Appendix of "Working Hard or Hardly Working: Challenges of Integrating Typology into Neural Dependency Parsers"

A Dependency Relations for Deriving the Liu Directionalities

Among all the 37 relation types defined in Universal Dependencies, we select the top-20 relations sorted by the number of languages in which the specific type appears, as listed in Table 1. For relation types that are missing in a specific language, we simply put its value (directionality) as 0.5.

cc, conj, case, nsubj, nmod, dobj, mark,
advcl, amod, advmod, neg, nummod, xcomp,
ccomp, cop, acl, aux, punct, det, appos,
iobj, dep, csubj, parataxis, mwe, name,
nsubjpass, compound, auxpass, csubjpass,
vocative, discourse

Table 1: Subset of universal dependency relations used for deriving the Liu directionalities.

B Feature Templates for Selective Sharing

We use the same set of selective sharing feature templates (Table 2) as Täckström et al. (2013).

$\texttt{d} \otimes \texttt{w.81A} \otimes \mathbb{1}\left[\texttt{h.p=VERB} \wedge \texttt{m.p=NOUN}\right]$
$\texttt{d} \otimes \texttt{w.81A} \otimes \mathbb{1}\left[\texttt{h.p=VERB} \wedge \texttt{m.p=PRON}\right]$
$\texttt{d} \otimes \texttt{w.85A} \otimes \mathbb{1}\left[\texttt{h.p=NOUN} \wedge \texttt{m.p=ADP}\right]$
$\texttt{d} \otimes \texttt{w.86A} \otimes \mathbb{1}\left[\texttt{h.p=PRON} \wedge \texttt{m.p=ADP}\right]$
$\texttt{d} \otimes \texttt{w.87A} \otimes \mathbb{1}\left[\texttt{h.p=NOUN} \wedge \texttt{m.p=ADJ}\right]$

Table 2: Arc-factored feature templates for selective sharing. Arc direction: $\texttt{d} \in \{\textsc{Left}, \textsc{Right}\}$; Part-of-speech tag of head / modifier: `h.p` / `m.p`. WALS features: `w.X` for `X=81A` (order of `Subject, Verb` and `Object`), `85A` (order of `Adposition` and `Noun`), `86A` (order of `Genitive` and `Noun`), `87A` (order of `Adjective` and `Noun`).

C Training details

To train our baseline parser and its typology-augmented variants, we use ADAM (Kingma and Ba, 2015) with a learning rate of 1e-3 for 200K updates (2M when using GD). We use a batch size of 500 tokens. Early stopping is also employed, based on the validation set in the training languages. Following Dozat and Manning (2017), we use a 3-layered bidirectional LSTM with a hidden size of 400. The hidden sizes of the MLPs for predicting arcs and dependency relations are 500 and 100, respectively.

Our baseline model shares all parameters across languages. During training, we truncate each training treebank to a maximum of 500K tokens for efficiency. Batch updates are composed of examples derived from a single language, and are sampled uniformly, such that the number of per-language updates are proportional to the size of each language's treebank. following (Wang and Eisner, 2018), when training on GD, we sample a batch from a real language with probability 0.2, and a batch of GD data otherwise.

For *fine-tune*, we perform 100 SGD updates with no early-stopping. When using K-Means to obtain language clusters, we set $K = 5$, based on cross-validation.

D LAS Results

Table 3 summarizes the LAS scores corresponding to Table 1 in the paper.

Language	B*	+$\mathbf{T}^*_S$	Our Baseline	Selective Sharing	+$\mathbf{T}_L$	+$\mathbf{T}_D$	+$\mathbf{T}_S$	Fine-tune
Basque	27.07	31.46	34.64	34.79	36.49	36.83	34.90	43.04
Croatian	48.68	52.29	61.34	61.41†	59.86	63.72	61.60	65.07
Greek	50.10	56.73	56.51	56.53†	55.16	60.18	56.59†	64.66
Hebrew	49.71	53.29	41.15	41.05	43.58	43.63	41.50	43.14
Hungarian	42.85	47.73	32.65	33.43	34.14	32.01	33.07	44.26
Indonesian	39.46	47.63	47.17	48.21	51.82	50.78	49.22	62.23
Irish	39.06	40.75	39.63	39.60†	43.02	42.14	40.24	48.58
Japanese	37.57	40.6	43.32	43.69	47.67	48.10	42.85	60.59
Slavonic	40.03	43.95	57.35	57.40†	56.89	56.69	57.19	53.88
Persian	30.06	24.6	35.72	35.59	32.85	39.78	34.93	49.72
Polish	50.08	54.85	61.67	61.57	64.69	57.20	61.71	65.68
Romanian	50.90	53.42	55.77	56.21	55.99†	59.28	56.48	59.12
Slovenian	57.09	61.48	70.86	70.01	70.44	70.03	70.29	73.81
Swedish	55.35	58.42	67.24	67.40	66.92	68.03	67.04	68.65
Tamil	28.39	37.81	33.81	34.57	34.96	36.61	34.70	47.46
AVG	43.09	47.00	49.26	49.43	50.30	51.00	49.49	56.66

Table 3: LAS results corresponding to Table 1 in the paper. Results with differences that are statistically *insignificant* compared to the baseline are marked with † (arc-level paired permutation test with $p \geq 0.05$).

E Rules for Deriving Corpus-specific WALS Features

Table 4 summarizes the rules we used to derive corpus-specific WALS features. The values are determined by the dominance of directionalities, e.g., if $\frac{\#\{\curvearrowright\}}{\#\{\curvearrowright\}+\#\{\curvearrowleft\}} > \delta$, then its typological feature is set to the right-direction value, vice versa. In-between values are set to `Mixed`. In our experiments, $\delta = 0.75$.

WALS ID	Condition	Values
82A	relation $\in$ {nsubj, csubj} $\wedge$ h.p=VERB $\wedge$ (m.p=NOUN $\vee$ m.p=PRON)	VS($\curvearrowright$), SV($\curvearrowleft$), Mixed
83A	relation $\in$ {dobj, iobj} $\wedge$ h.p=VERB $\wedge$ (m.p=NOUN $\vee$ m.p=PRON)	VO($\curvearrowright$), OV($\curvearrowleft$), Mixed
85A	(h.p=NOUN $\vee$ h.p=PRON) $\wedge$ m.p=ADP	Prepositions($\curvearrowleft$), Postpositions($\curvearrowright$)
86A	h.p=NOUN $\wedge$ m.p=NOUN	Noun-Genitive($\curvearrowright$), Genitive-Noun($\curvearrowleft$), Mixed
87A	h.p=NOUN $\wedge$ m.p=ADJ	Adjective-Noun($\curvearrowleft$), Noun-Adjective($\curvearrowright$), Mixed
88A	relation $\in$ {det} $\wedge$ m.p=DET	Demonstrative-Noun($\curvearrowleft$), Noun-Demonstrative($\curvearrowright$), Mixed

Table 4: Rules for determining the dependency arc set of each specific WALS feature type. The arc direction specificed in the parenthesis of each value indicates the global directional tendency of the corresponding typological feature.

F Examples of Mismatching between WALS and Corpus Statistics

Table 5 shows some examples of mismatching between WALS and corpus statistics. Substantial variations exist for some typological features, and for UD v1.2 in several cases, the dominant word order specified by linguists is questionable or even reversed (cf. Arabic subject-verb order).

Language	WALS		UD	
	ID	Value	#{↷}	#{↶}
Arabic	82A	SV (↶)	4,875	2,489
Czech	82A	SV (↶)	13,925	32,510
Czech	83A	VO (↷)	37,034	20,246
Spanish	83A	VO (↷)	10,745	6,119
Finnish	86A	G-N (↶)	6,010	8,134

Table 5: Example of mismatching between WALS and arc directionalities collected from UD v1.2. G-N is short for Genitive-Noun.

References

Timothy Dozat and Christopher D. Manning. 2017. Deep Biaffine Attention for Neural Dependency Parsing. In *ICLR*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *NAACL*, pages 1061–1071, Atlanta, Georgia. Association for Computational Linguistics.

Dingquan Wang and Jason Eisner. 2018. Surface statistics of an unknown language indicate how to parse it. *TACL*, 6:667–685.