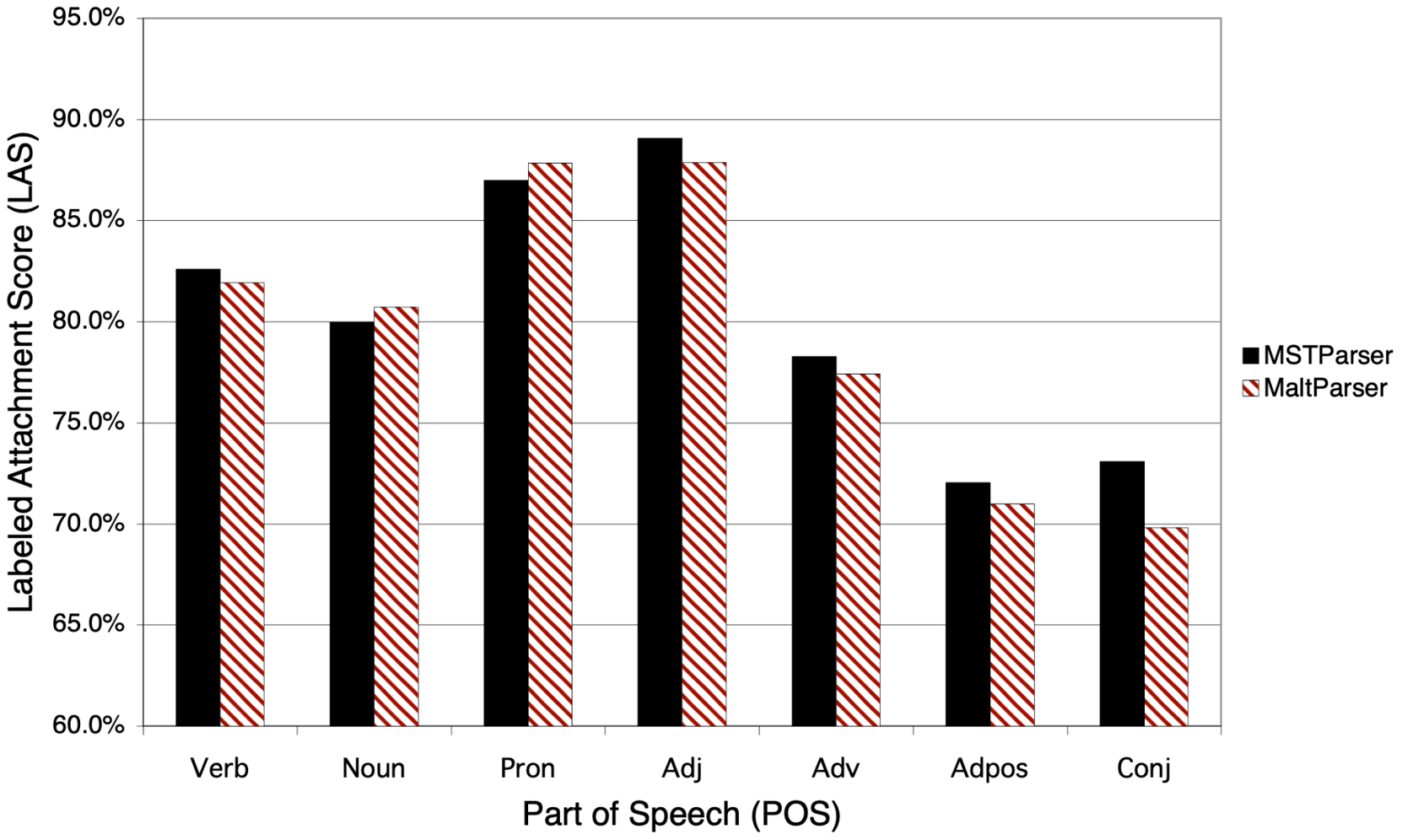
Ryan McDonald and Joakim Nivre. 2007.
Characterizing the Errors of Data-Driven Dependency Parsing Models.
***EMNLP*: 122–131.**



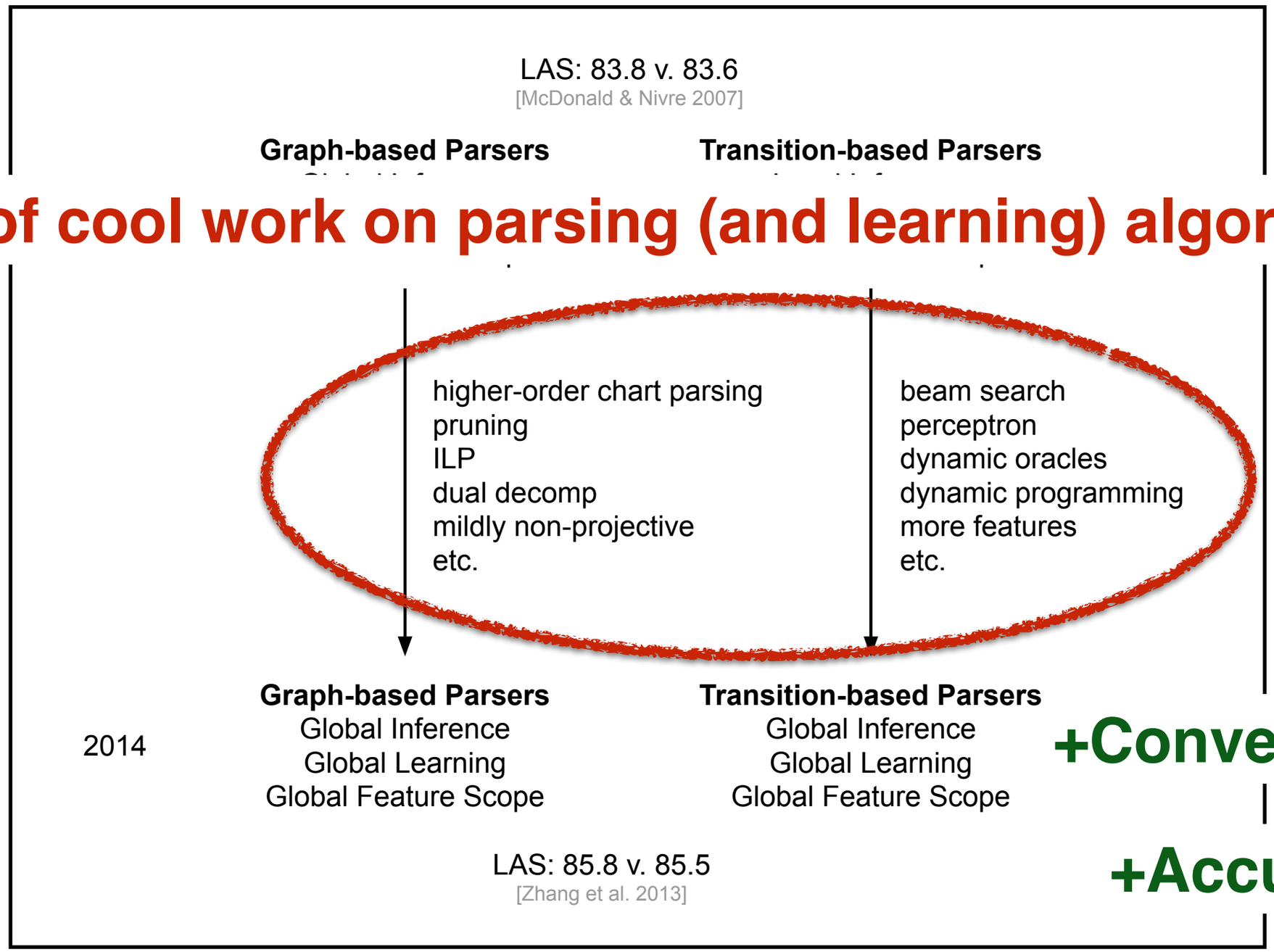
Ryan McDonald and Joakim Nivre. 2007.
Characterizing the Errors of Data-Driven Dependency Parsing Models.
***EMNLP*: 122–131.**



Ryan McDonald and Joakim Nivre. 2007.
Characterizing the Errors of Data-Driven Dependency Parsing Models.
***EMNLP*: 122–131.**



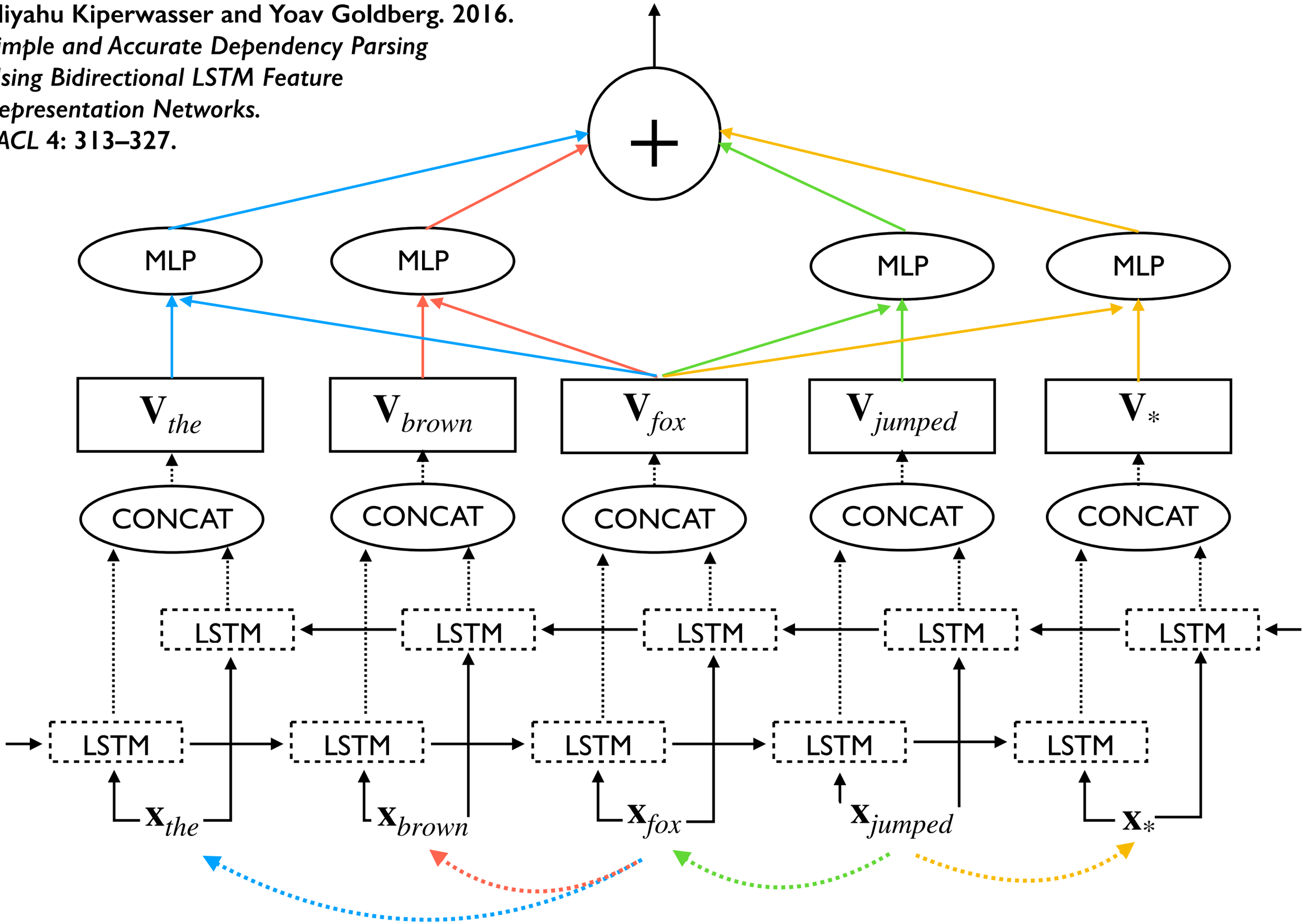
Ryan McDonald and Joakim Nivre. 2014.
Recent Advances in Dependency Parsing.
Tutorial at EACL.

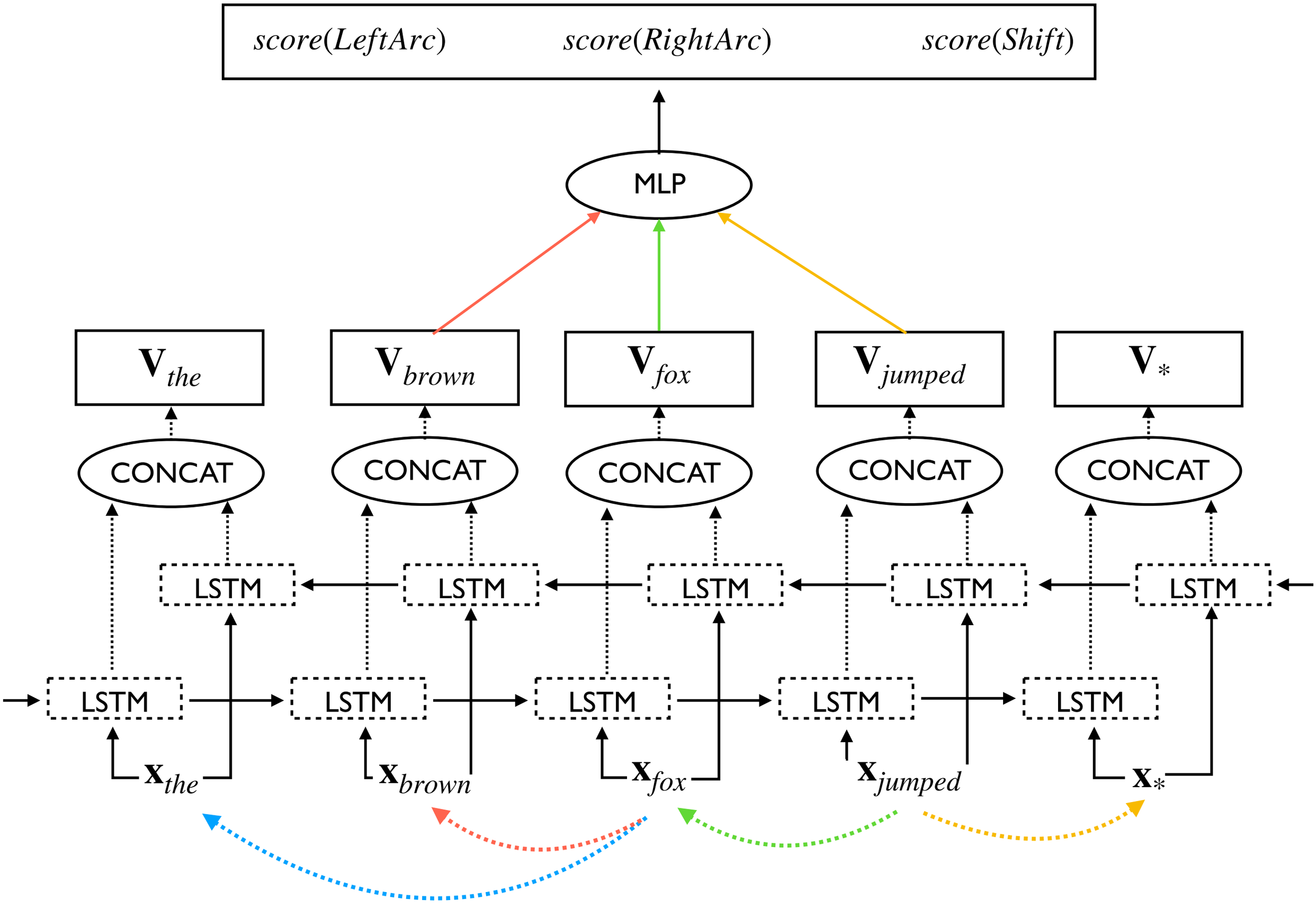


2014 and beyond…

- Neural Networks!
- Continuous word representations (`word2vec`)
- Convergence in feature encoding (`BiLSTM`)
- Simplified parsing architectures
- Convergence in error profile?

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




score(LeftArc)
score(RightArc)
score(Shift)
MLP
V the
V brown
V fox
V jumped
V *
CONCAT
CONCAT
CONCAT
CONCAT
CONCAT
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
x the
x brown
x fox
x jumped
x *

2018 and beyond…

- Language model pre-training
- ELMo, *BERT*, GPT 1&2, XLNet, etc.
- Dynamic word representations
- Features across layers of abstraction
- Capture approximate syntax out-of-the-box

Deep Contextualized Word Embeddings in Transition-Based and Graph-Based Dependency Parsing – A Tale of Two Parsers Revisited*

Artur Kulmizev Miryam de Lhoneux Johannes Gontrum Elena Fano Joakim Nivre
Department of Linguistics and Philology, Uppsala University
{artur.kulmizev,miryam.de_lhoneux,joakim.nivre}@lingfil.uu.se
{johannes.gontrum.4608,elena.fano.3249}@student.uu.se

Research Questions

1. What do the error profiles of contemporary, neural parsers look like?

2. Do deep contextualised word representations reduce errors more effectively in transition-based parsing?

3. If so, is this true for…

…longer dependencies?

…dependencies closer to root?

…longer sentences?

Parsers

transition-based (TB):

TB model from Kiperwasser and Goldberg (2016)

arc-hybrid transition system

SWAP transition (**uuparser**)

graph-based (GB):

GB model from Kiperwasser and Goldberg (2016)

arc-factored

CLE algorithm over score matrix for
max spanning tree (**uuparser**)

Input Representations

baseline:

pretrained fastText word vectors

BiLSTM character embeddings

$$\mathbf{w}_k = \mathbf{x}_k \circ \mathbf{BiLSTM}(c_{1:M})$$

Input Representations

ELMo  **:**

pretrained models by Che et al. (2018)

20 mil. words sampled from WikiDump and CommonCrawl

44 languages

$$\mathbf{ELMo}_k = \sum_{j=0}^{3} s_j \mathbf{h}_{j,k}$$

Input Representations

BERT  **:**

multilingual BERT Cased:

concatenated WikiDump for 104 languages

12 layers, 12 attn. heads

$$\mathbf{BERT}_k = \sum_{j=4}^{8} s_j \mathbf{h}_{j,k}$$

Input Representations

model	input rep.	D
baseline	$\mathbf{w}_k$	$300 + (2 * 100) = 500$
	$\mathbf{w}_k \circ \mathbf{ELMo}_k$	$300 + (2 * 100) + 1024 = 1524$
	$\mathbf{w}_k \circ \mathbf{BERT}_k$	$300 + (2 * 100) + 768 = 1278$

Languages

language	treebank	family	word order	train # sents
arabic	PADT	afro-asiatic	VSO	6.1k
basque	BDT	basque	SOV	5.4k
chinese	GSD	sino-tibetan	SVO	4.0k
english	EWT	indo-european	SVO	12.5k
finnish	TDT	finno-ugric	SVO	12.2k
hebrew	HTB	afro-asiatic	SVO	5.2k
hindi	HDTB	indo-european	SOV	13.3k
italian	ISDT	indo-european	SVO	13.1k
japanese	GSD	japanese	SOV	7.1k
korean	GSD	korean	SOV	4.4k
russian	SynTagRus	indo-european	SVO	48.8k
swedish	Talbanken	indo-european	SVO	4.3k
turkish	IMST	altaic	SOV	3.7k

Results

4.0

vs

2.9

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**

4.5

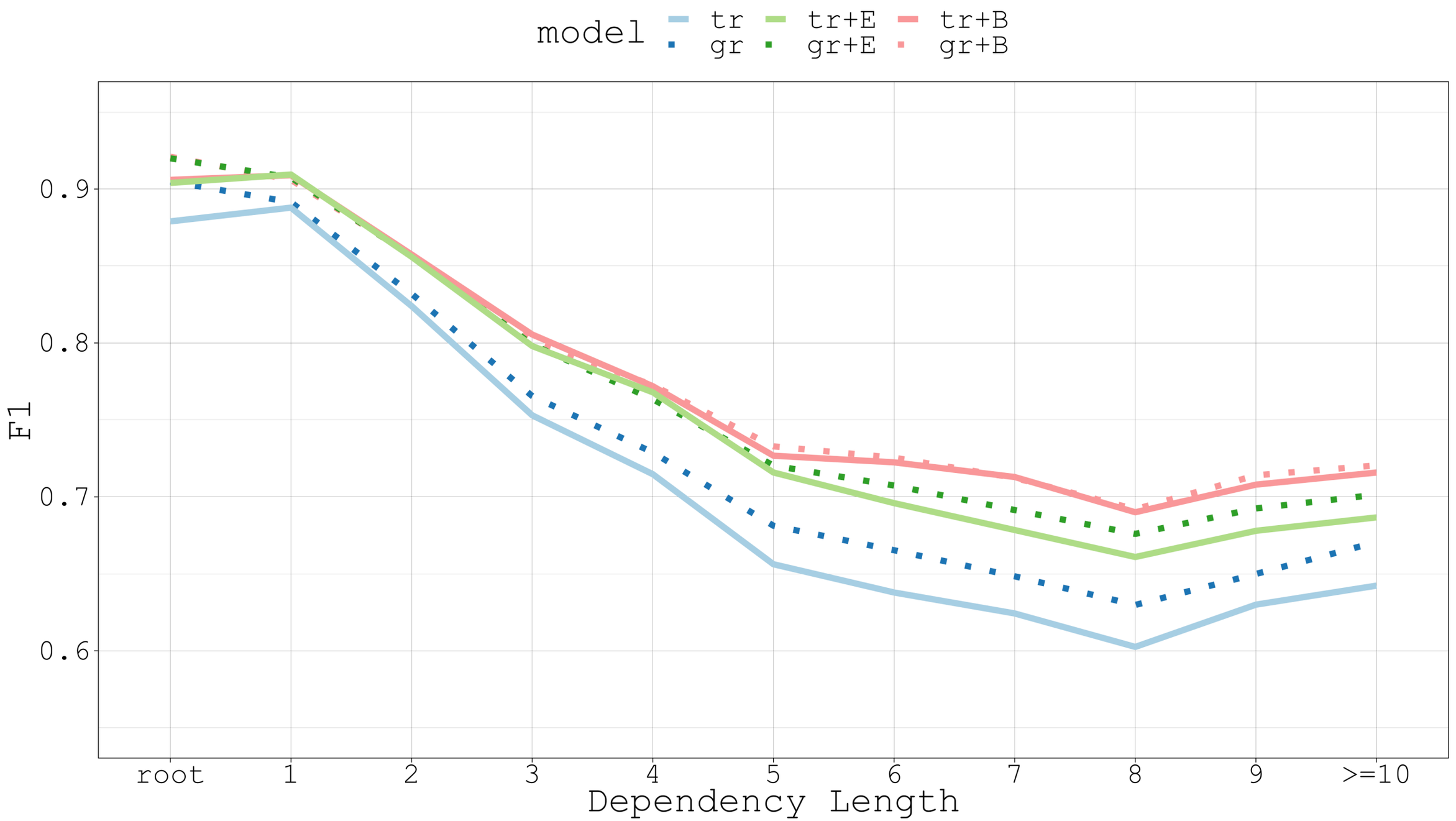
vs

3.1

Error Analysis

TB still suffers from search errors on long dependencies, though differences are smaller

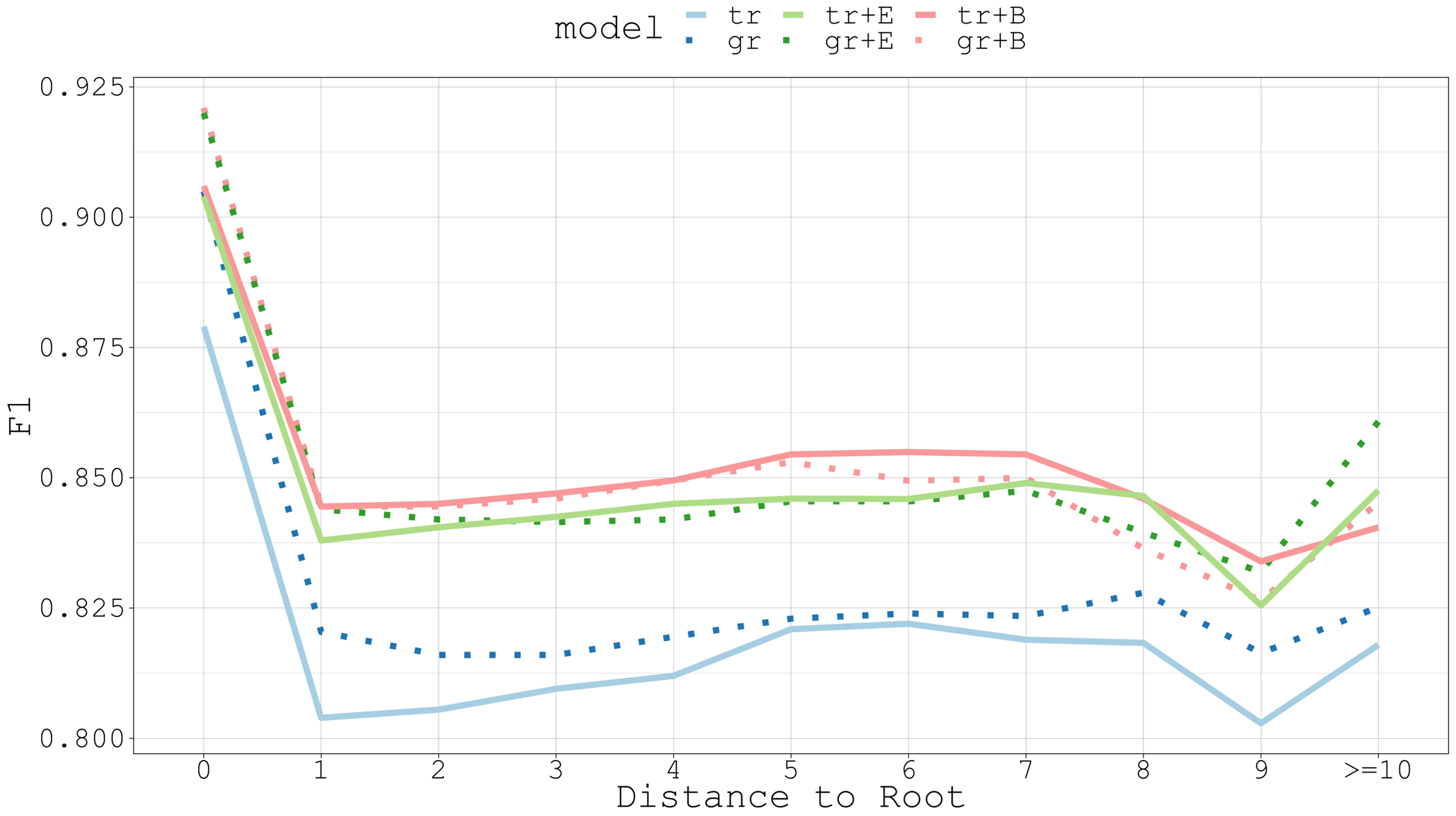
BiLSTM adds uniformity in feature reps



Error Analysis

Inverse correlation between deplen and distance to root

tr+B superior from 5-onwards

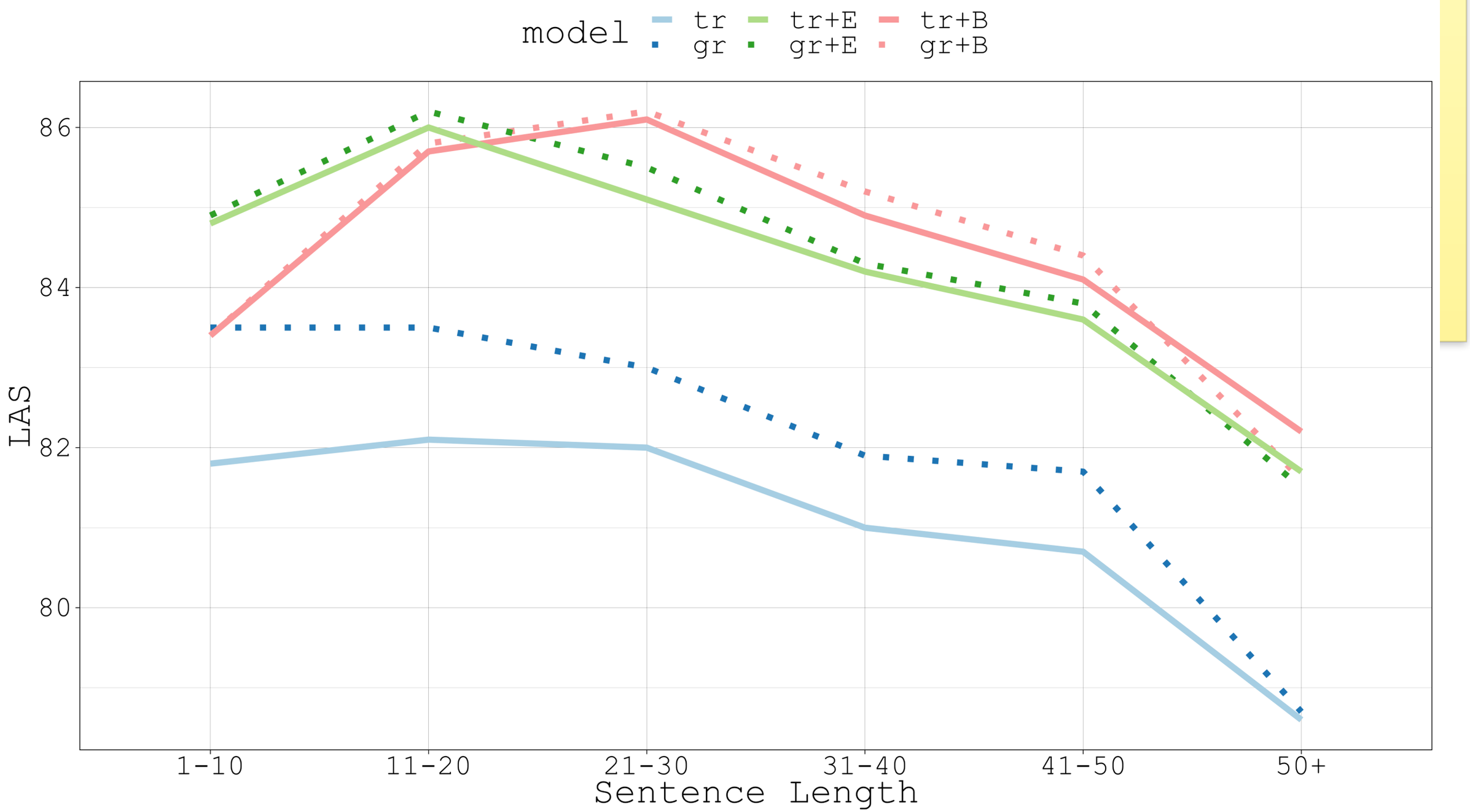


Error Analysis

GR degrades faster than TR

TR doesn't suffer from search errors as much, though

Higher accuracy sentences have a larger distribution of sentences of



Research Questions

1. What do the error profiles of contemporary, neural parsers look like?

similar for baseline, near-convergence for +E/+B

2. Do deep contextualised word representations reduce errors more effectively in transition-based parsing?

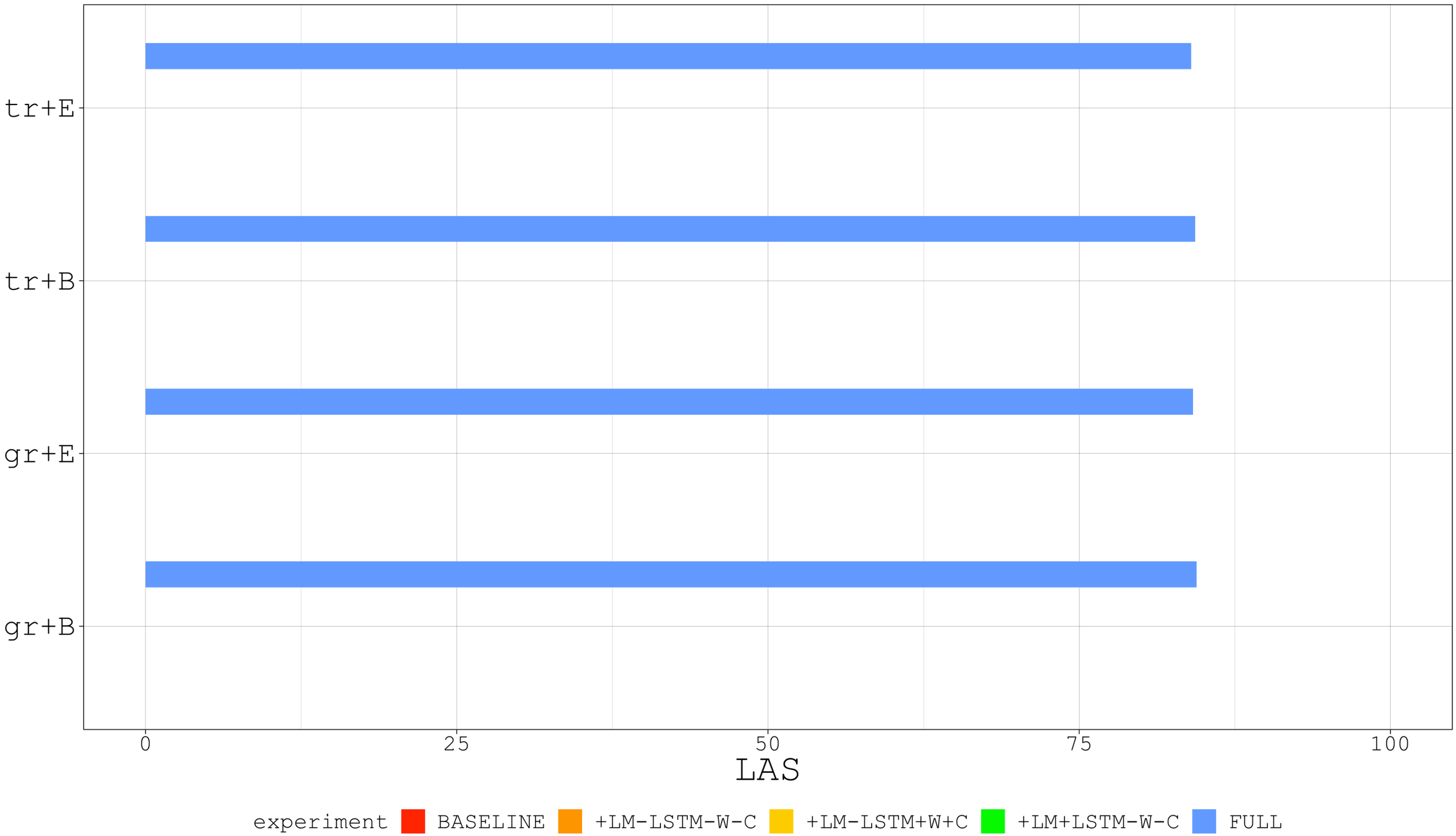
yes

… what happens if we take out the BiLSTM, etc. ?

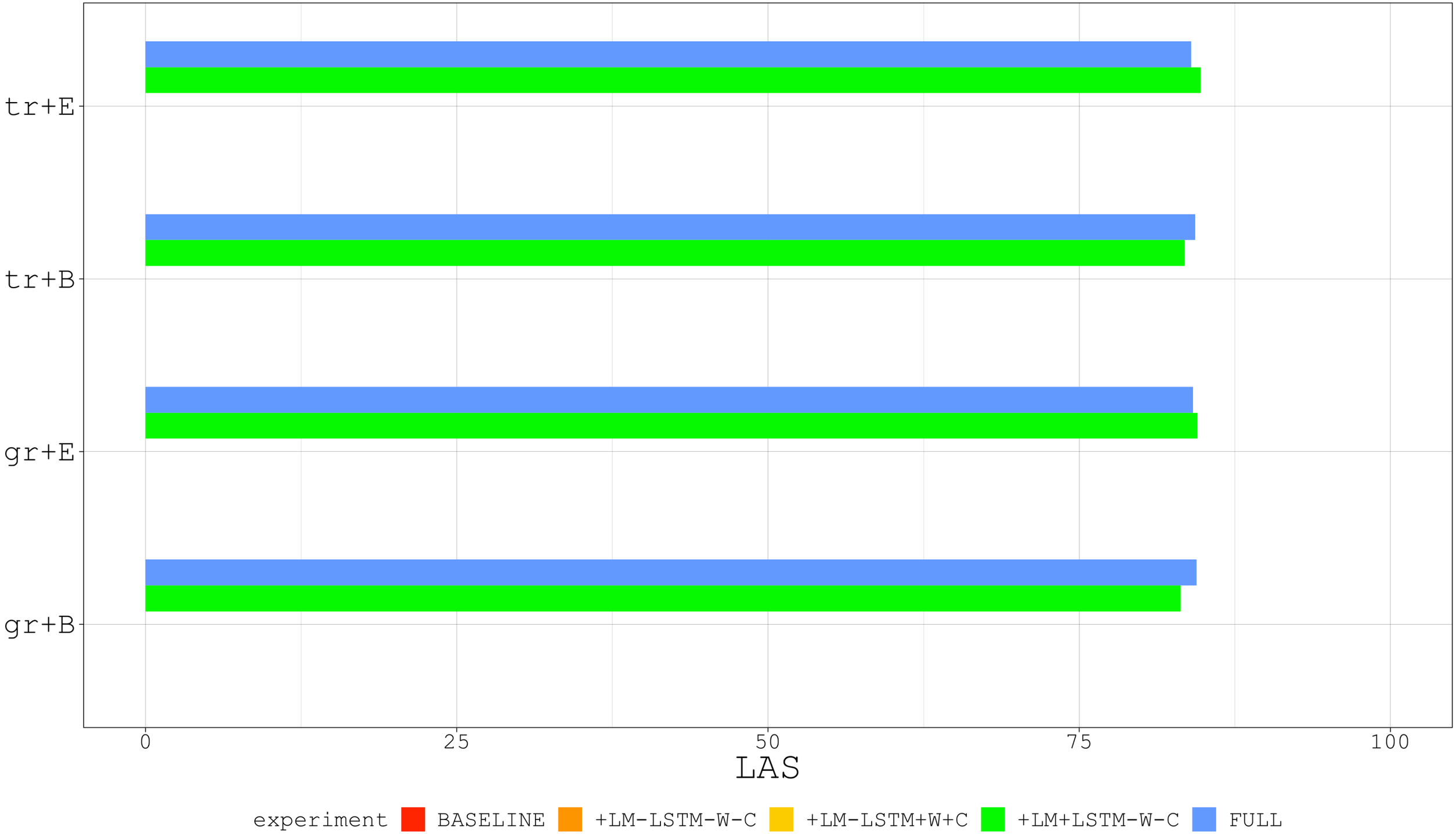
3. If so, is this true for…

…longer dependencies?	**yes**
…dependencies closer to root?	**yes**
…longer sentences?	**maybe?**

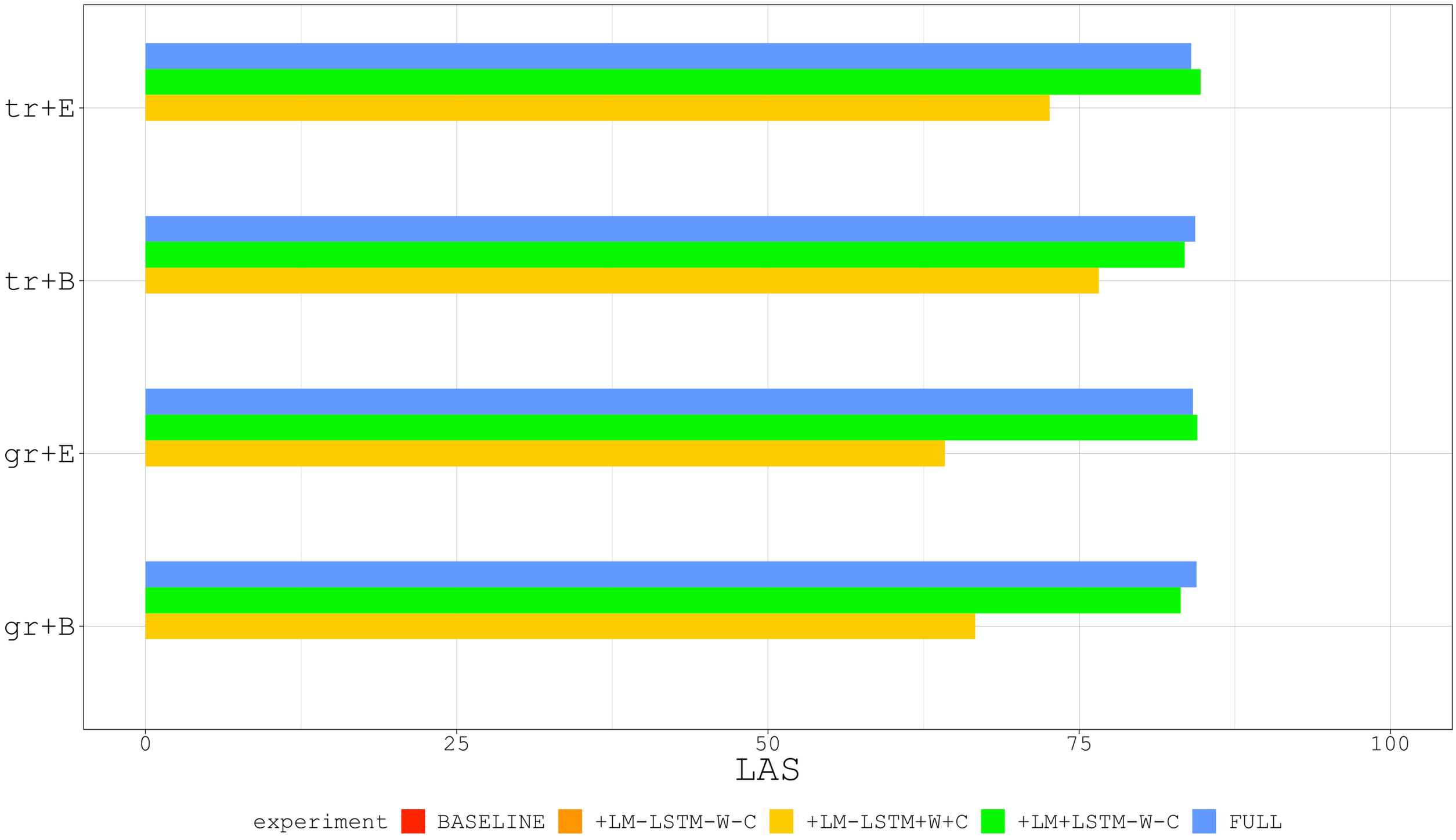
full model


tr+E
tr+B
gr+E
gr+B
0
25
50
75
100
LAS
experiment
BASELINE
+LM-LSTM-W-C
+LM-LSTM+W+C
+LM+LSTM-W-C
FULL

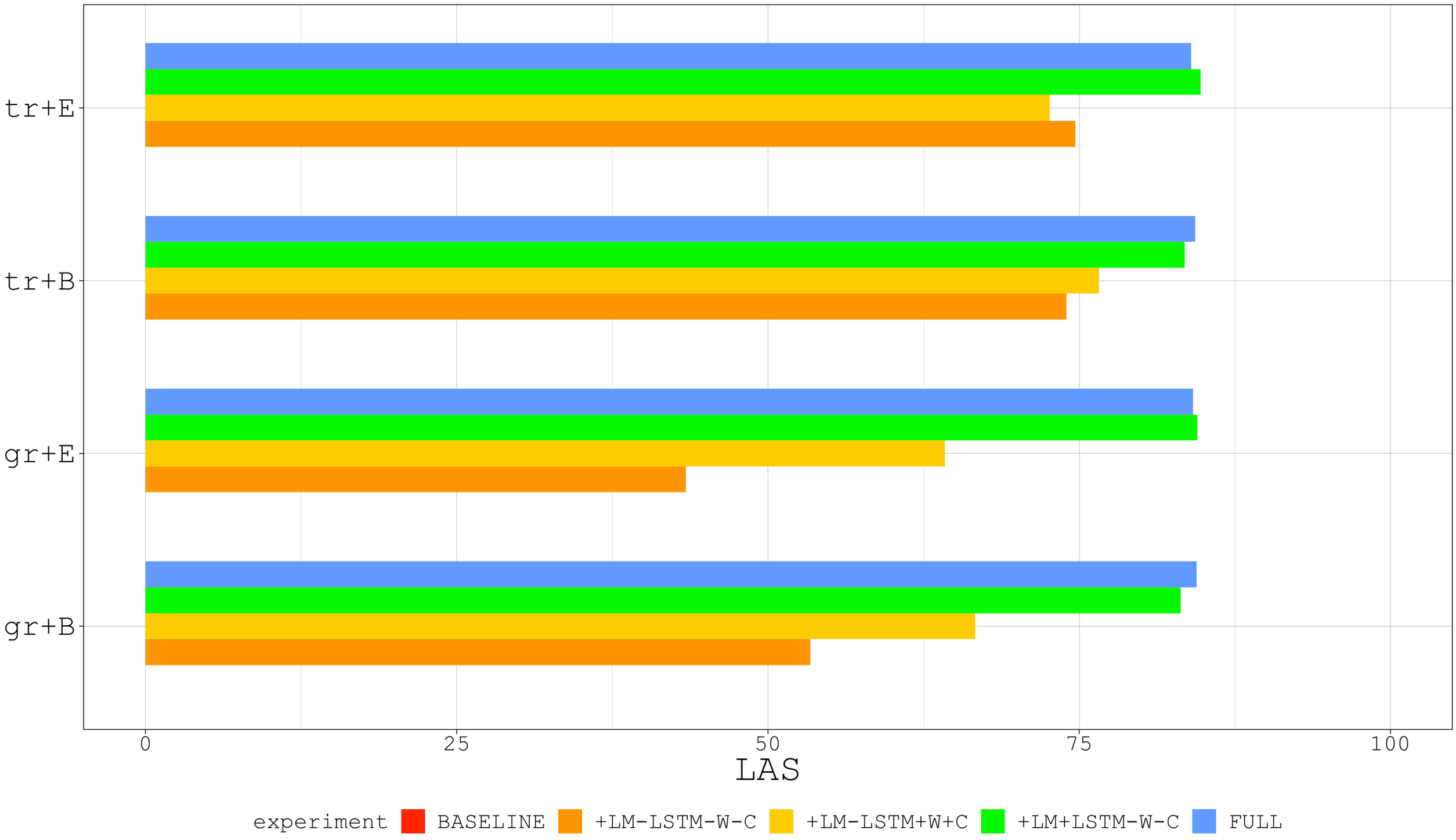
keep LSTM, remove W + C embeds



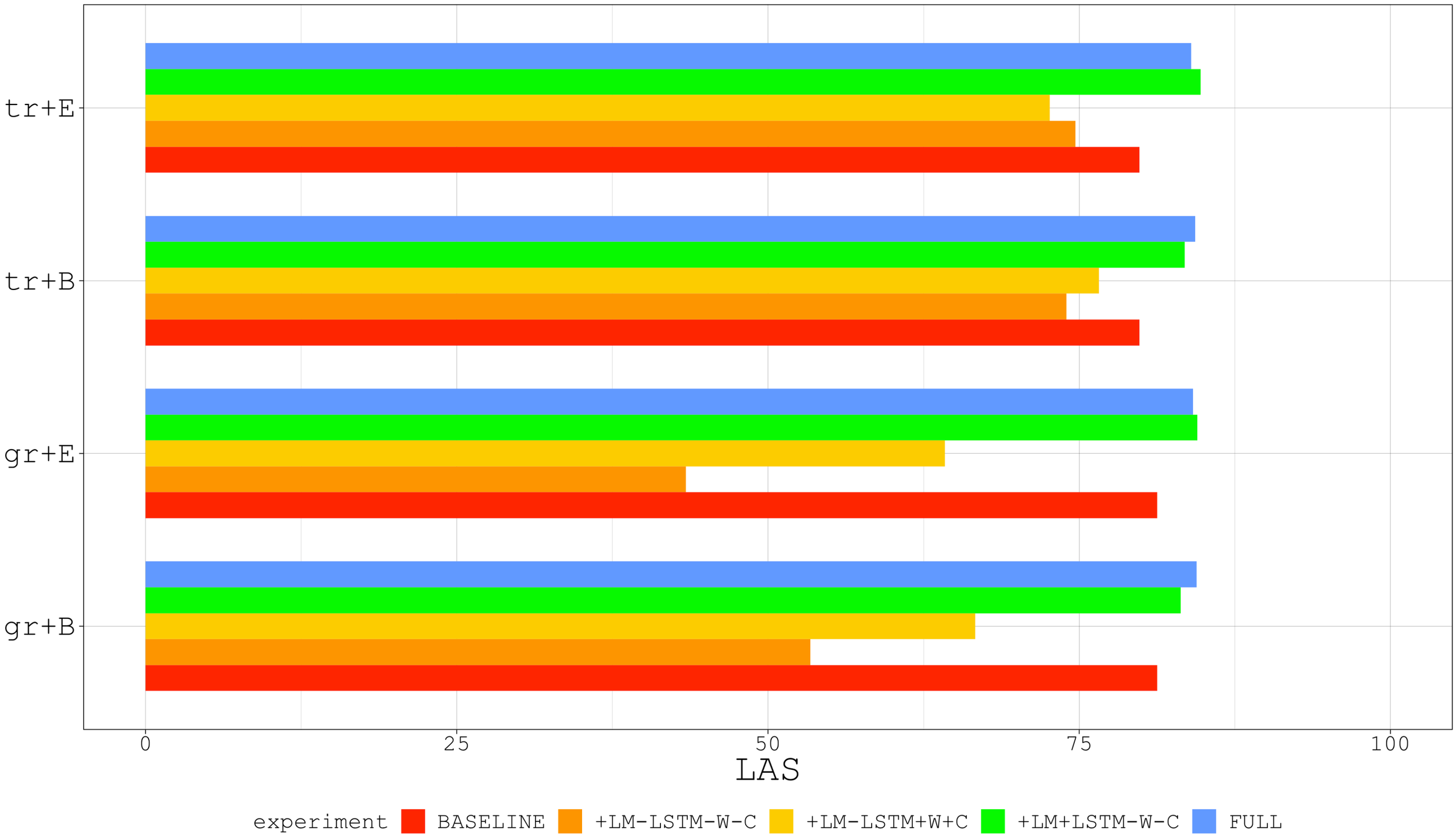
remove LSTM, keep W + C embeds



remove LSTM, remove W + C embeds



baseline


tr+E
tr+B
gr+E
gr+B
0
25
50
75
100
LAS
experiment
BASELINE
+LM-LSTM-W-C
+LM-LSTM+W+C
+LM+LSTM-W-C
FULL

Language Models

- Obvious boon for dependency parsing
- Ubiquitous in contemporary NLP
- SOA on virtually all tasks when pretraining / fine-tuning

BERTology 😔

BERT Rediscovers the Classical NLP Pipeline

Ian

{ifter

What Does BERT Look At?
An Analysis of BERT's Attention

Kevin Clark[†]

{kev

Open Sesame: Getting Inside BERT's Linguistic Knowledge

Yongjie Lin[a,*] and **Yi Chern Tan**[a,*] and **Robert Frank**[b]
[a]Department of Computer Science, Yale University

Investigating BERT's Knowledge of Language:
Five Analysis Methods with NPIs

{y

Alex Warstadt,[†,1,2] **Y**
Anna Alsop,[†,2] **Shikha Bordi**
Anhad Mohananey,

WHAT DO YOU LEARN FROM CONTEXT? PROBING FOR SENTENCE STRUCTURE IN CONTEXTUALIZED WORD REPRESENTATIONS

Ian Tenney,[*1] **Patric**
R. Thomas McCoy,[2]
Dipanjan Das,[1] **and**

[1]Google AI Language
[4]New York University

Revealing the Dark Secrets of BERT

Olga Kovaleva, Alexey Romanov, Anna Rogers, Anna Rumshisky
Department of Computer Science
University of Massachusetts Lowell
Lowell, MA 01854
{okovalev,arum,aromanov}@cs.uml.edu

… and many more…

A Structural Probe for Finding Syntax in Word Representations

John Hewitt
Stanford University
johnhew@stanford.edu

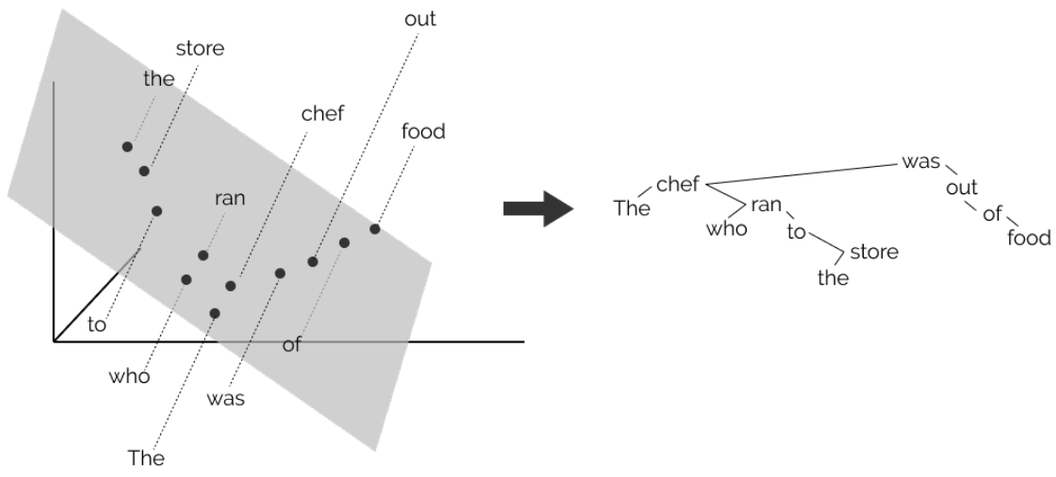
Christopher D. Manning
Stanford University
manning@stanford.edu

much evidence that linguistic hierarchy is captured by neural LMs

if models capture syntax, they don't devote entire spaces to do so

distance and depth natural properties of syntax trees and vector spaces

theoretically possible to train probe to retrieve these properties



A Structural Probe for Finding Syntax in Word Representations

John Hewitt
Stanford University
johnhew@stanford.edu

Christopher D. Manning
Stanford University
manning@stanford.edu

distance probe

$$\min_{B} \sum_{l=1}^{L} \frac{1}{|n^l|^2} \sum_{i,j} |\, d_{T^l}(w_i^l, w_j^l) - d_B(\mathbf{h}_i^l, \mathbf{h}_j^l)^2 \,|$$

depth probe

$$\min_{B} \sum_{l=1}^{L} \frac{1}{n_l} \sum_{i}^{n} (\|w_i^l\| - \|B\mathbf{h}_i^l\|^2)$$

UD embraced by NLP community

(some) linguists remain skeptical

SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD

Kim Gerdes*, Bruno Guillaume†, Sylvain Kahane◊, Guy Perrier†
*LPP, Sorbonne Nouvelle & CNRS
†Loria, Université de Lorraine & CNRS & INRIA, Nancy;
◊Modyco, Université Paris Nanterre & CNRS
kim@gerdes.fr, bruno.guillaume@inria.fr,
sylvain@kahane.fr, guy.perrier@loria.fr

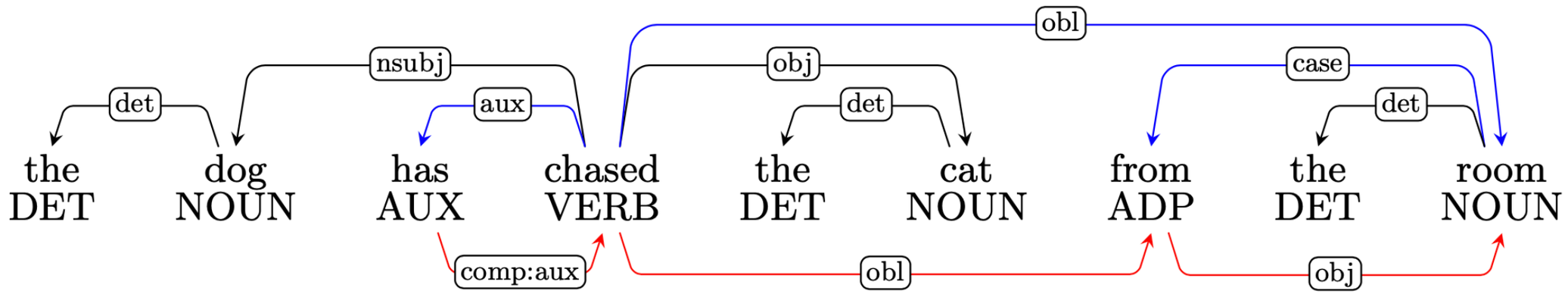
treebank by definition less interesting for typological research on syntax. In particular, UD does not account for the hierarchy between functional words and tends to flatten syntactic structures. The content-word-centric annotation is also problematic for the internal cohesion of the treebank (cf. the difficulty of coherently annotating complex prepositions that usually contain a content word, Gerdes & Kahane 2016) and it marks a break with syntactic traditions, where headedness is defined by distributional properties of individual words (Bloomfield 1933), see Section 2.[1]

an (isomorphic) alternative to UD

function-word heads

chain-like analysis of conjuncts

UD vs. SUD


obl
nsubj
obj
case
det
aux
det
det
the DET
dog NOUN
has AUX
chased VERB
the DET
cat NOUN
from ADP
the DET
room NOUN
comp:aux
obl
obj

Research Questions

1. Does the syntax captured by LMs align more with UD or SUD representations?

2. Do different languages play a role in how well parse trees are captured by such models?

Input Representations

ELMo  **:**

pretrained models by Che et al. (2018)

20 mil. words sampled from WikiDump and CommonCrawl

44 languages

$$\mathbf{ELMo}_k = \sum_{j=0}^{3} s_j \mathbf{h}_{j,k}$$

Input Representations

BERT  **:**

multilingual BERT Cased:

concatenated WikiDump for 104 languages

12 layers, 12 attn. heads

$$\mathbf{BERT}_k = \sum_{j=1}^{12} s_j \mathbf{h}_{j,k}$$

Directed Dependency Trees

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

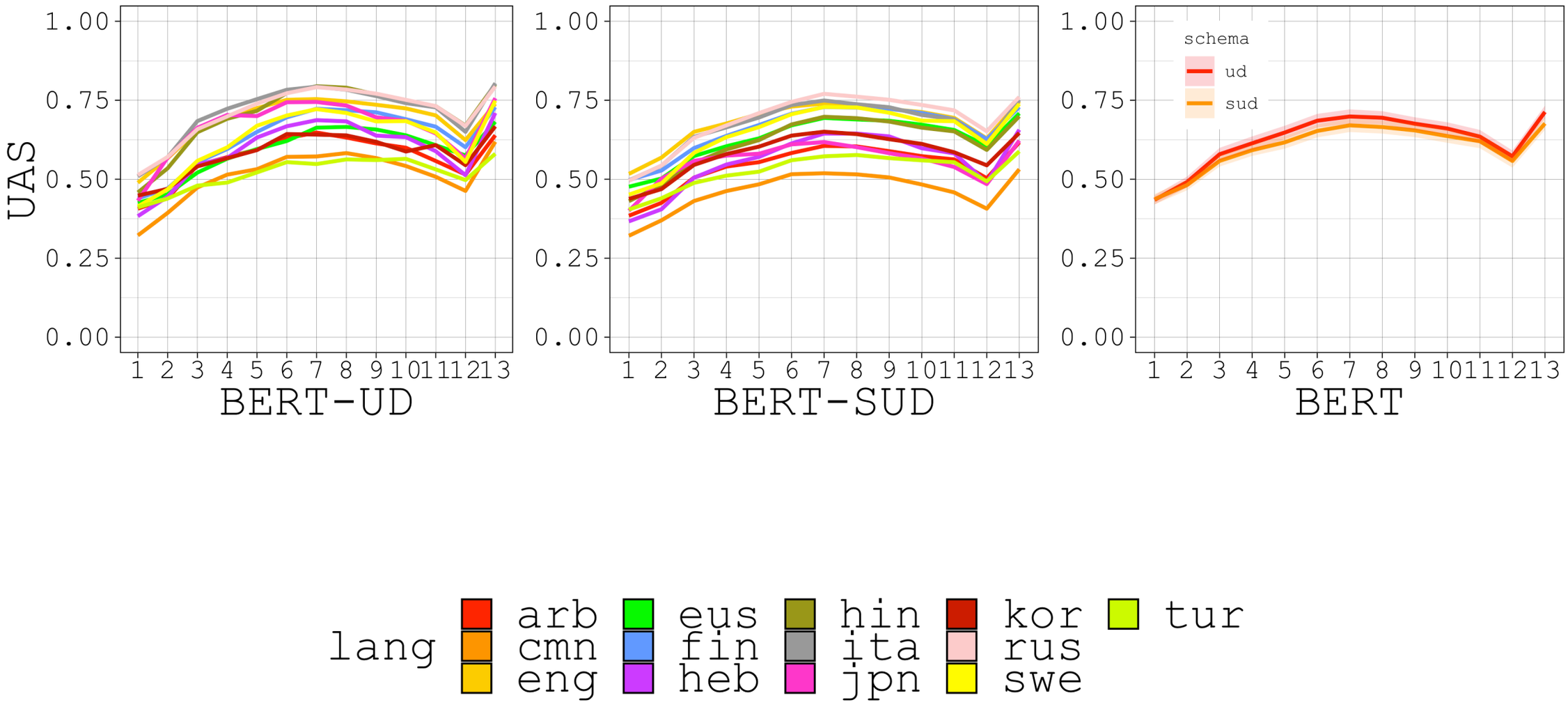
$$score(w_i, w_j) = \begin{cases} -dist(w_i, w_j) & \text{if } depth(w_i) < depth(w_j) \\ -\infty & \text{otherwise} \end{cases}$$

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

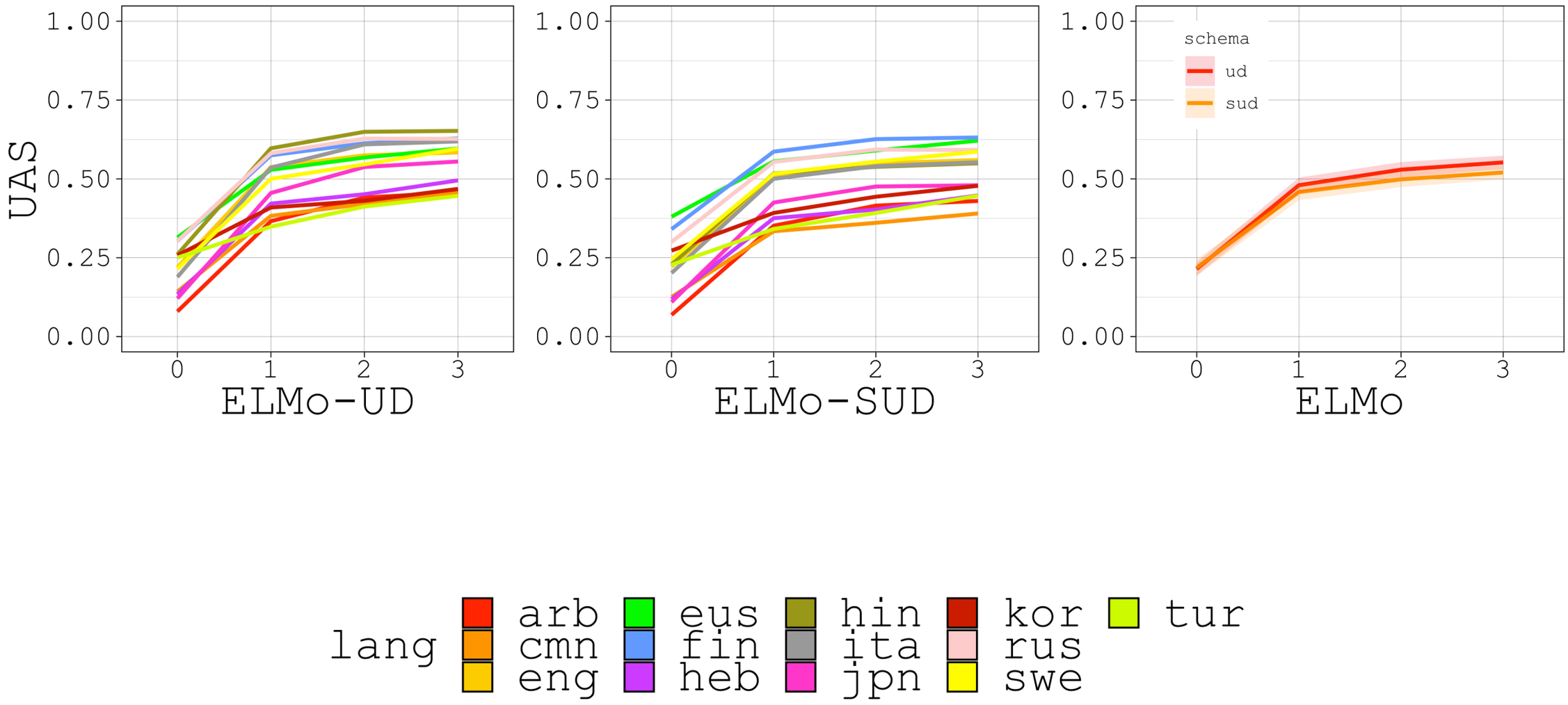
Data

Language	Code	Treebank	# Sents	%ADP	%AUX	%ContRel		Dep Len		Height	
						UD	SUD	UD	SUD	UD	SUD
Arabic	arb	PADT	6075	15	1	37	24	4.17	3.92	7.20	9.82
Chinese	cmn	GSD	3997	5	3	37	30	3.72	3.74	4.30	6.56
English	eng	EWT	12543	8	6	20	12	3.13	2.94	3.48	5.11
Basque	eus	BDT	5396	2	13	34	25	2.99	2.90	3.49	4.18
Finnish	fin	TDT	12217	2	7	35	30	2.98	2.91	3.42	4.22
Hebrew	heb	HTB	5241	14	2	28	14	3.76	3.53	5.07	7.30
Hindi	hin	HDTB	13304	22	9	26	10	3.44	3.05	4.25	7.41
Italian	ita	ISDT	13121	14	5	21	8	3.30	3.12	4.21	6.28
Japanese	jap	GSD	7125	25	14	31	10	2.49	2.08	4.40	8.18
Korean	kor	GSD	4400	2	0	58	57	2.20	2.17	3.86	4.07
Russian	rus	SynTagRus	48814	10	1	31	22	3.28	3.13	4.21	5.24
Swedish	swe	Talbanken	4303	12	5	29	17	3.14	2.98	3.50	5.02
Turkish	tur	IMST	3664	3	2	33	30	2.21	2.12	3.01	3.37

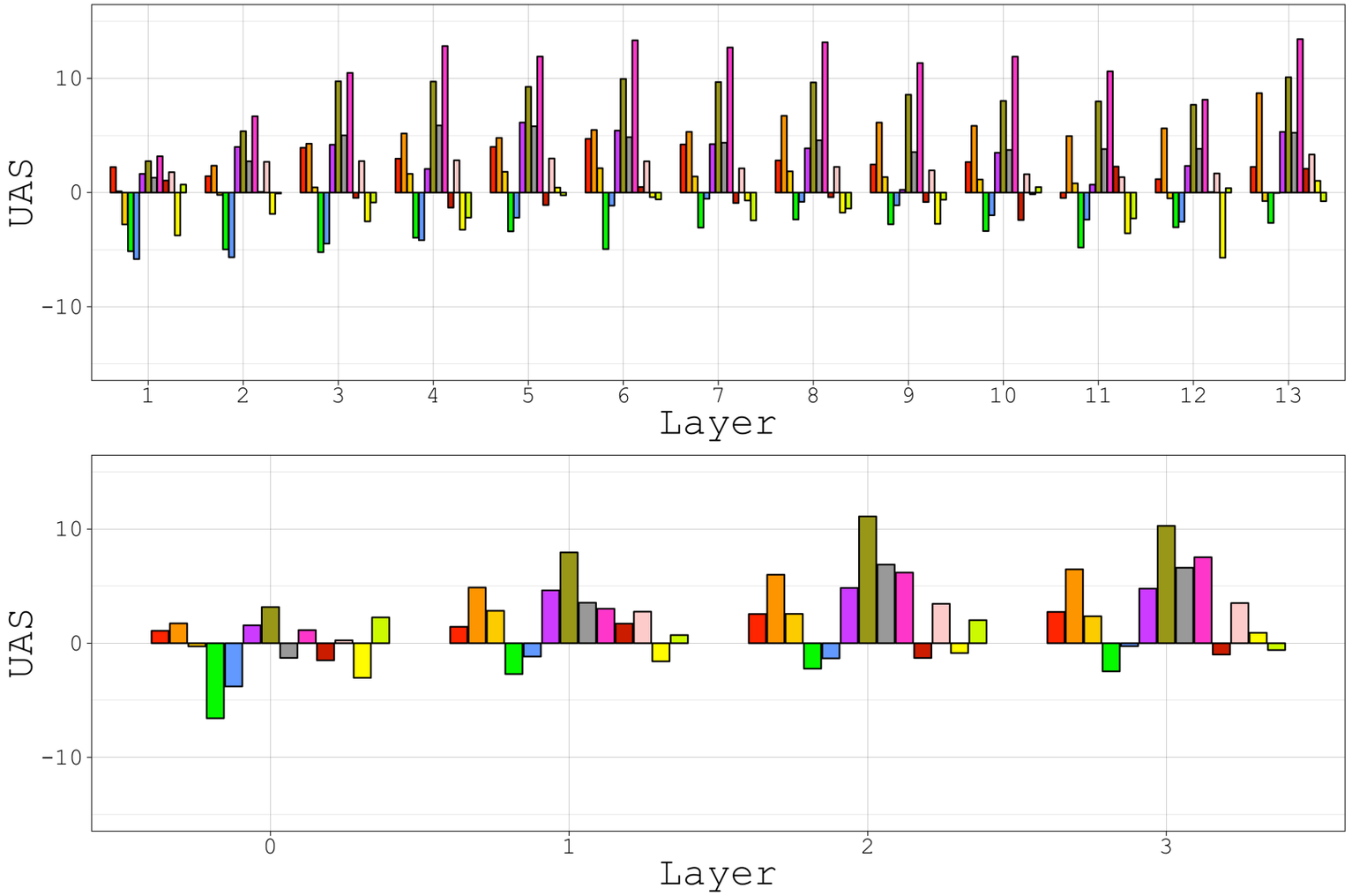
Results


UAS
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13
BERT-UD
BERT-SUD
BERT
schema
ud
sud
lang
arb
cmn
eng
eus
fin
heb
hin
ita
jpn
kor
rus
swe
tur

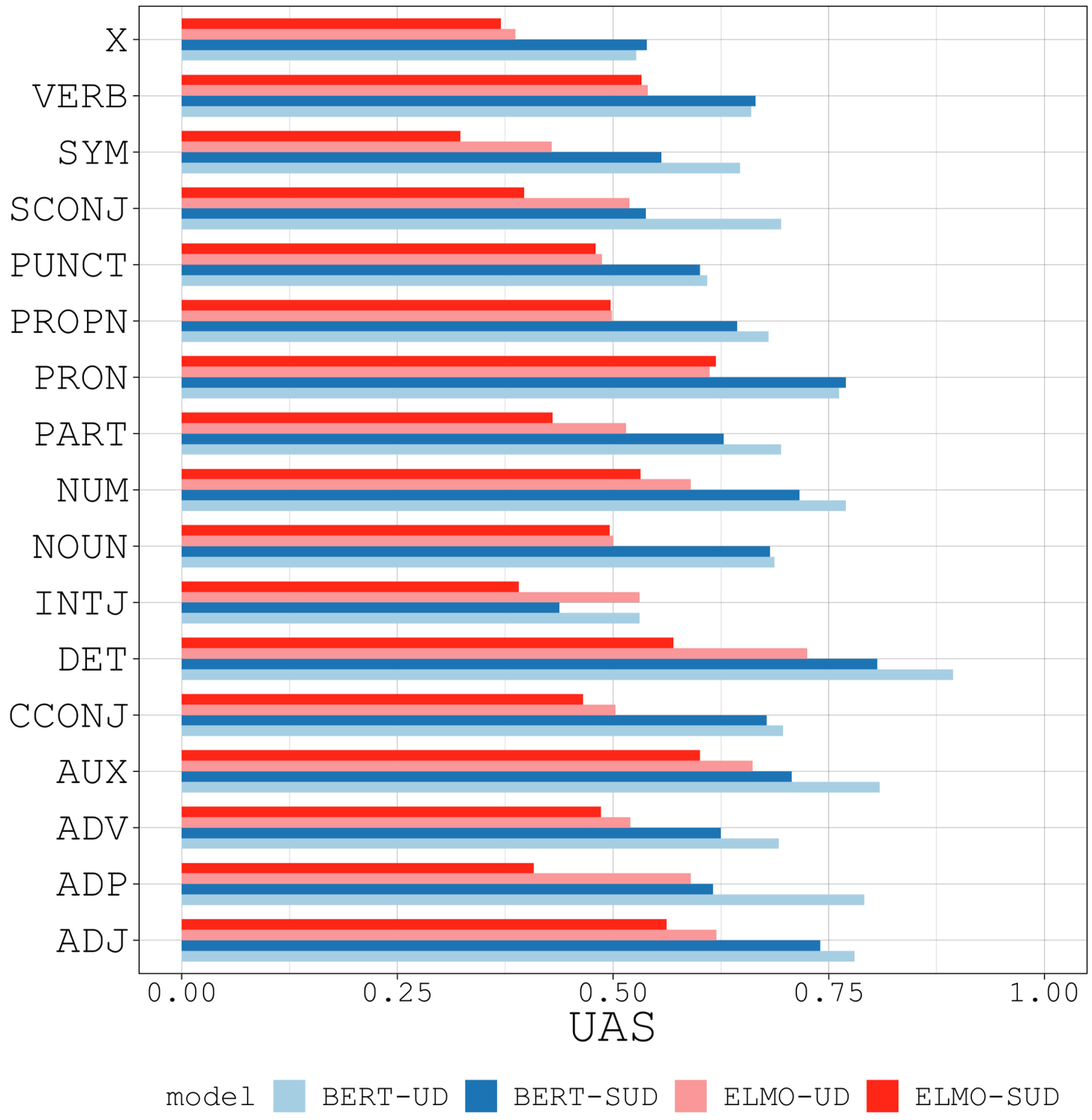
Results

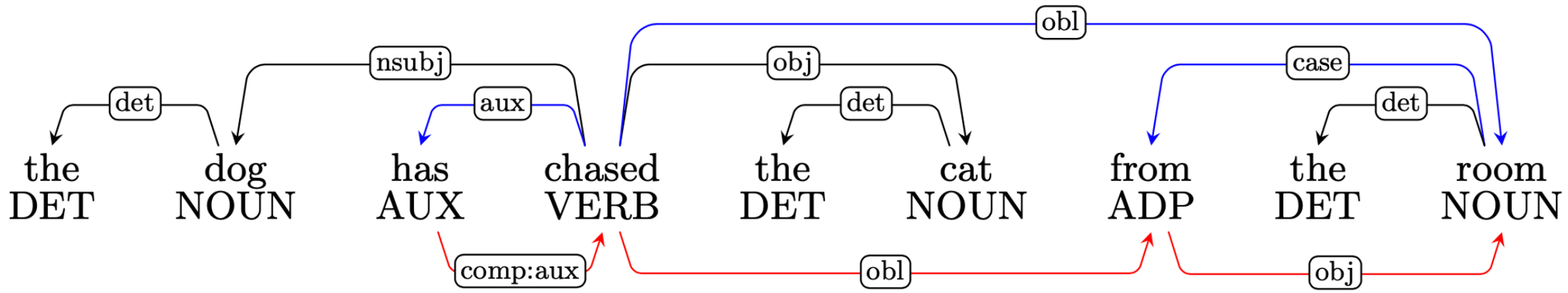

UAS
1.00
0.75
0.50
0.25
0.00
0
1
2
3
ELMo-UD
ELMo-SUD
ELMo
schema
ud
sud
lang
arb
cmn
eng
eus
fin
heb
hin
ita
jpn
kor
rus
swe
tur

Results

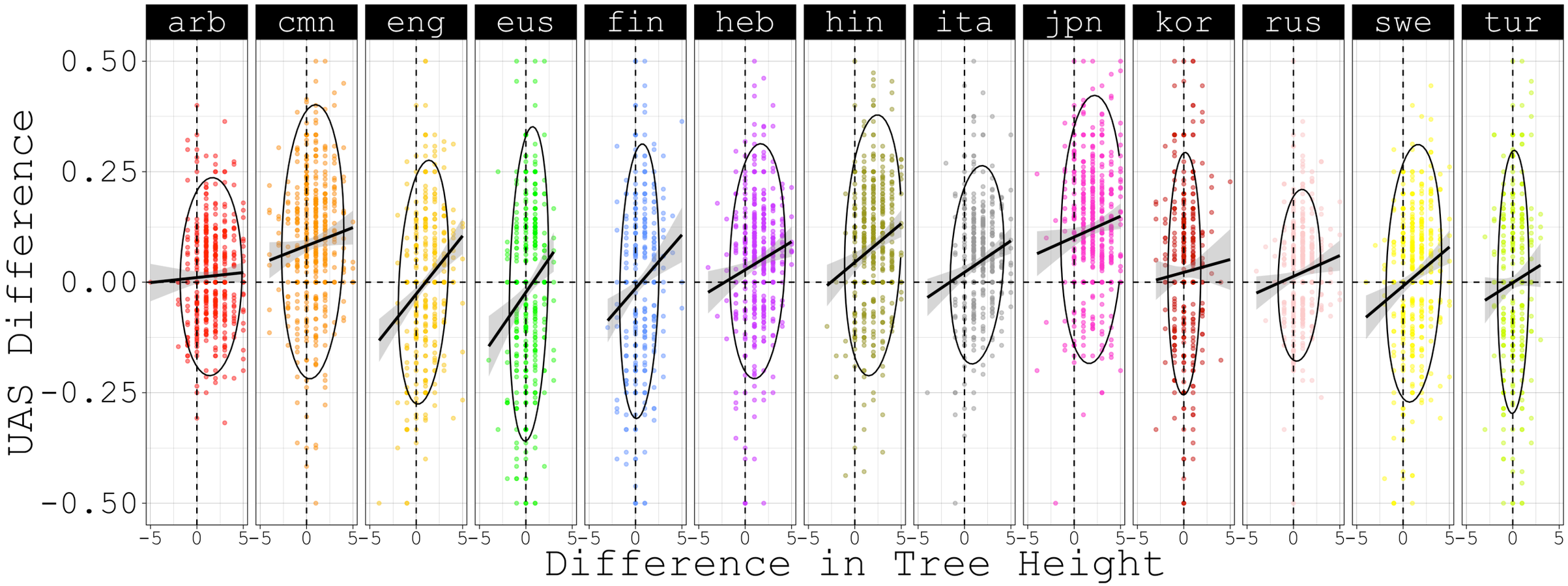

UAS
10
0
-10
1
2
3
4
5
6
7
8
9
10
11
12
13
Layer
UAS
10
0
-10
0
1
2
3
Layer
lang
arb
cmn
eng
eus
fin
heb
hin
ita
jpn
kor
rus
swe
tur

Part-of-speech Tags




obl
nsubj
obj
case
det
aux
det
det
the
DET
dog
NOUN
has
AUX
chased
VERB
the
DET
cat
NOUN
from
ADP
the
DET
room
NOUN
comp:aux
obl
obj

Difference in Tree Height



Research Questions

1. Does the syntax captured by LMs align more with UD or SUD representations?

the probe heavily favors UD, albeit with interesting variations across languages

Research Questions

2. Do different languages play a role in how well parse trees are captured by such models?

languages with many function words (jpn, ita, hin) overwhelmingly fit better to UD

agglutinative languages show no preference in aggregate

tree height is highly correlated with probe accuracy — UD's flatter trees are a better fit

Thank you!