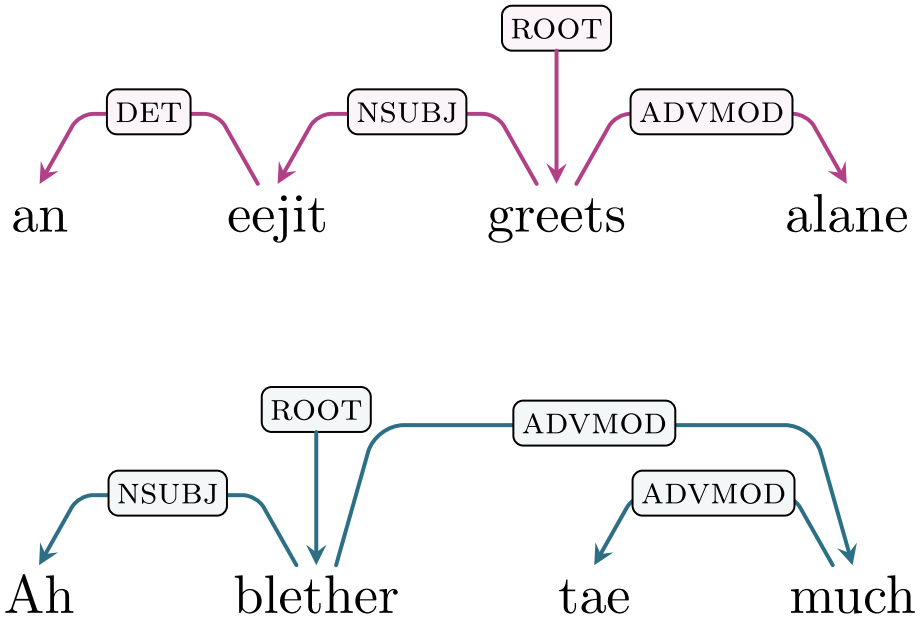
Replicating and Extending "*Because Their Treebanks Leak*": Graph Isomorphism, Covariants, and Parser Performance

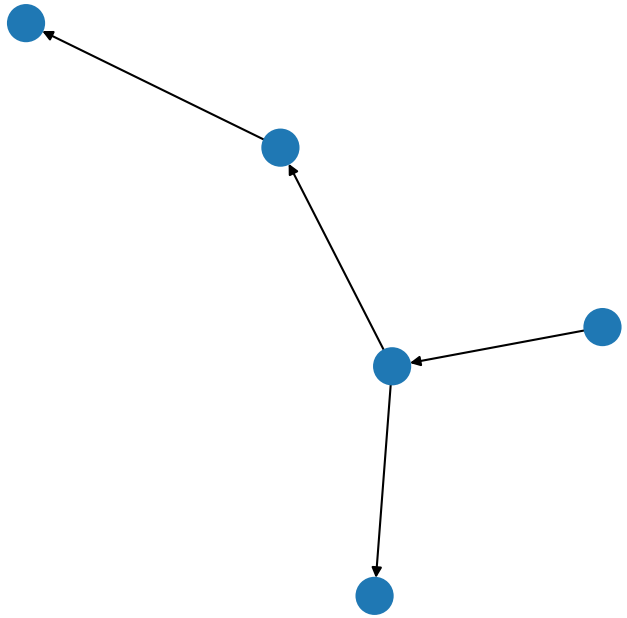
Mark Anderson, Anders Søgaard, Carlos Gómez Rodríguez

Directed Unlabelled Graph Isomorphism (DUG)

Dependency trees



Directed unlabelled graph



Data and Systems

All use Universal Dependency treebanks

- CoNLL18 - best performing systems, 33 treebanks
- UDPipe 1.2 - transition-based NN, 94 treebanks (UD v2.5)
- UDPipe 2.0 - graph-based NN, competitive, 90 treebank (UD v2.7)

Explained variance from LR fit

	Original			10 Seeds		
	CoNLL18	UDPipe 1.2	UDPipe 2.0	CoNLL18	UDPipe 1.2	UDPipe 2.0
Training size	0.014	0.100	0.060	-0.019	-0.346	-0.005
+ DUG	0.228	0.061	0.097	-0.004	-0.553	0.091
+ $\langle L_{test} \rangle$	0.195	0.169	0.146	-0.007	-0.370	0.140
All	-0.078	0.157	0.086	-0.413	-0.138	0.106

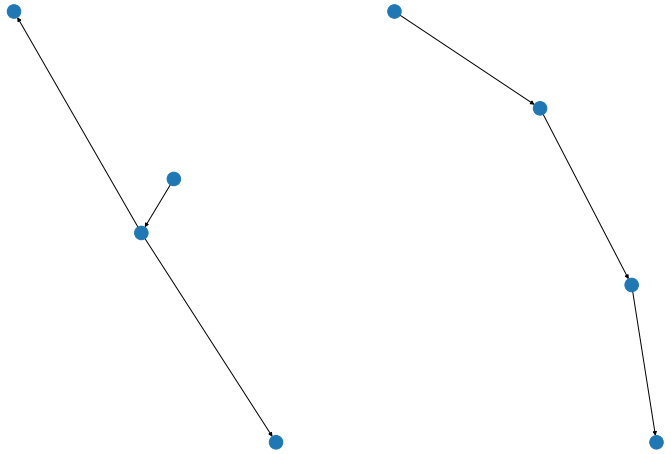
Better with log

	CoNLL18	UDPipe 1.2	UDPipe 2.0
log-size	0.055	0.319	0.126
+ DUG	0.132	0.410	0.277
+ $\langle L_{test} \rangle$	0.106	0.452	0.294
All	-0.184	0.412	0.229

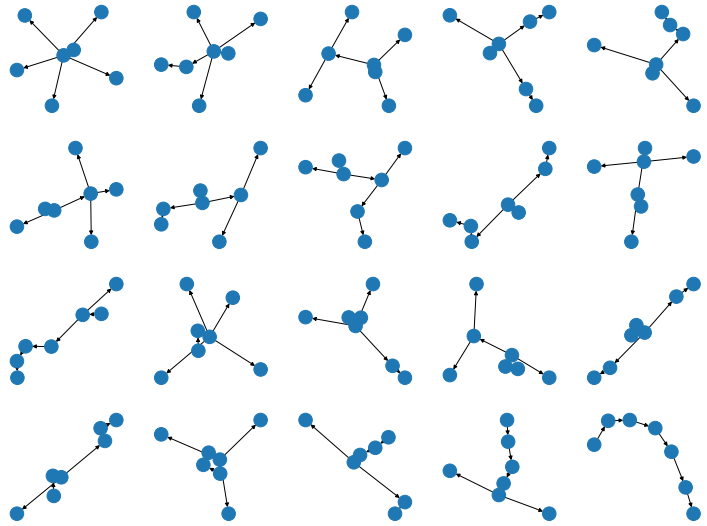
10 seeds CV (3 splits)

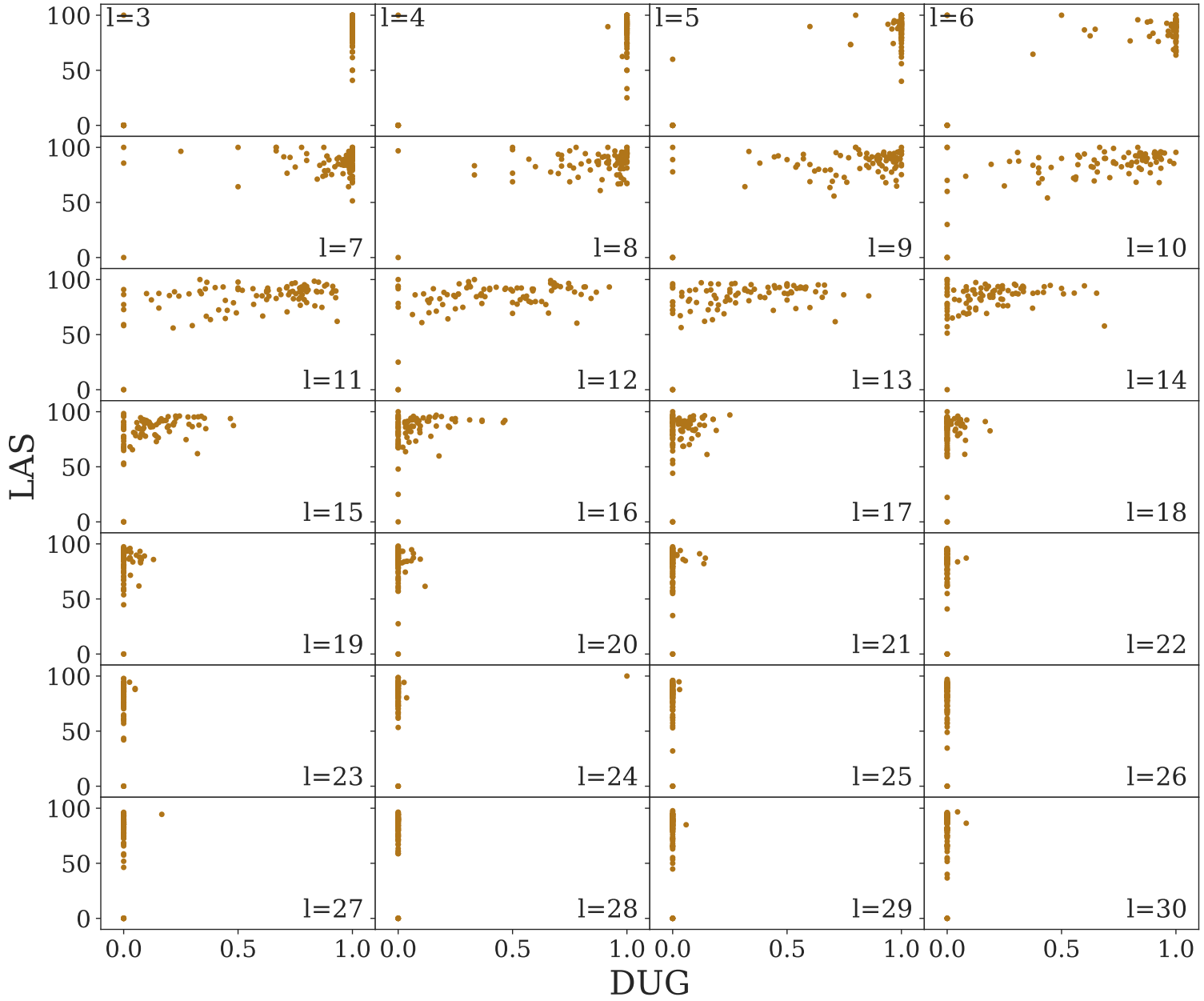
Isomorphisms

Sentences with 3 tokens



Sentences with 6 tokens




l=3
l=4
l=5
l=6
l=7
l=8
l=9
l=10
l=11
l=12
l=13
l=14
l=15
l=16
l=17
l=18
l=19
l=20
l=21
l=22
l=23
l=24
l=25
l=26
l=27
l=28
l=29
l=30
100
50
0
0.0
0.5
1.0
LAS
DUG

DUG Coefficients

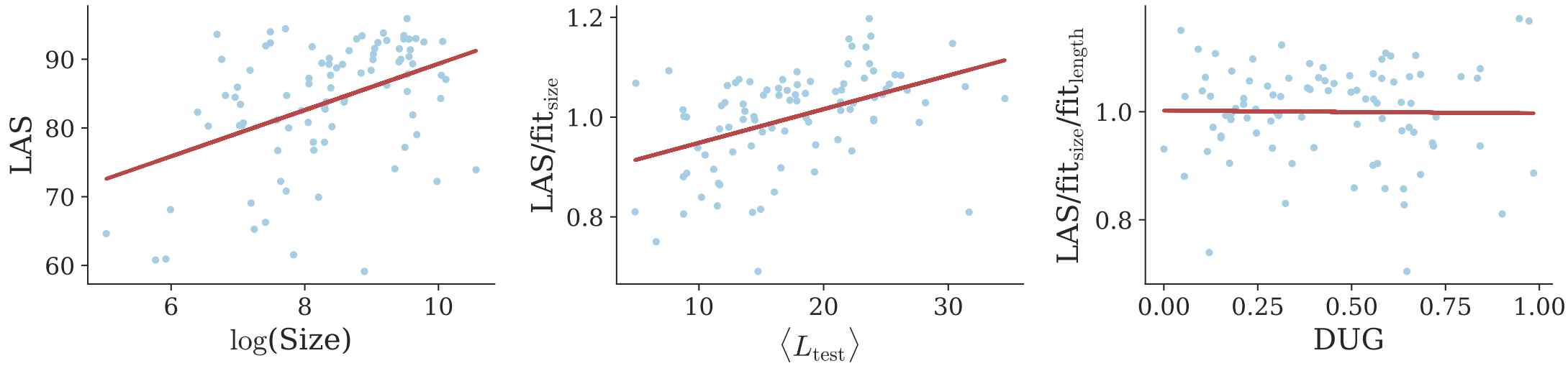
	CoNLL18	UDPipe 1.2	UDPipe 2.0
LAS	-0.13 (p=0.458)	-0.13 (p=0.213)	-0.18 (p=0.083)
Training size	0.44 (p=0.011)	0.42 (p<0.001)	0.46 (p<0.001)
$\langle L_{test} \rangle$	-0.96 (p<0.001)	-0.91 (p<0.001)	-0.92 (p<0.001)

Spearman's ρ

Partial Coefficients

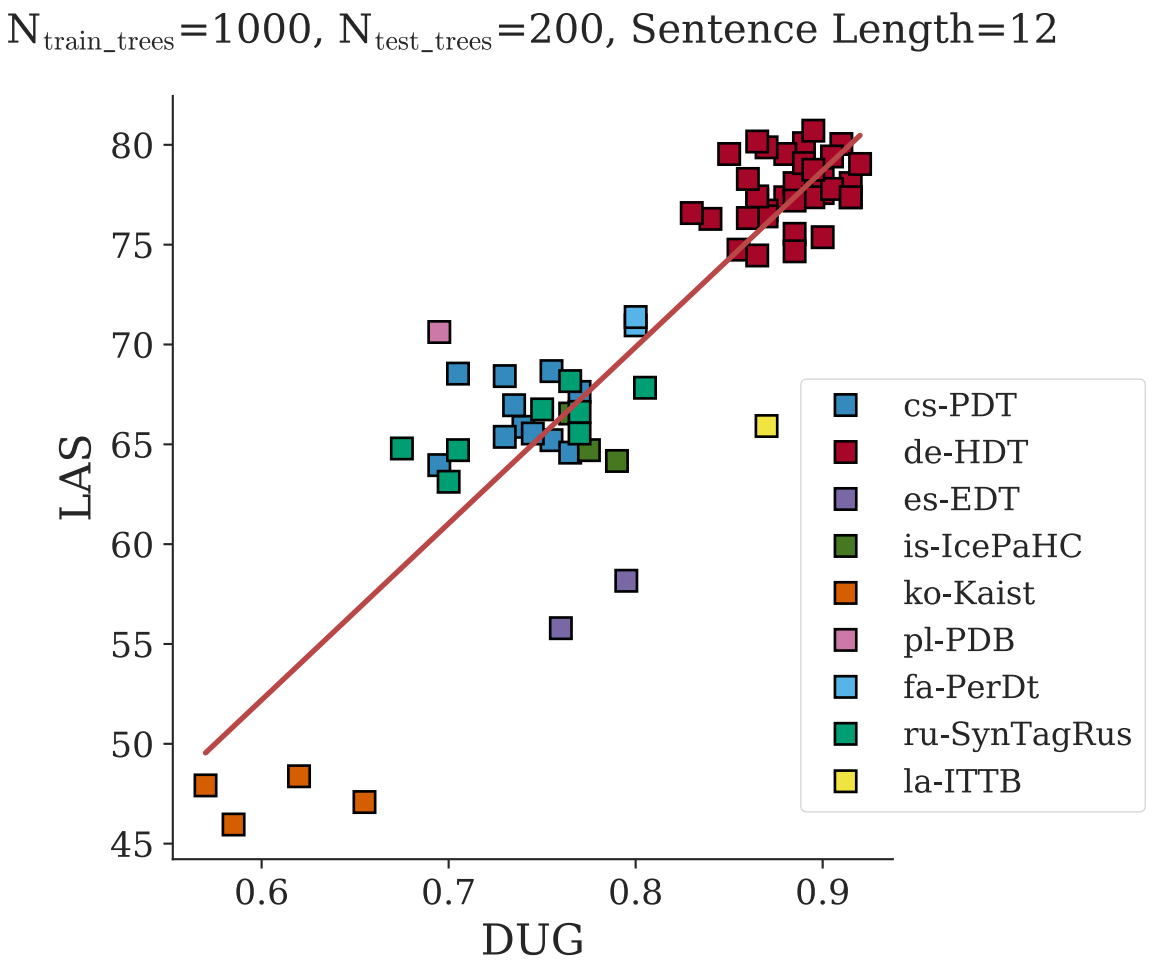
	CoNLL18	UDPipe 1.2	UDPipe 2.0
DUG	-0.13 (p=0.458)	-0.13 (p=0.213)	-0.18 (p=0.083)
Training size	-0.44 (p=0.010)	-0.50 (p<0.001)	-0.46 (p<0.001)
$\langle L_{test} \rangle$	0.18 (p=0.329)	-0.13 (p=0.213)	0.21 (p=0.049)
both	-0.27 (p=0.126)	0.01 (p=0.915)	-0.12 (p=0.245)

Background removal



Controlled experiment

- Find treebanks with sufficient number of sentences with 12 tokens (>1200).
- Create unique splits of 1000 training instances and 200 test.
- Train UDPipe 1.2 models for each.
- Measure LAS and DUG for each split.
- Spearman's ρ: 0.82 ($p<0.001$)



End