Character CNNs	
Char embedding size	16
(# Window Size, # Filters)	(1, 32), (2, 32), (3, 68), (4, 128), (5, 256), 6, 512), (7, 1024)
Activation	Relu
Word-level LSTM	
LSTM size	2048
# LSTM layers	2
LSTM projection size	256
Use skip connections	Yes
Inter-layer dropout rate	0.1
Training	
Batch size	128
Unroll steps (Window Size)	20
# Negative samples	64
# Epochs	10
Adagrad (Duchi et al., 2011) lrate	0.2
Adagrad initial accumulator value	1.0

Table 7: Language model hyperparameters.

Input	
Input dropout rate	0.3
Word-level BiLSTM	
LSTM size	400
# LSTM layers	3
Recurrent dropout rate	0.3
Inter-layer dropout rate	0.3
Use Highway Connection	Yes
Multilayer Perceptron, Attention	
Arc MLP size	500
Label MLP size	100
# MLP layers	1
Activation	Relu
Training	
Batch size	80
# Epochs	80
Early stopping	50
Adam (Kingma and Ba, 2015) lrate	0.001
Adam β_1	0.9
Adam β_2	0.999

Table 8: UD parsing hyperparameters.

A Training Parameters

In this section, we provide hyperparameters used in our models and training details for ease of replication.

A.1 Language Models

Seen in Table 7 is a list of hyperparameters for our language models. We use the publicly available code of Peters et al. (2018) for training.[14] Following Mulcaire et al. (2019), we reduce the LSTM and projection sizes to expedite training and to compensate for the greatly reduced training data—the hyperparameters used in Peters et al. (2018) were tuned for the One Billion Word Corpus (Chelba et al., 2013), while we used only 5% as much text (approximately 50M tokens) per language. Contextual representations from language models trained with even less text are still effective (Che et al., 2018; Schuster et al., 2019), suggesting that the method used in this work would apply to even lower-resource languages that have scarce text in addition to scarce or nonexistent annotation, though at the cost of some of the performance.

A.2 UD Parsing

For UD parsing, we generally follow the hyperparameters used for the dependency parsing demo in AllenNLP (Gardner et al., 2018). See a list of hyperparameters in Table 8. We use stratified sampling so that each training mini-batch has an equal number of sentences from the source and target languages.

A.3 Multilingual Word Vectors

We train our word type representations used for non-contextual baselines with fastText (Bojanowski et al., 2017). We use window size 5 and a minimum count of 5, with 300 dimensions.

B Other Low-Resource Simulations

In addition to the 100-sentence condition, we simulated low-resource experiments with 500 and 1000 sentences of target language data, and zero-target-treebank experiments in which the parser was trained with only source language data, but with multilingual representations allowing crosslingual transfer. See Table 9 for these results. The additional low-resource results confirm our analysis in Section 4.2: polyglot training is more effective the less target-language data is available, with a slight advantage for related languages.

C UD Treebanks

Additional statistics about the languages and treebanks used are given in Table 10.

D Additional Experiments

Semantic Role Labeling For SRL, we again follow the hyperparameters given in AllenNLP (Table 11). The one exception is that we used 4 layers of alternating BiLSTMs instead of 8 layers to expedite the training process.

[14] github.com/allenai/bilm-tf

	$\|D_\tau\| = 0$		$\|D_\tau\| = 100$			$\|D_\tau\| = 500$			$\|D_\tau\| = 1000$		
target	+eng	+rel.	mono	+eng	+rel.	mono	+eng	+rel.	mono	+eng	+rel.
ARA	10.31	**20.47**	62.50	73.39	**73.43**	76.15	**79.55**	79.16	79.43	81.38	**81.49**
HEB	23.76	**24.89**	64.53	74.86	**75.69**	79.27	82.35	**82.92**	82.59	84.59	**84.70**
HRV	48.69	**67.67**	63.49	79.21	**82.00**	80.80	84.92	**85.89**	84.14	86.27	**86.66**
RUS	38.69	**73.24**	59.51	75.63	**79.29**	77.38	83.16	**84.60**	82.90	85.68	**86.99**
NLD	61.68	**72.90**	57.12	74.90	**77.01**	75.19	**82.42**	81.33	81.41	**84.93**	83.23
DEU	51.18	**68.66**	60.26	72.52	**73.45**	72.94	**77.88**	77.68	76.46	**78.67**	78.57
SPA	55.85	**75.88**	64.97	80.86	**81.55**	79.67	**84.88**	84.63	82.97	86.69	**86.81**
ITA	59.71	**78.12**	69.17	**84.63**	83.51	82.96	**88.96**	87.91	87.03	**90.22**	89.32
CMN	**8.16**	5.34	53.36	**63.63**	61.47	71.94	74.88	**74.98**	77.42	**79.07**	78.96
JPN	4.12	**11.66**	72.37	**80.94**	80.24	86.20	**87.74**	**87.74**	88.74	89.08	**89.32**

Table 9: LAS for UD parsing with additional simulated low-resource and zero-target-treebank settings.

Named Entity Recognition We again use the hyperparameter configurations provided in AllenNLP. See Table 12 for details.

Lang	Code	WALS Genus	WALS 81A	Size (# sents.)	Treebank	Genre
English	eng	Germanic	SVO		EWT	blog, email, reviews, social
Simulation Pairs						
Arabic	ara	Semitic	VSO/SVO	5241	PADT	news
Hebrew	heb	Semitic	SVO		HTB	news
Croatian	hrv	Slavic	SVO	6983	SET	news, web, wiki
Russian	rus	Slavic	SVO		SynTagRus	contemporary fiction, popular, science, newspaper, journal articles, online news
Dutch	nld	Germanic	SOV/SVO	12269	Alpino	news
German	deu	Germanic	SOV/SVO		GSD	news, reviews, wiki
Spanish	spa	Romance	SVO	12543	GSD	blog, news, reviews, wiki
Italian	ita	Romance	SVO		ISDT	legal, news, wiki
Chinese	cmn	Chinese	SVO	3997	GSD	wiki
Japanese	jpn	Japanese	SOV		GSD	wiki
Truly Low Resource and Related Languages						
Hungarian	hun	Ugric	SOV/SVO	910	Szeged	news
Finnish	fin	Finnic	SVO	12217	TDT	news, wiki, blog, legal, fiction, grammar-examples
Vietnamese	vie	Viet-Muong	SVO	1400	VTB	news
Uyghur	uig	Turkic	SOV	1656	UDT	fiction
Kazakh	kaz	Turkic	SOV (not in WALS)	31	KTB	wiki, fiction, news
Turkish	tur	Turkic	SOV	3685	IMST	nonfiction, news

Table 10: List of the languages and their UD treebanks used in our experiments. Each shaded/unshaded section corresponds to a pair of *related* languages. WALS 81A denotes Feature 81A in WALS, Order of Subject, Object, and Verb (Dryer and Haspelmath, 2013). Size represents the downsampled size in # of sentences used for source treebanks.

Input	
Predicate indicator embedding size	100
Word-level Alternating BiLSTM	
LSTM size	300
# LSTM layers	4
Recurrent dropout rate	0.1
Use Highway Connection	Yes
Training	
Batch size	80
# Epochs	80
Early stopping	20
Adadelta (Zeiler, 2012) lrate	0.1
Adadelta ρ	0.95
Gradient clipping	1.0

Table 11: SRL hyperparameters.

Char-level LSTM	
Char embedding size	25
Input dropout rate	0.5
LSTM size	128
# LSTM layers	1
Word-level BiLSTM	
LSTM size	200
# LSTM layers	3
Inter-layer dropout rate	0.5
Recurrent dropout rate	0.5
Use Highway Connection	Yes
Multilayer Perceptron	
MLP size	400
Activation	tanh
Training	
Batch size	64
# Epochs	50
Early stopping	25
Adam (Kingma and Ba, 2015) lrate	0.001
Adam β_1	0.9
Adam β_2	0.999
L2 regularization coefficient	0.001

Table 12: NER hyperparameters.