

# Poverty of the Stimulus with CHILDES: Supplementary Materials

Aditya Yedetore

## CHILDES data pre-processing

All extraneous marks from the CHILDES NA-Eng corpora were removed.

**training, validation, and test split** 20% of the files in the CHILDES Treebank (Valian, Soderstrom, Brown, Suppes) were randomly selected and placed in the test set, for purposes of creating a test set for fine-tuning. 3% of the remaining files (by number of non-child utterances) were allocated to the validation and test sets each. The rest were allocated to the training set.

For the fine tuning data set, I

## Hyper-parameters and further model details

**LSTM** For LSTMs I explored the following hyper-parameters for a total of 144 models.

1. layers: 2
2. hidden and embedding size: 200, 800
3. batch size: 20, 80
4. dropout rate: 0.0, 0.2, 0.4, 0.6
5. learning rate: 5.0, 10.0, 20.0
6. random seed: 1001, 1002, 1003 (...)

Each had it's own random seed, which ranged from 1001 to 1144.

The 5 LSTM models with the lowest perplexities after 40 training epochs are reported in Table 1.

nlayers	nhidden/embed	lr	batch_size	dropout	seed	test loss	test ppl
2	800	20	80	0.4	1135	3.25	25.70
2	800	10	20	0.4	1095	3.25	25.84
2	800	5	20	0.4	1093	3.26	25.98
2	800	10	80	0.4	1131	3.26	26.06
2	800	20	20	0.4	1097	3.26	26.13