Colorless green recurrent networks dream hierarchically: Supplementary Material

Treebank data post-processing

In Hebrew, for consistency with the Wikipedia training data, we tokenized the sentences based on UD morpheme annotation, by treating orthographic clitics that can precede the word (conjunctions, prepositions and definite articles) as separate tokens. We also modified the morphological annotation to reflect the definiteness of nouns with possessive suffixes, to ensure that indefinite nouns are only substituted by other indefinites. Finally, since the Hebrew treebank is the smallest one in our set, we use the automatically parsed Hebrew Wikipedia to extract words annotated with their morphological information.

For English, the UD treebank only marks third-person singular present tense verbs. We added explicit plural features to the other finite present tense verb forms.

Hyperparameters and further model details

LSTM For LSTMs we explored the following hyperparameters, for a total of 68 combinations (we only considered two-layer models):

1. hidden and embedding size: 200 and 650
2. batch size: 20 (only for 200-units models), 64, 128
3. dropout rate: 0.0, 0.1, 0.2, 0.4 (for 650-units models only)
4. learning rate: 1.0, 5.0, 10.0, 20.0

The LSTM results reported in the paper are averaged over the 5 models that achieved the lowest validation perplexity after 40 training epochs. Their individual accuracies are reported in Table 1.

	accuracy			ppl	hidden/	batch size	dropout rate	learning rate
	total	orig	nonce		embedding size			
Italian	85.4	90.8	84.8	44.9	650	64	0.2	20.0
	87.2	93.3	86.6	45.0	650	64	0.2	10.0
	86.2	93.3	85.4	45.1	650	128	0.2	20.0
	86.5	93.3	85.7	45.3	650	64	0.1	20.0
	85.6	89.9	85.2	45.6	650	128	0.1	20.0
English	72.7	80.5	71.8	51.9	650	128	0.2	20.0
	76.8	82.9	76.2	51.9	650	64	0.2	10.0
	75.4	78.0	75.1	52.1	650	64	0.2	20.0
	74.1	80.5	73.4	52.1	650	64	0.1	20.0
	74.9	82.9	74.0	52.6	650	128	0.1	20.0
Hebrew	81.3	95.2	79.8	42.3	650	64	0.1	20.0
	82.3	94.4	81.0	42.4	650	128	0.2	20.0
	83.2	94.9	81.9	42.5	650	64	0.2	20.0
	82.3	94.9	80.8	42.7	650	64	0.1	10.0
	81.7	94.4	80.3	42.7	650	128	0.1	20.0
Russian	90.8	96.6	90.2	48.1	650	64	0.2	20.0
	89.6	97.1	88.8	48.6	650	64	0.1	20.0
	88.6	95.9	87.8	49.1	650	64	0.1	10.0
	89.0	95.2	88.3	49.2	650	128	0.1	20.0
	89.8	95.7	89.1	49.6	650	128	0.2	20.0

Table 1: Top 5 LSTMs.

sRNN For sRNNs we explored the following hyperparameters, for a total of 24 models:

1. hidden/embedding size: 650
2. batch size: 64, 128
3. dropout rate: 0.0, 0.1, 0.2
4. learning rate: 0.5, 1.0, 5.0, 10.0

Results obtained in all languages by the top-5 sRNNs by perplexity after 60 training epochs are reported in Table 2.

5-gram LSTM The 5-gram LSTM is trained as a count-based n-gram model, that is, during training it observes all 5-grams in the text, moving ahead of one token at each time step. For each 5-gram, the LSTM is re-initialized with a blank (all zeroes) hidden state. We tested only models with 650 units in the hidden and embedding layers, and explored the following hyperparameters, for a total of 24 models:

1. batch size: 256, 512, 1024
2. dropout rate: 0.0, 0.1, 0.2
3. learning rate: 1.0, 5.0, 10.0, 20.0

The 5-gram LSTM results reported in the paper are averaged over the 5 models that achieved the lowest validation perplexities, reported in Table 3.

Results by pattern

See Table 4 and Table 5 for Hebrew and Russian, respectively (Italian and English results are available in the main paper).

Human data

Subjects were recruited through the standard Amazon Mechanical Turk interface.[1] Instructions were presented in Italian, and subjects were paid 2 dollar cents per assignment. We did not record personal data about subjects, and our data set only contains aggregated data.

[1] `https://www.mturk.com/`

	accuracy			ppl	hidden/	batch size	dropout rate	learning rate
	total	orig	nonce		embedding size			
Italian	77.2	86.6	76.2	64.0	650	64	0.0	1.0
	77.7	84.0	77.0	64.5	650	128	0.0	1.0
	76.8	84.0	76.0	65.8	650	64	0.0	0.5
	78.5	85.7	77.7	66.1	650	64	0.1	1.0
	78.2	87.4	77.1	67.8	650	64	0.1	0.5
English	65.1	73.2	64.2	69.8	650	64	0.0	1.0
	65.6	82.9	63.7	70.6	650	128	0.0	1.0
	69.5	70.7	69.4	72.9	650	64	0.1	1.0
	63.9	68.3	63.4	73.3	650	64	0.0	0.5
	65.9	75.6	64.8	74.4	650	128	0.1	1.0
Hebrew	75.8	88.5	74.4	60.9	650	128	0.0	1.0
	75.5	88.7	74.0	62.8	650	64	0.0	1.0
	74.8	88.2	73.3	63.3	650	64	0.0	0.5
	75.7	88.2	74.3	63.7	650	64	0.1	1.0
	76.0	87.9	74.6	63.9	650	128	0.1	1.0
Russian	82.4	87.6	81.9	65.1	650	64	0.1	1.0
	81.6	87.1	80.9	65.6	650	64	0.0	1.0
	81.3	86.4	80.7	68.2	650	128	0.0	5.0
	81.7	87.1	81.1	68.4	650	128	0.1	10.0
	82.0	87.6	81.3	68.6	650	128	0.1	1.0

Table 2: Top 5 sRNNs.

	accuracy			ppl	hidden/	batch size	dropout rate	learning rate
	total	orig	nonce		embedding size			
Italian	79.5	83.2	79.1	62.4	650	1024	0.2	10.0
	78.5	77.3	78.6	62.4	650	256	0.2	5.0
	76.6	79.8	76.3	62.4	650	512	0.2	5.0
	77.8	84.0	77.1	62.7	650	1024	0.2	5.0
	79.7	84.9	79.1	62.9	650	1024	0.1	10.0
English	58.0	70.7	56.6	71.6	650	512	0.2	5.0
	61.5	65.9	61.0	71.7	650	1024	0.1	10.0
	56.6	63.4	55.8	71.8	650	1024	0.2	10.0
	61.5	78.0	59.6	71.8	650	1024	0.2	5.0
	59.5	73.2	58.0	72.5	650	1024	0.1	5.0
Hebrew	78.4	90.1	77.1	59.7	650	256	0.1	5.0
	80.1	92.5	78.7	59.7	650	1024	0.2	10.0
	79.2	90.3	77.9	59.8	650	256	0.2	5.0
	77.8	89.8	76.5	60.1	650	512	0.1	5.0
	78.7	91.7	77.2	60.1	650	1024	0.1	10.0
Russian	86.2	91.9	85.5	61.1	650	1024	0.2	10.0
	86.7	91.0	86.3	61.6	650	512	0.1	10.0
	87.0	91.6	86.5	61.7	650	512	0.2	5.0
	85.6	91.4	85.0	62.5	650	1024	0.2	5.0
	85.9	91.9	85.2	62.8	650	512	0.2	10.0

Table 3: Top 5-gram LSTMs.

construction	acc original	*std*	acc nonce	*std*	size
NOUN ADJ PROPN VERB	100.0	0.0	61.8	9.5	50
NOUN NOUN ADJ VERB	100.0	0.0	90.7	3.5	60
NOUN NOUN PROPN VERB	100.0	0.0	80.0	4.2	30
NOUN PROPN VERB	100.0	0.0	100.0	0.0	10
NOUN VERB PUNCT VERB	95.0	11.2	81.7	4.2	40
NOUN VERB VERB	86.7	9.3	65.7	4.1	90
NOUN ADJ DET ADJ	99.4	0.8	96.4	1.0	710
NOUN ADJ PUNCT SCONJ VERB	98.8	2.8	89.6	1.8	160
NOUN DET ADV PUNCT ADJ	93.3	5.3	81.2	4.6	300
NOUN NOUN DET ADJ	94.2	1.0	77.6	0.9	1520
NOUN NOUN PUNCT SCONJ VERB	90.5	6.7	73.9	2.7	210
NOUN PUNCT NOUN CCONJ NOUN	96.7	4.6	50.7	3.7	120
NOUN PUNCT NOUN PUNCT NOUN	98.3	3.7	68.3	4.1	120
VERB NOUN CCONJ VERB	83.3	5.9	83.1	2.8	240
VERB NOUN NOUN CCONJ VERB	100.0	0.0	80.0	9.3	10
VERB NOUN PUNCT CCONJ VERB	100.0	0.0	73.9	5.4	40
VERB VERB CCONJ VERB	90.0	22.4	73.3	9.1	20

Table 4: Average top 5 LSTM accuracies for Hebrew.

construction	acc original	*std*	acc nonce	*std*	size
ADJ ADJ NOUN	99.1	0.9	94.1	0.9	910
ADJ DET NOUN	100.0	0.0	96.9	0.5	370
ADJ NOUN ADJ NOUN	97.3	3.7	96.6	1.5	150
ADJ NOUN NOUN	88.6	2.6	69.6	2.2	420
ADJ PUNCT ADJ NOUN	99.4	0.9	94.5	0.8	620
ADJ PUNCT NOUN NOUN	72.0	11.0	71.1	4.4	50
ADJ PUNCT VERB NOUN	80.0	18.3	84.4	3.1	30
ADJ VERB NOUN	80.0	11.2	83.3	5.9	40
DET ADJ ADJ NOUN	100.0	0.0	99.2	1.1	190
DET ADJ NOUN	97.1	2.6	94.6	1.8	210
DET VERB NOUN	97.5	5.6	75.8	5.8	80
NOUN PUNCT VERB ADV VERB	100.0	0.0	85.6	3.0	40
VERB ADJ NOUN	94.0	5.5	86.7	2.2	100
VERB NOUN ADJ NOUN	100.0	0.0	98.5	1.6	90
VERB NOUN NOUN	87.0	5.3	66.7	3.4	230
NOUN PUNCT CCONJ PUNCT NOUN	93.3	14.9	83.7	7.2	30
VERB NOUN CCONJ PART VERB	98.2	4.1	91.7	2.3	110
VERB NOUN CCONJ VERB	95.2	1.9	86.7	1.6	670
VERB NOUN NOUN CCONJ VERB	100.0	0.0	96.1	1.5	40
VERB PROPN CCONJ VERB	100.0	0.0	93.9	4.6	40

Table 5: Average top 5 LSTM accuracies for Russian.