A Appendix

A.1 Hyper-Parameters for Dependency Parsing

We use the same Hyper-Parameters as StackPtr over all languages for dependency parsing. Table 1 summarizes all hyper parameters we are using.

Layer	Hyper-parameter	Value
CNN	window size	3
	number of filters	50
LSTM	encoder layers	3
	encoder size	512
	decoder layers	1
	decoder size	512
MLP	arc MLP size	512
	label MLP size	128
Dropout	embeddings	0.33
	LSTM hidden states	0.33
	LSTM layers	0.33
Learning	Optimizer	Adam
	initial learning rate	0.01
	(β_1, β_2)	(0.9, 0.9)
	decay rate	0.75
	gradient clipping	5.0

Table 1: Hyper-parameters for all experiments

A.2 UD Treebanks

Table 2 shows the UD Treebank corpora used for language.

Language	Corpora
Bulgarian	BTB
Catalan	AnCora
English	EWT
French	GSD
German	GSD
Italian	ISDT
Romanian	RRT

Table 2: UD Treebanks corpora of the 7 languages that we have tested.

A.3 Hyper-Parameters for Discourse Parsing

The hyper-parameters used for discourse parsing experiments are listed in table 3.

Hyper-parameters	Value
Minibatch size	64
Embedding size	1024
Encoder hidden size	64
Decoder hidden size	64
ELMo dropout rate	0.5
Encoder dropout rate	0.4
Decoder dropout rate	0.6
Classifier dropout rate	0.5
Initial Learning rate	0.001
Adam β_1	0.9
Adam β_2	0.95
L_2 Regularization strength	0.0005

Table 3: Optimal hyper-parameter settings for parser