Learning Agreement with Deep Neural Networks

Jean-Philippe Bernardy and Shalom Lappin

University of Gothenburg

`jean-philippe.bernardy@gu.se, shalom.lappin@gu.se`

1. Introduction

We consider the extent to which different deep neural network (DNN) configurations can learn syntactic relations, by taking up Linzen et al.'s (2016) work on subject-verb agreement with LSTM RNNs. We test their methods on a much larger corpus than they used: a ∼24 million example part of the WaCky corpus (Baroni et al., 2009), instead of their ∼1.35 million example corpus, both drawn from Wikipedia. We experiment with several different DNN architectures (LSTM RNNs, GRUs, and CNNs), and alternative parameter settings for these systems. We also try out our own unsupervised DNN language model. Our results are broadly compatible with those that Linzen et al. report. However, we discovered some interesting, and in some cases, surprising features of DNNs and language models in their performance of the agreement learning task. In particular, we found that DNNs require large vocabularies to form substantive lexical embeddings in order to learn structural patterns. This finding has significant consequences for our understanding of the way in which DNNs represent syntactic information.

As Linzen et al. observe, the agreement task increases in difficulty in relation to the length of the sequence of NPs with the wrong number feature that occur between a subject and the verb that it controls. They refer to such intervening NPs as *attractors*.

(1a) *The students submit* a final project to complete the course.

(b) *The students* enrolled in **the program** *submit* a final project to complete the course.

(c) *The students* enrolled in **the program** in **the Department** *submit* a final project to complete the course.

(d) *The students* enrolled in **the program** in **the Department** where **my colleague** teaches *submit* a final project to complete the course.

Our main objective in the work that we report here is to explore the capabilities, and the limitations of DNNs for learning complex syntactic relations which depend on structural properties of sentences.

We used methods similar to Linzen et al.'s to test several DNN models on a much larger corpus. We experimented with different DNN architectures, and with alternative values for the following parameters: (i) ratio of training to testing as a partition of the corpus, (ii) number of hidden units (memory size), (iii) vocabulary size (iv) number of layers, (v) dropout rate, and (vi) lexical embedding dimension size.

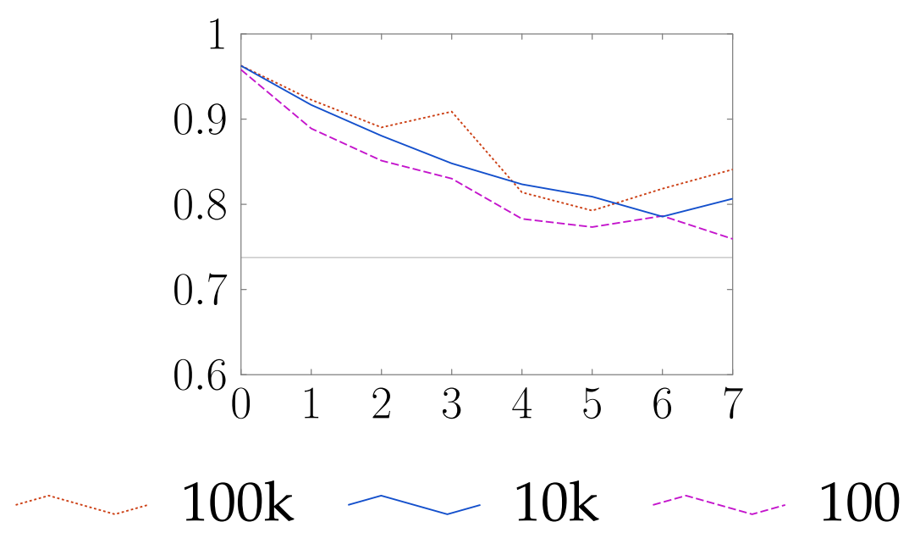
In addition we applied our own language model to the number prediction task, testing two distinct methods of predicting verb number from the model's probability distribution.

Specifically, by working with different vocabulary sizes, lexical encodings, and embedding sizes we discovered that our supervised DNN models learn agreement patterns more effectively from rich word embeddings than from abstract syntactically annotated input. We also found that our models required larger amounts of training relative to testing than Linzen et al. describe for their system, in order to reach the performance that they report. Increasing the values of hyperparameters generally improves accuracy, although after a point, overfitting is observed. Changing an individual hyperparameter does not create dramatic effects, but the global effect of hypeparameter optimisation is significant.

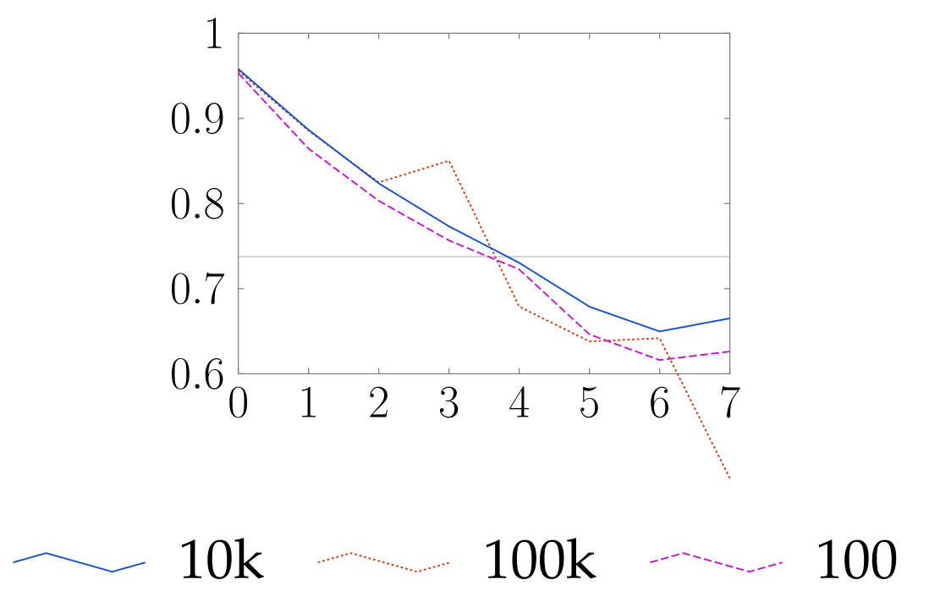
Finally, we were able to construct a language model with significantly better (unsupervised) prediction. Yet, our model is much smaller than the Google LM. All of our supervised DNNs outperformed our language model on the number prediction task.

2. Results

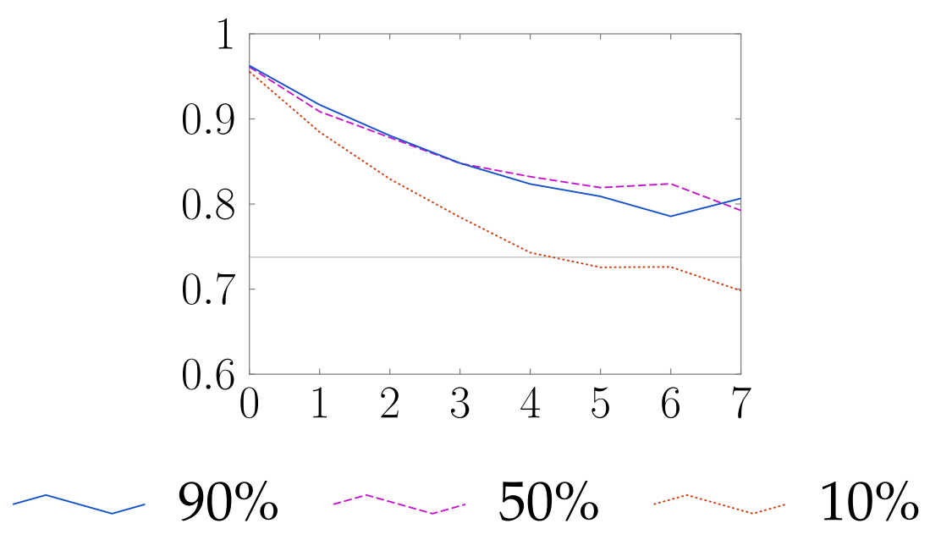
Except for the final graph, the y-axis gives the accuracy of agreement prediction rate, and the x-axis the number of NP attractors. Unless otherwise indicated, results are for a benchmark LSTM RNN with one layer of 150 units, no dropout, lexical embeddings of dimension 50, and training on 90% of the corpus.



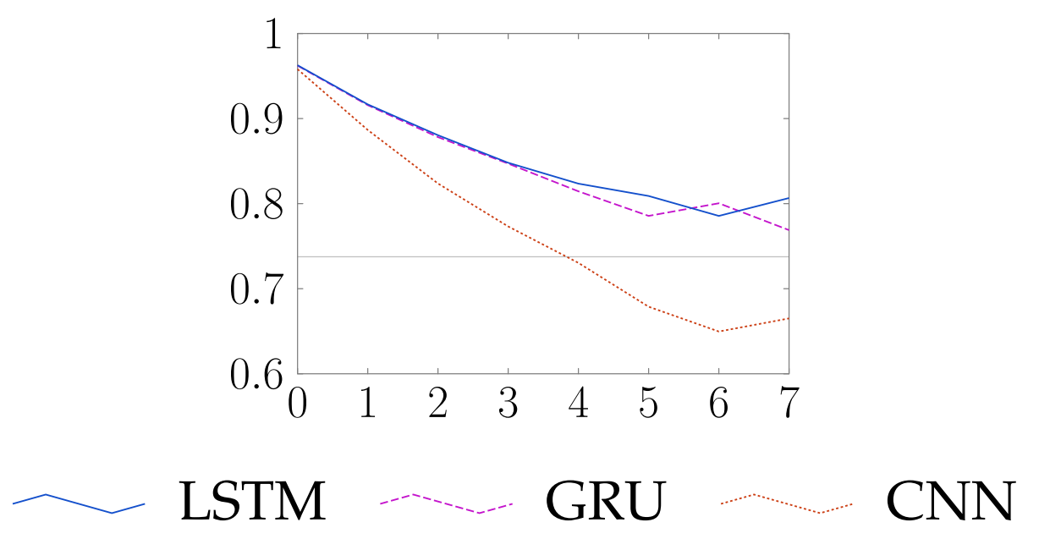
Comparing voc. size for benchmark LSTM RNN



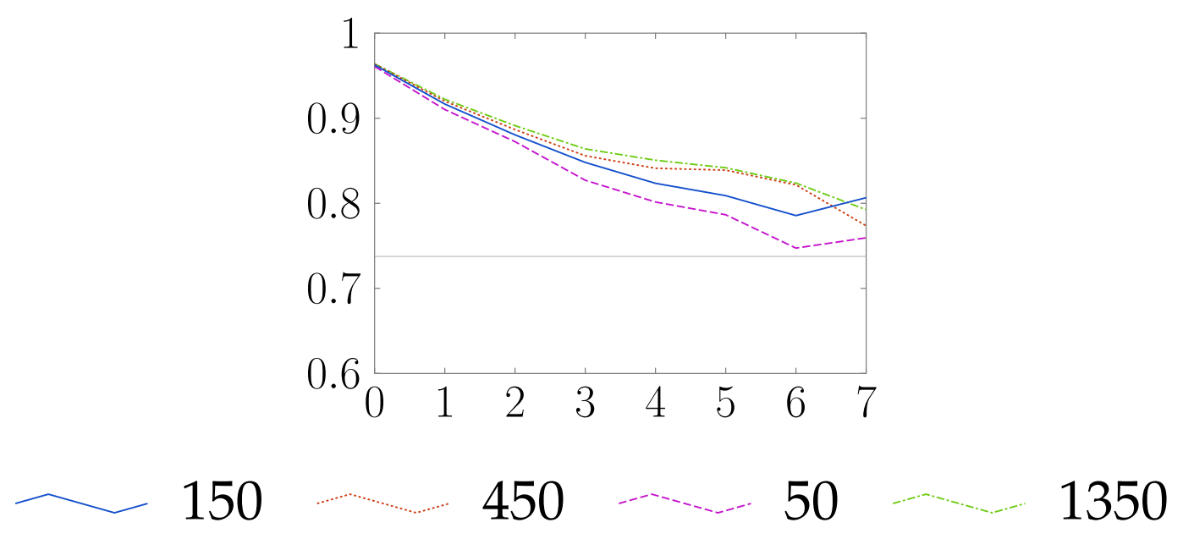
Comparing voc. size, for CNN architecture. Training set 90%



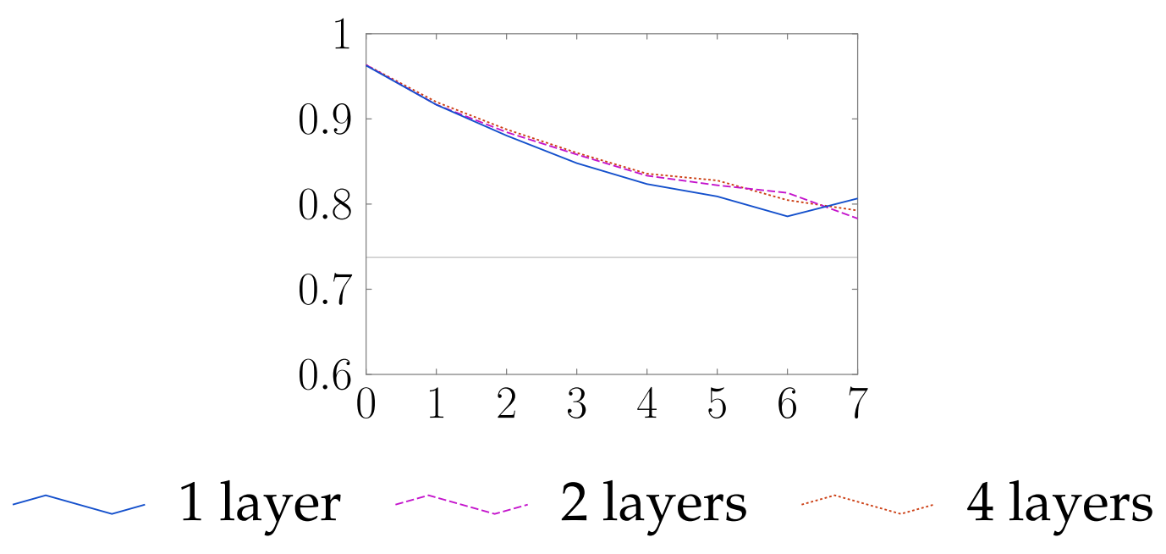
Comparing size of training set for an LSTM RNN



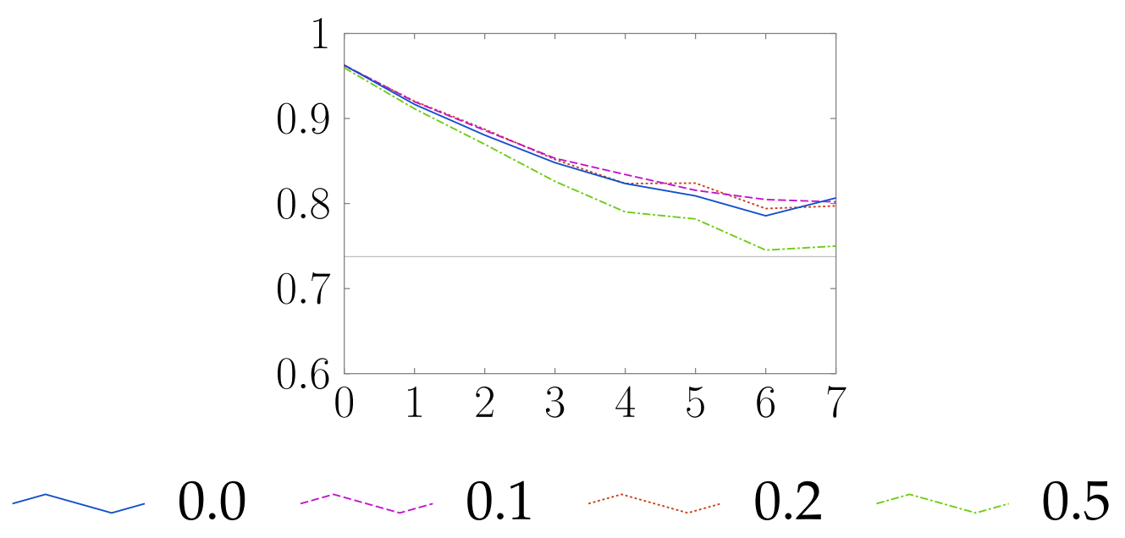
Comparing architectures. (Both RNNs use 150 units.)



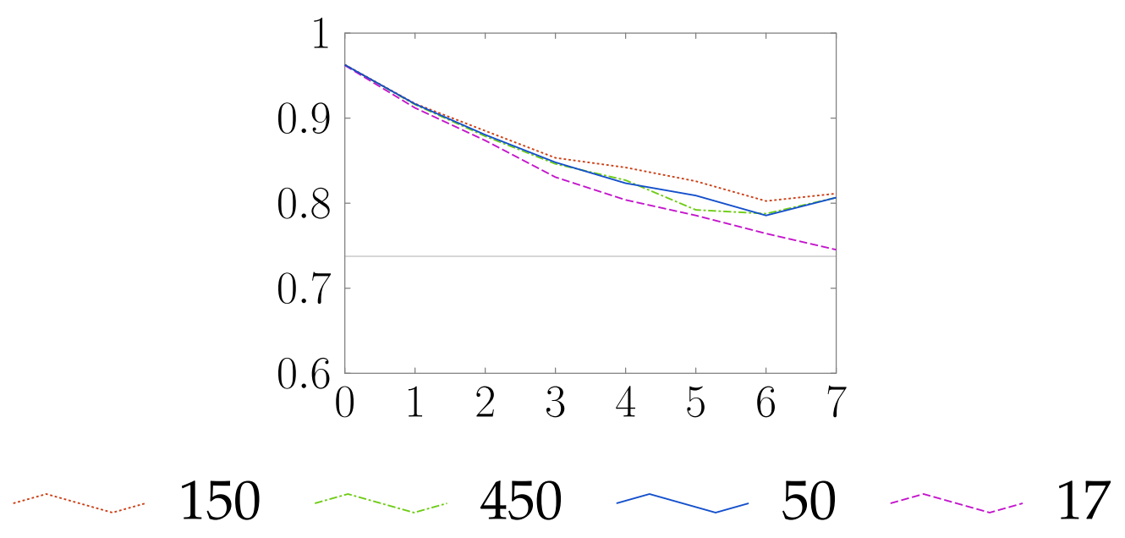
Comparing memory size



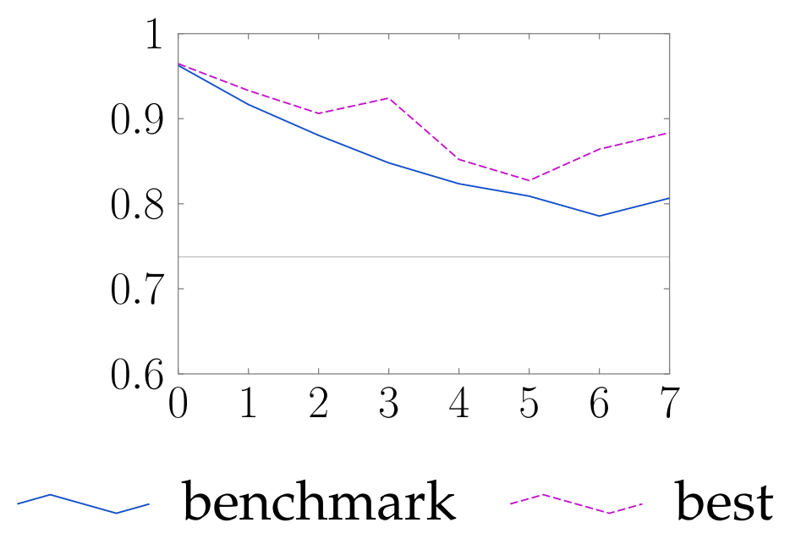
Comparing number of layers



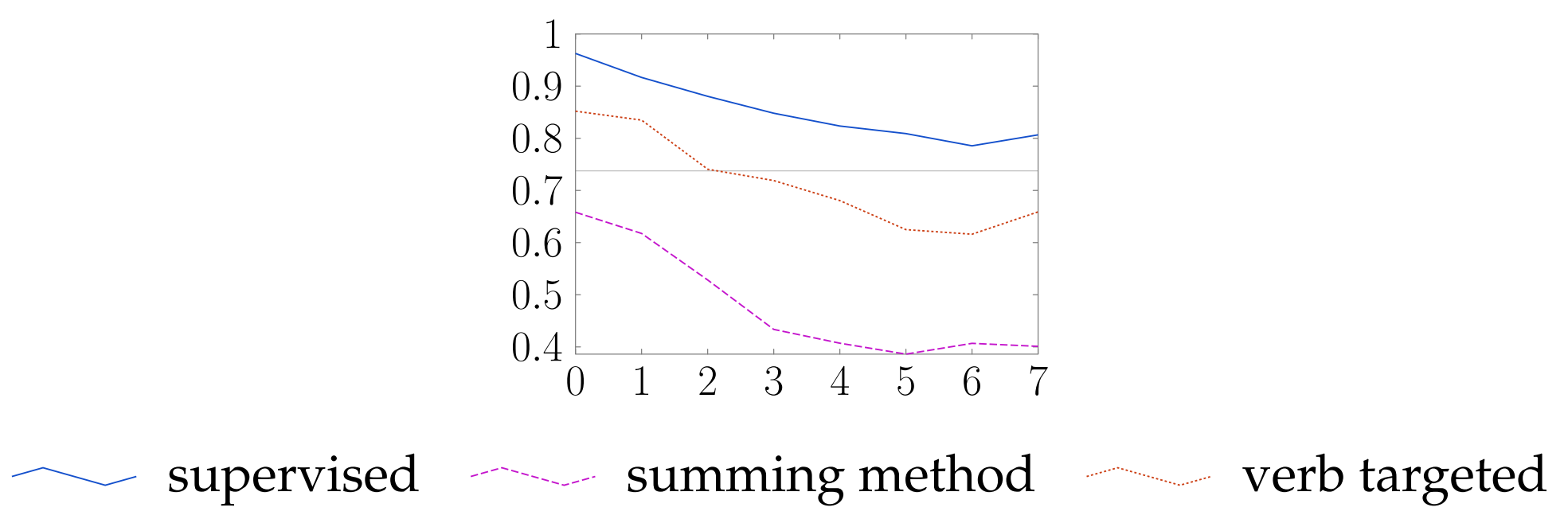
Comparing dropout rates



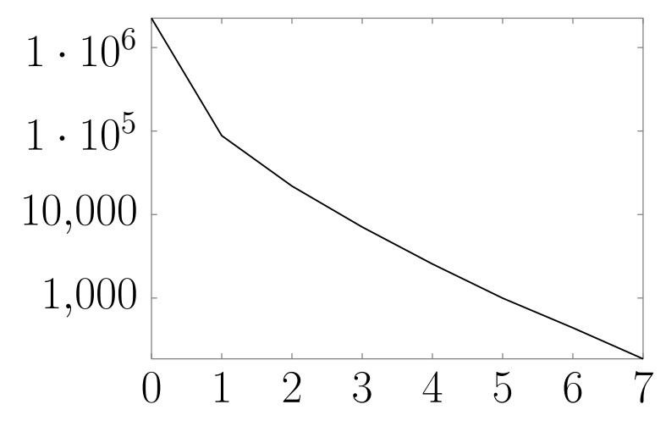
Varying embedding dimensions



Configuration with best parameters values: LSTM RNN with 2 layers of 1350 units, dropout rate 0.1, vocabulary size 100k, training on 90%, and lexical embedding dimension size 150



Comparing LSTM trained language model (with voc. size 100 and 1000 units) for the two methods of predicting verb number. The solid blue line represents our (supervised) benchmark LSTM RNN.



Number of test examples per number of attractors

3. Discussion

It could be that an RNN with a more structured memory that incorporates the equivalent of a stack for encoding the beginning of a dependency and a pop mechanism for releasing it later in a sequence (Grefenstette et al., 2015) would yield even better results. In general, there is considerable room for exploring alternative architectures before drawing strong conclusions on the capabilities of the entire class of DNNs for learning syntactic relations.

One of our most striking results is that training DNNs on data that is lexically impoverished, but highlights the syntactic elements between which a relation is to be acquired does not facilitate learning, but degrades it. DNNs learn better from data populated by richer lexical sequences. This suggests that DNNs are not efficient at picking up abstract syntactic patterns when they are explicitly marked in the data. Instead they extract them incrementally from lexical embeddings through recognition of their distributional regularities. It is also possible that they use the lexical semantic cues that larger vocabularies introduce to determine agreement preferences for a verb.

It is interesting to note that some recent work in neurolinguistics indicates that syntactic knowledge is distributed through different language centres in the brain, and closely integrated with lexical-semantic representations (Blank et al., 2016). This lexically encoded and distributed way of representing syntactic information is consistent with the role of rich lexical embeddings in DNN syntactic learning.

Finally our results show that a language model can achieve not entirely unreasonable results on the number agreement prediction task, if an appropriate method is applied for comparing the conditional probabilities of alternative number markings on verbs.

4. Conclusions and Future Work

Our experimental results strengthen Linzen et al.'s conclusion that DNNs are able to learn long distance syntactic relations to a fairly high degree of accuracy, across extended complex sequences of potentially distracting phrases. We also found that accuracy in the supervised version of this task scales with the amount of training data used, and with the size of the lexical embedding vocabulary.

Performance improves with an increase in the number of hidden units. This effect may be even more pronounced when tracking more complex syntactic relations with multiple features. This is a question that we will explore in future work.

We also found that it is possible to obtain reasonable results with unsupervised learning through a comparatively small language model, when we use a targeted procedure for predicting verb number. The performance of this model, with this procedure, significantly exceeds that of the two models that Linzen et al. present.

One of our main concerns will be to explore syntactic dependencies involving several agreement features. In languages in which gender, and person, as well as number are morphologically realised on verbs the agreement prediction task is more difficult. It requires accuracy across three feature dimensions rather than one. Testing DNNs on agreement in such languages will provide a better sense of their capacity to learn and represent syntactic information.

Acknowledgments

The research reported here was supported by a grant from the Swedish Research Council, which funds the Centre for Linguistic Theory and Studies in Probability in the Department of Philosophy, Linguistics, and Theory of Science at the University of Gothenburg. We are grateful to Tal Linzen for valuable discussion of some of the ideas presented in this paper. Earlier versions of the paper were presented to the Chalmers Machine Learning Seminar in April 2017 and the colloquium of the Cambridge Linguistics Society in May 2017. We thank the audiences of these venues for helpful comments and suggestions.

References

Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation 43*, 209–226.

Blank, I., Z. Balewski, K. Mahowald, and E. Fedorenko (2016). Syntactic processing is distributed across the language system. *NeuroImage 127*, 307–323.

Grefenstette, E., K. M. Hermann, M. Suleyman, and P. Blunsom (2015). Learning to transduce with unbounded memory. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, Cambridge, MA, USA, pp. 1828–1836. MIT Press.

Linzen, T., E. Dupoux, and Y. Golberg (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association of Computational Linguistics 4*, 521–535.

ISCOL, Hebrew University of Jeruslaem, September 25, 2017