Supplementary Material

Number Agreement data

For our experiments we use several 7 synthetically generated datasets, that each contain sentences with one particular syntactic structure, and varied lexical material. Each construction is stored in a distinct file. Each file contains 4 columns, that are separated by tabs:

1. the first column contains the sentence
2. the second column contains the number of the subject (singular or plural)
3. the third column contains whether the main verb of the sentence agrees with the subject ('correct') or not ('wrong'). Each of the sentences is present twice, once with the correct agreement and once with the wrong one.
4. the fourth column contains the sentence id (which is the same for both correct and wrong version of the same sentence).

The results of the paper can be reproduced using the scripts in our github repository, which will be made available upon acceptance.

Tree Depth Data

For our regression experiments we used a large corpus with sentences with unambiguous but varied syntactic structures, generated by a script that follows a predefined context-free grammar. The output of this script is a four column file containing:

1. The generated sentence.
2. The syntactic parse tree of the sentence, according to the grammar that was used to generate it.
3. The number of open nodes (syntactic tree-depth), following Nelson et. al, 2017.
4. The number of adjacent open and closing brackets before each word

The columns of the file are separated with the character |.

We first processed a large corpus of such sentence with lengths between 2 and 25 words with our LSTM language model and stored the activations of all gates, the 2 hidden layers and memory cells of the model in a pickled dictionary. Aside from the sentence and the corresponding activations, this dictionary contains also the position, log probability, syntactic depth, frequency count and open nodes count for each word in this sentence, and the length and structure of the sentence.

Since syntactic depth is naturally correlated with word position, we filtered the processed words such that all position-depth combinations within positions 7-12 and depths 3-8 are uniformly represented in our final dataset. Note that the datapoints for our regression analysis are thus (word activation, tree depth) pairs. As a results from our sampling strategy, only one or a few pairs for each sentence in the original dataset are included. Our final dataset contains 4,033 positions from 1,303 sentences.

We provide the script that generates these sentences, as well as the files containing the sentences and filtered data based on the above decorrelation method used in our study.