Supplementary Material: Appendices

A Details of UD Treebanks

The statistics of the Universal Dependency treebanks we used are summarized in Table 1.

Language	Lang. Family	Treebank		#Sent.	#Token(w/o punct)
Arabic (ar)	Afro-Asiatic	PADT	train	6075	223881(206041)
			dev	909	30239(27339)
			test	680	28264(26171)
Bulgarian (bg)	IE.Slavic	BTB	train	8907	124336(106813)
			dev	1115	16089(13822)
			test	1116	15724(13456)
Catalan (ca)	IE.Romance	AnCora	train	13123	417587(371981)
			dev	1709	56482(50452)
			test	1846	57902(51459)
Chinese (zh)	Sino-Tibetan	GSD	train	3997	98608(84988)
			dev	500	12663(10890)
			test	500	12012(10321)
Croatian (hr)	IE.Slavic	SET	train	6983	154055(135206)
			dev	849	19543(17211)
			test	1057	23446(20622)
Czech (cs)	IE.Slavic	PDT,CAC, CLTT,FicTree	train	102993	1806230(1542805)
			dev	11311	191679(163387)
			test	12203	205597(174771)
Danish (da)	IE.Germanic	DDT	train	4383	80378(69219)
			dev	564	10332(8951)
			test	565	10023(8573)
Dutch (nl)	IE.Germanic	Alpino, LassySmall	train	18058	261180(228902)
			dev	1394	22938(19645)
			test	1472	22622(19734)
English (en)	IE.Germanic	EWT	train	12543	204585(180303)
			dev	2002	25148(21995)
			test	2077	25096(21898)
Estonian (et)	Uralic	EDT	train	20827	287859(240496)
			dev	2633	37219(30937)
			test	2737	41273(34837)
Finnish (fi)	Uralic	TDT	train	12217	162621(138324)
			dev	1364	18290(15631)
			test	1555	21041(17908)
French (fr)	IE.Romance	GSD	train	14554	356638(316780)
			dev	1478	35768(31896)
			test	416	10020(8795)
German (de)	IE.Germanic	GSD	train	13814	263804(229338)
			dev	799	12486(10809)
			test	977	16498(14132)
Hebrew (he)	Afro-Asiatic	HTB	train	5241	137680(122122)
			dev	484	11408(10050)
			test	491	12281(10895)
Hindi (hi)	IE.Indic	HDTB	train	13304	281057(262389)
			dev	1659	35217(32850)
			test	1684	35430(33010)
Indonesian (id)	Austronesian	GSD	train	4477	97531(82617)
			dev	559	12612(10634)
			test	557	11780(10026)
Italian (it)	IE.Romance	ISDT	train	13121	276019(244632)
			dev	564	11908(10490)
			test	482	10417(9237)
Japanese (ja)	Japanese	GSD	train	7164	161900(144045)
			dev	511	11556(10326)
			test	557	12615(11258)
Korean (ko)	Korean	GSD, Kaist	train	27410	353133(312481)
			dev	3016	37236(32770)
			test	3276	40043(35286)
Latin (la)	IE.Latin	PROIEL	train	15906	171928(171928)
			dev	1234	13939(13939)
			test	1260	14091(14091)
Latvian (lv)	IE.Baltic	LVTB	train	5424	80666(66270)
			dev	1051	14585(11487)

			test	1228	15073(11846)
Norwegian (no)	IE.Germanic	Bokmaal, Nynorsk	train	29870	489217(432597)
			dev	4300	67619(59784)
			test	3450	54739(48588)
Polish (pl)	IE.Slavic	LFG, SZ	train	19874	167251(136504)
			dev	2772	23367(19144)
			test	2827	23920(19590)
Portuguese (pt)	IE.Romance	Bosque, GSD	train	17993	462494(400343)
			dev	1770	42980(37244)
			test	1681	41697(36100)
Romanian (ro)	IE.Romance	RRT	train	8043	185113(161429)
			dev	752	17074(14851)
			test	729	16324(14241)
Russian (ru)	IE.Slavic	SynTagRus	train	48814	870474(711647)
			dev	6584	118487(95740)
			test	6491	117329(95799)
Slovak (sk)	IE.Slavic	SNK	train	8483	80575(65042)
			dev	1060	12440(10641)
			test	1061	13028(11208)
Slovenian (sl)	IE.Slavic	SSJ, SST	train	8556	132003(116730)
			dev	734	14063(12271)
			test	1898	24092(22017)
Spanish (es)	IE.Romance	GSD, AnCora	train	28492	827053(730062)
			dev	3054	89487(78951)
			test	2147	64617(56973)
Swedish (sv)	IE.Germanic	Talbanken	train	4303	66645(59268)
			dev	504	9797(8825)
			test	1219	20377(18272)
Ukrainian (uk)	IE.Slavic	IU	train	4513	75098(60976)
			dev	577	10371(8381)
			test	783	14939(12246)

Table 1: Statistics of the UD Treebanks we used. For language family, "IE" is the abbreviation for Indo-European. "(w/o) punct" means the numbers of the tokens excluding "PUNCT" and "SYM".

B Hyper-Parameters

Table 2 summarizes the hyper-parameters that we used in our experiments. Most of them are similar to those in (Dozat and Manning, 2017) and (Ma et al., 2018).

	Layer	Hyper-Parameter	Value
Input	Word	dimension	300
	POS	dimension	50
RNN	Encoder	encoder layer	3
		encoder size	300
	MLP	arc MLP size	512
		label MLP size	128
	Training	Dropout	0.33
		optimizer	Adam
		learning rate	0.001
		batch size	32
Self-Attention	Encoder	encoder layer	6
		d_{model}	350
		d_{ff}	512
	MLP	arc MLP size	512
		label MLP size	128
	Training	Dropout	0.2
		optimizer	Adam
		learning rate	0.0001
		batch size	80

Table 2: Hyper-parameters in our experiments.

C Details about augmented dependency types

Type	Avg. Freq. (%)	#Lang.	Type	Avg. Freq. (%)	#Lang.
(ADP, NOUN, case)	7.47	31	(PROPN, VERB, nsubj)	0.81	30
(PUNCT, VERB, punct)	6.91	30	(PRON, VERB, obj)	0.77	30
(NOUN, NOUN, nmod)	4.97	31	(NOUN, ROOT, root)	0.66	31
(ADJ, NOUN, amod)	4.92	31	(VERB, VERB, xcomp)	0.61	28
(DET, NOUN, det)	4.69	30	(VERB, VERB, ccomp)	0.60	30
(VERB, ROOT, root)	4.31	31	(ADP, PRON, case)	0.57	29
(NOUN, VERB, obl)	3.96	30	(AUX, NOUN, cop)	0.57	28
(NOUN, VERB, obj)	3.10	31	(ADV, ADJ, advmod)	0.54	29
(NOUN, VERB, nsubj)	2.89	31	(AUX, ADJ, cop)	0.50	27
(PUNCT, NOUN, punct)	2.75	30	(PROPN, VERB, obl)	0.48	29
(ADV, VERB, advmod)	2.43	31	(PRON, VERB, obl)	0.44	30
(AUX, VERB, aux)	2.29	28	(ADV, NOUN, advmod)	0.41	28
(PRON, VERB, nsubj)	1.53	30	(ADJ, ROOT, root)	0.39	29
(ADP, PROPN, case)	1.46	29	(PRON, NOUN, nmod)	0.39	22
(NOUN, NOUN, conj)	1.32	30	(NOUN, ADJ, obl)	0.37	25
(VERB, NOUN, acl)	1.31	31	(PROPN, PROPN, conj)	0.35	29
(SCONJ, VERB, mark)	1.27	28	(NOUN, ADJ, nsubj)	0.35	30
(CCONJ, VERB, cc)	1.18	30	(CCONJ, ADJ, cc)	0.29	28
(PROPN, NOUN, nmod)	1.14	30	(PUNCT, NUM, punct)	0.26	24
(CCONJ, NOUN, cc)	1.13	30	(NOUN, NOUN, nsubj)	0.25	31
(NUM, NOUN, nummod)	1.11	31	(ADJ, ADJ, conj)	0.25	26
(PROPN, PROPN, flat)	1.09	26	(CCONJ, PROPN, cc)	0.22	26
(VERB, VERB, conj)	1.05	30	(PRON, VERB, iobj)	0.21	21
(PUNCT, PROPN, punct)	0.94	29	(ADV, ADV, advmod)	0.19	21
(VERB, VERB, advcl)	0.89	30	(NOUN, NOUN, appos)	0.18	23
(PUNCT, ADJ, punct)	0.89	30	(PROPN, VERB, obj)	0.17	24

Table 3: Selected augmented dependency types sorted by their average frequencies. "#Lang." denotes in how many languages the specific type appears. Since the augmented dependency types can be in hundreds or larger than 1k, but mostly infrequent, we prune them according to average frequency and number of appearing languages. Our pruning criterion is "$Freq > 0.1\%$ and $\#Lang \geq 20$".

D Punctuation-included Evaluation on the test sets

Language	SelfAtt-Graph	RNN-Graph	SelfAtt-Stack	RNN-Stack
en	89.29/87.52	89.46/87.54	89.16/87.26	**90.83/89.07**
no	78.47/71.38	78.47/71.50	78.11/70.84	**79.61/72.10**
sv	79.70/72.69	79.94/72.99	79.24/72.24	**81.44/73.98**
fr	75.58/71.05	**76.11/71.79**	74.32/69.87	73.56/69.16
pt	**73.07**/65.30	72.82/**65.38**	71.61/63.96	71.21/63.76
da	74.03/66.52	74.99/67.67	73.76/66.15	**75.81/67.76**
es	70.98/63.84	**71.50/64.40**	69.54/62.44	69.73/62.37
it	78.19/73.77	**78.63/74.31**	76.52/72.11	78.29/73.84
hr	**60.58/52.60**	58.60/50.28	59.03/50.65	59.27/50.72
ca	70.47/62.37	**70.96/62.85**	68.91/60.87	68.79/60.45
pl	**74.78/64.68**	71.73/60.83	73.82/63.19	72.24/62.11
uk	**57.57/51.16**	56.32/50.25	54.58/48.18	57.31/50.81
sl	**66.50/55.84**	64.55/53.84	64.83/53.88	66.07/55.03
nl	66.92/59.59	66.45/59.54	66.05/58.59	**68.10/61.01**
bg	**76.15/66.48**	74.85/65.01	74.92/65.23	75.69/65.96
ru	55.85/48.47	55.40/47.84	54.10/46.62	**55.88/48.52**
de	**69.61/61.27**	67.60/58.86	68.18/59.73	68.02/59.36
he	**53.53/46.98**	53.04/46.16	51.53/44.76	53.26/40.83
cs	**60.95/53.03**	59.56/51.80	58.88/50.86	59.63/51.13
ro	**63.11/53.54**	61.19/51.45	60.31/50.63	59.38/49.61
sk	**65.11/57.76**	63.66/56.38	63.68/56.21	64.97/57.08
id	**49.00/44.07**	47.08/42.78	47.03/42.17	47.12/42.38
lv	66.53/49.52	**66.95/49.66**	64.50/47.72	65.98/48.46
fi	64.83/49.83	**65.04/49.98**	63.41/48.61	64.97/49.63
et	**63.50/45.88**	63.08/45.45	61.74/44.12	62.15/44.57
zh*	**40.46/25.52**	39.54/24.74	38.37/23.55	39.26/24.25
ar	**37.15/27.79**	32.37/25.42	31.69/23.46	32.04/24.73
la	**47.96/35.21**	45.96/33.91	45.49/33.19	43.85/31.25
ko	**33.96/17.99**	33.08/16.96	31.68/16.04	32.81/16.17
hi	**36.90/28.52**	30.94/23.55	32.65/24.92	26.80/19.49
ja*	**27.83/21.25**	18.39/12.59	20.33/13.56	15.01/9.75
Average	**62.21/53.27**	60.91/52.12	60.26/51.34	60.62/51.46

Table 4: Evaluations with punctuation included (average UAS%/LAS% over 5 runs) on the test sets. The patterns are similar to the punctuation-excluded evaluations in the main content. (Languages are sorted by the word-ordering distance to English, '*' refers to results of delexicalized models.)

E Results on the original training sets

Language	SelfAtt-Graph	RNN-Graph	SelfAtt-Stack	RNN-Stack
en°	90.35/88.40	90.44/88.31	90.18/88.06	**91.82/89.89**
no	80.72/72.45	80.59/72.41	80.06/71.60	**81.46/72.75**
sv	80.07/71.91	80.42/**72.39**	79.45/71.28	**80.87**/72.25
fr	79.31/74.73	**79.99/75.52**	78.62/74.02	76.84/72.22
pt	77.06/69.33	**77.33/69.91**	75.84/68.22	75.39/67.75
da	75.75/67.12	75.95/67.41	75.18/66.55	**76.98/67.50**
es	73.91/66.48	**74.39/67.03**	72.84/65.38	72.46/64.78
it	80.37/75.48	**80.89/75.99**	79.15/74.17	79.05/73.91
hr	**61.57/52.40**	59.74/50.37	59.94/50.43	60.44/50.68
ca	74.40/65.73	**74.94/66.21**	73.01/64.42	72.75/63.68
pl	**75.32/63.26**	73.12/59.76	74.28/61.46	73.21/61.02
uk	65.70/**57.48**	64.77/56.40	64.10/55.83	**65.82**/57.13
sl	**69.13/58.92**	67.35/56.87	67.74/57.08	68.95/58.26
nl	68.98/60.00	68.37/59.52	68.22/59.02	**69.16/60.11**
bg	**80.25/68.88**	78.39/67.03	79.19/67.66	79.66/68.22
ru	60.50/51.35	59.55/50.17	59.01/49.71	**60.71/51.57**
de	**67.23/58.27**	66.64/57.48	66.10/56.89	65.88/56.63
he	58.32/**49.80**	57.75/49.07	56.36/47.62	**58.79**/43.83
cs	**63.04/53.92**	61.75/52.91	61.11/51.91	62.21/52.48
ro	**65.31/54.22**	63.17/52.16	63.03/51.95	61.78/50.52
sk	**76.07/62.75**	74.67/61.15	75.93/61.97	75.37/60.94
id	**47.92/41.93**	45.07/39.91	46.23/40.16	45.62/39.67
lv	71.69/50.43	**72.48/50.85**	70.24/48.97	71.60/49.56
fi	64.64/46.21	64.63/**46.22**	63.07/44.82	**64.74**/46.09
et	**66.63/45.58**	65.78/45.01	64.94/44.04	65.06/44.33
zh*	**41.05/23.85**	40.11/23.02	39.49/22.68	39.89/22.49
ar	**38.74/28.24**	33.66/25.44	34.25/24.69	33.31/24.86
la	**49.04/35.48**	47.12/34.36	46.78/33.56	45.26/31.97
ko	**34.62/15.14**	33.91/14.16	32.70/13.77	32.95/13.14
hi	**36.01/27.24**	29.59/21.75	32.02/23.79	26.37/18.56
ja*	**28.19/21.74**	18.23/12.68	20.53/13.78	15.21/10.37
Average	**64.57**/54.14	63.25/52.94	62.88/52.44	62.88/52.16

Table 5: Results (average UAS%/LAS% over 5 runs, excluding punctuation) on the original training sets. (Languages are sorted by the word-ordering distance to English, '*' refers to results of delexicalized models, 'en°' means that for English we use results on the test set since models are trained with the English training set.)

F Results on Google Universal Dependency Treebanks v2.0

We also ran our models on Google Universal Dependency Treebanks v2.0 (McDonald et al., 2013), which is an older dataset that was used by (Guo et al., 2015). The results show that our models perform better consistently.

Language	SelfAtt-Graph	RNN-Graph	SelfAtt-Stack	RNN-Stack	(Guo et al., 2015)
German	65.03/55.03	64.60/54.57	63.63/54.40	**65.51/55.82**	60.35/51.54
French	74.45/63.28	**76.75/65.20**	73.63/62.76	75.13/64.44	72.93/63.12
Spanish	72.00/61.50	73.99/63.46	71.73/61.42	**74.13/64.00**	71.90/62.28

Table 6: Comparisons (UAS%/LAS%) on Google Universal Dependency Treebanks v2.0.

G Results on specific dependency types for Czech

In table 7, we show results of Czech on some dependency types with evaluation breakdowns on dependency directions. We select Czech mainly for two reasons: (1) It has the largest dataset; (2) Czech is famous for relatively flexible word order. Generally, we can see that models that are more flexible on word ordering perform better. Interestingly, for objective and subjective types, we can see that LAS scores for all models are quite low even when the correct heads are predicted. The reason might be that even the relative-positional self-attention encoder can capture some positional information which further reveals word ordering information in some way.

(ADP, NOUN, case): (mod-first% in English is 99.92%.)					
Direction	Percentage	SelfAtt-Graph	RNN-Graph	SelfAtt-Stack	RNN-Stack
mod-first	99.99%	**75.34/75.34**	74.62/74.61	74.46/74.43	74.17/74.08
head-first	0.01%	–	–	–	–
all	100.00%	**75.33/75.33**	74.61/74.61	74.45/74.43	74.17/74.07
(NOUN, NOUN, nmod): (mod-first% in English is 4.72%.)					
Direction	Percentage	SelfAtt-Graph	RNN-Graph	SelfAtt-Stack	RNN-Stack
mod-first	0.97%	–	–	–	–
head-first	99.03%	21.38/17.85	18.55/16.20	20.49/16.61	**22.51/19.16**
all	100.00%	21.64/17.68	18.86/16.05	20.77/16.45	**22.78/18.98**
(ADJ, NOUN, amod): (mod-first% in English is 99.01%.)					
Direction	Percentage	SelfAtt-Graph	RNN-Graph	SelfAtt-Stack	RNN-Stack
mod-first	92.99%	88.93/88.92	**89.42/89.41**	85.39/85.21	87.26/86.37
head-first	7.01%	**41.80/37.03**	36.52/32.36	34.82/27.19	40.59/19.85
all	100.00%	85.63/85.29	**85.72/85.41**	81.85/81.14	83.98/81.71
(NOUN, VERB, obl): (mod-first% in English is 9.62%.)					
Direction	Percentage	SelfAtt-Graph	RNN-Graph	SelfAtt-Stack	RNN-Stack
mod-first	37.80%	48.84/40.33	46.39/38.49	48.75/41.08	**50.16/41.64**
head-first	62.20%	**62.81/55.97**	60.38/53.41	62.22/55.37	61.73/55.32
all	100.00%	**57.53**/50.06	55.09/47.77	57.13/49.97	57.36/**50.15**
(NOUN, VERB, obj): (mod-first% in English is 0.72%.)					
Direction	Percentage	SelfAtt-Graph	RNN-Graph	SelfAtt-Stack	RNN-Stack
mod-first	20.65%	55.56/**0.64**	53.75/0.46	54.08/0.37	**60.34**/0.18
head-first	79.35%	**73.18/65.24**	71.30/62.28	72.12/63.81	72.76/64.65
all	100.00%	69.54/**51.90**	67.68/49.52	68.39/50.71	**70.20**/51.34
(NOUN, VERB, nsubj): (mod-first% in English is 85.07%.)					
Direction	Percentage	SelfAtt-Graph	RNN-Graph	SelfAtt-Stack	RNN-Stack
mod-first	60.22%	**61.42**/58.33	58.12/54.51	60.88/58.24	60.67/**58.98**
head-first	39.78%	**64.07**/3.83	62.93/3.18	62.38/2.97	59.94/**4.42**
all	100.00%	**62.47**/36.65	60.03/34.09	61.48/36.25	60.38/**37.28**
(ADV, VERB, advmod): (mod-first% in English is 58.82%.)					
Direction	Percentage	SelfAtt-Graph	RNN-Graph	SelfAtt-Stack	RNN-Stack
mod-first	70.15%	**88.23/87.49**	86.43/85.48	86.65/85.30	86.64/83.72
head-first	29.85%	**65.79/65.28**	65.02/64.33	65.33/64.35	61.93/60.53
all	100.00%	**81.53/80.86**	80.04/79.17	80.29/79.05	79.26/76.80
(AUX, VERB, aux): (mod-first% in English is 99.64%.)					
Direction	Percentage	SelfAtt-Graph	RNN-Graph	SelfAtt-Stack	RNN-Stack
mod-first	83.71%	88.78/**88.19**	84.44/83.52	**89.03**/86.59	82.54/76.33
head-first	16.29%	**68.18/65.28**	54.59/50.87	63.96/54.02	56.67/20.24
all	100.00%	**85.42/84.46**	79.57/78.20	84.94/81.28	78.32/67.19
(VERB, VERB, advcl): (mod-first% in English is 31.02%.)					
Direction	Percentage	SelfAtt-Graph	RNN-Graph	SelfAtt-Stack	RNN-Stack
mod-first	41.75%	57.51/**55.61**	56.98/55.60	**57.54**/55.03	54.74/51.66
head-first	58.25%	**71.52/56.68**	67.39/56.08	67.27/54.17	65.93/54.13
all	100.00%	**65.67/56.23**	63.04/55.88	63.21/54.53	61.26/53.10

Table 7: Evaluation breakdowns (UAS%/LAS%) on dependency directions for Czech on some specific dependency types. "mod-first" means the dependency edges whose modifier is before head, "head-first" means the opposite, and "all" indicates both "mod-first" and "head-first". "–" replaces results that are unstable because of rare appearance (below 1%).

References

Timothy Dozat and Christopher D Manning. 2017. Deep biaffine attention for neural dependency parsing. *Internation Conference on Learning Representations* .

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. volume 1, pages 1234–1244.

Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard Hovy. 2018. Stack-pointer networks for dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of ACL-2013*. Sofia, Bulgaria, pages 92–97.