

# Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains

Samira Abnar   Lisa Beinborn   Rochelle Choenni   Willem Zuidema

Institute for Logic, Language and Computation  
University of Amsterdam

{abnar, l.beinborn}@uva.nl, rochelle.choenni@student.uva.nl, zuidema@uva.nl

## Abstract

In this paper, we define and apply *representational stability analysis* (ReStA), an intuitive way of analyzing neural language models. ReStA is a variant of the popular *representational similarity analysis* (RSA) in cognitive neuroscience. While RSA can be used to compare representations in models, model components, and human brains, ReStA compares instances of the *same* model, while systematically varying single model parameter. Using ReStA, we study four recent and successful neural language models, and evaluate how sensitive their internal representations are to the amount of prior context. Using RSA, we perform a systematic study of how similar the representational spaces in the first and second (or higher) layers of these models are to each other and to patterns of activation in the human brain. Our results reveal surprisingly strong differences between language models, and give insights into where the *deep* linguistic processing, that integrates information over multiple sentences, is happening in these models. The combination of ReStA and RSA on models and brains allows us to start addressing the important question of what kind of linguistic processes we can hope to observe in fMRI brain imaging data. In particular, our results suggest that the data on story reading from Wehbe et al. (2014) contains a signal of *shallow* linguistic processing, but show no evidence on the more interesting *deep* linguistic processing.

## 1 Representational Similarity

Representational similarity analysis (RSA) is a technique which allows us to compare heterogeneous representational spaces (Laakso and Cottrell, 2000). It is very common in cognitive neuroscience because it allows researchers to study the relation between patterns of activation in the brain and representations of stimuli in a computational model (Kriegeskorte et al., 2008). The key idea

is simple: instead of directly trying to map models to brains, we first construct two similarity matrices that record how similar brain responses are to each other for different stimuli, and how similar the computational model’s representations for each stimulus are to each other. The representational similarity score is then defined as the similarity (typically: Pearson’s correlation) of the two similarity matrices (or equivalently: the similarity of two distance matrices).

RSA can also be applied to deep learning models (Laakso and Cottrell, 2000; Dharmaretnam and Fyshe, 2018; Alvarez-Melis and Jaakkola, 2018; Wang et al., 2018; Chrupała and Alishahi, 2019). In this paper, we present a large-scale study and comparison of both neural language models and fMRI data from brain imaging experiments with human subjects, using RSA. However, we extend standard RSA using an approach we call *Representational Stability Analysis* (ReStA). The idea is again simple: we apply RSA to compare instances of the *same* model, while systematically varying a model parameter.

We focus on a single parameter: the length of the prior context presented to the model. Varying the amount of context allows us to quantify the degree of context-dependence of different neural language models, and different components of those models. If internal representations are similarly organized regardless of how much additional context is presented to the model, context-dependence is low. If, on the other hand, representations change with each additional amount of context included, context-dependence is high. Using this approach, we find intriguing differences between some recent, successful neural language models (GoogleLM, ELMO, BERT and the Universal Sentence Encoder; Table 1), and between the first and deeper layers of those models.

Context-dependence, in turn, gives us a handle on an important question in the research that tries

Model	Objective	Corpus	Rep.Dim.	Architecture
GloVe (Pennington et al., 2014)	Predicting co-occurrence probabilities	Wikipedia	300	Bag of words
ELMO (Peters et al., 2018)	Bidirectional Language Modelling	1B benchmark	1,024	BiLSTM
GoogleLM (Jozefowicz et al., 2016)	Language Modelling	1B benchmark	1,024	LSTM
UniSentEnc. (Cer et al., 2018)	Skip-Thought/Classification	Variety of web sources / SNLI	512	Transformer Encoder
BERT (base) (Devlin et al., 2019)	Masked Language Modelling / Next Sent. Pred.	BooksCorpus / English Wikipedia	768	Transformer Encoder

Table 1: Details of the third party computational models used in this paper, including a brief characterization of the optimization objective, the training corpus, and the dimensionality of representations we extract from them.

to link neural language models to brain activation: which aspects of language processing in the brain can we hope to observe in fMRI data using NLP and machine learning tools?

## 2 Bridging NLP Models and Neurolinguistics

An important motivation behind our work is to contribute to answering a big question in computational linguistics: how do we establish a relationship between NLP models and data on the human brain activation while they process language? Pioneering work of Mitchell et al. (2008) showed that techniques from distributional semantics could be used to predict and decode brain activation. In the decade since that paper, many efforts have been reported using brain data to evaluate computational models, or using NLP models to build predictive models of the human brain, or both (Murphy et al., 2012; Wehbe et al., 2014a; Ruan et al., 2016; Sogaard, 2016; Xu et al., 2016; Fyshe et al., 2014; Bingel et al., 2016; Bulat et al., 2017; Abnar et al., 2018; Pereira et al., 2018; Huth et al., 2016).

Most of that work is focused on lexical representations, reporting promising results for concrete nouns, presented in isolation. More recently researchers have tried to adapt the methodology to address words in context, in sentence and story processing tasks. Pereira et al. (2018), for instance, used a bag of words model of sentence meaning to decode sentences from brain activation. Wehbe et al. (2014b); Qian et al. (2016) use the internal states of LSTMs trained for language modelling for encoding. Jain and Huth (2018) report that the higher layers of the LSTM are better at predicting the activation of brain regions that are known for higher level language functions (a find-

ing seemingly at odds with results from section 5).

In this effort, however, we run into a number of major conceptual, methodological and technical challenges. Most importantly: how do we determine what we are really observing in the brain data? Are we really seeing signatures of linguistic processes, or just neural correlates of general cognitive processes evoked by a correct understanding of the linguistic input? How do we adequately control for alternative explanations of the observed correlations? And how do we deal with the intricate temporal dynamics and the overwhelmingly high dimensionality of the brain, and the very indirect, delayed and/or coarse measurements that neuroimaging gives us of the processes in the brain? Merely demonstrating a correlation between two black boxes is clearly not sufficient.

We argue that experiments to find the model best correlated with brain activations should be accompanied by efforts for interpreting the internal representations and operations of the models. Applying ReStA for the prior context parameter gives us a way to roughly characterize the *depth* of linguistic processing in different language models and different components of these models. If a model component only tracks the lexical semantics of the current word, the representations it forms should not be sensitive to the amount of prior context. On the other hand, If a model component tracks long-distance syntactic dependen-

Block	Words	Unique words	Sentences	Sent Length	Scans
1	1583	553	115	11	326
2	1711	560	163	8	338
3	1411	461	134	8	265
4	1853	583	177	8	366

Table 2: Statistics of the Harry Potter dataset.

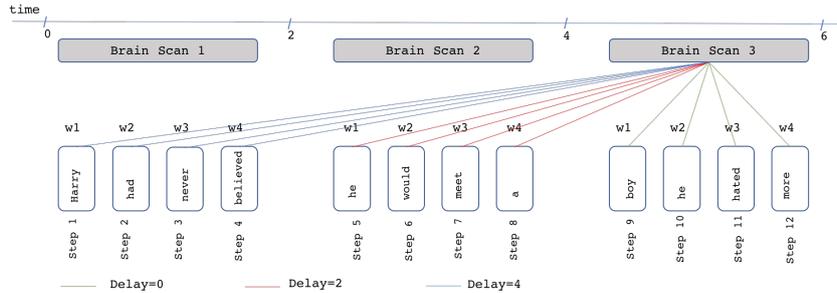


Figure 1: Alignment of the words in the story and the brain vectors. Each fMRI scan lasts for 2 seconds during which the subject is reading four words sequentially. Delay is the amount of time in seconds between the time the first of the four word is shown to the subject and when the fMRI scan is started to be taken.

cies, semantic polarity, named entities, topics or story arcs, resolves anaphora or builds up situation models, its representations will be different whenever different amounts of prior context are available. Hence, in this paper, we will interpret context-dependence as an imperfect but useful signature of deep linguistic processing.

### 3 Models and Data

In this section, we explain the language encoding models we study in our experiments and the dataset from which we get the language stimuli and their corresponding brain data.

#### 3.1 Neural Language Models

We study language models with different architectures trained with different objective functions (see Table 1). As a word level embedding model, we use GloVe (Pennington et al., 2014). We consider a sentence as a bag of words and take the average of the GloVe embeddings of its individual words.

We employ two high performing LSTM based language models: ELMO (Peters et al., 2018) and GoogleLM (Jozefowicz et al., 2016). Both of these models have two LSTM layers; however, ELMO uses bidirectional LSTM layers, whereas in the GoogleLM the LSTM layers are uni-directional. From these models, we take the internal states of each of the LSTM layers as two different representation spaces.

In our comparisons, we also use BERT and the Universal Sentence Encoder (UniSentEnc), as Transformer based models. BERT is trained on masked language modelling and next sentence prediction tasks (Devlin et al., 2019) while the Universal Sentence Encoder is trained on a different objective than language modelling. The parameters of this model are optimized with respect

to different language tasks such that it can better encode the meaning of complete sentences. These two models do not have the recurrent inductive bias of LSTMs, and hence the representations they learn can be completely different.

To study how and where the models integrate information over time, we modify the amount of context provided to the models to obtain the contextualized word representations. We do this at the sentence level. Thus, for the context length of 0, we only feed the target words to the models; For context length 1 we feed all the previous words in the current sentence to the models. For context length  $i$  where  $i > 1$ , in addition to the current sentence we feed all the words in the last  $i$  sentences. We operate on the sentence level to feed the model with independently meaningful pieces of text.

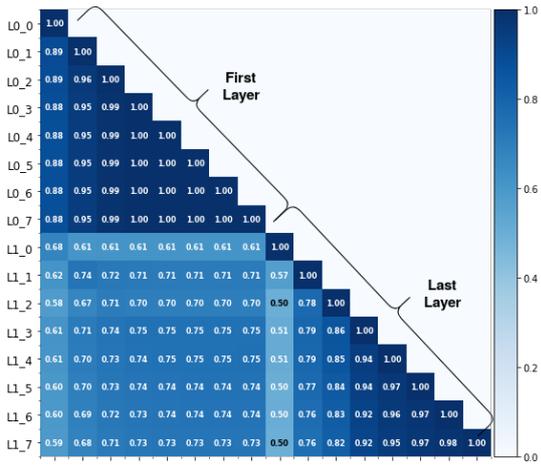
From prior work, we expect a relation between the depth of the layers and the level of abstraction of their representations. We study this intuition here empirically by analyzing the different layers of the models, and we focus on the first and last layers. Note that the last layer corresponds to the second layer for the LSTM architectures, but to the 12th layer for Bert.

#### 3.2 Brain Data

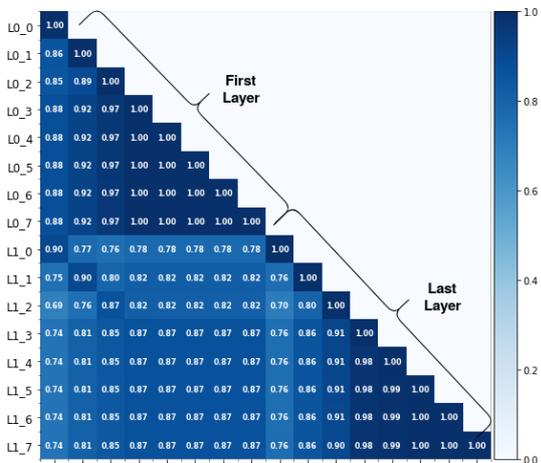
We compare the representations of our model to human brain activations captured while reading a story. We use the dataset by (Wehbe et al., 2014a) which consists of the fMRI scans of 8 participants reading chapter 9 of *Harry Potter and the Sorcerer’s stone* (Rowling, 1998).<sup>1</sup>

The story was presented to the participants word

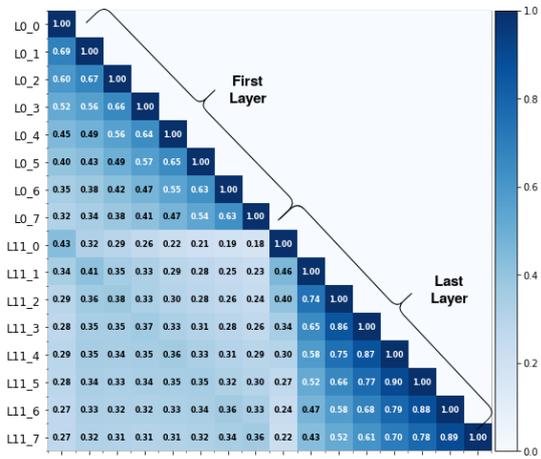
<sup>1</sup>The data is available at <http://www.cs.cmu.edu/~fmri/plosone/>. Further information on the pre-processing steps is described in the supplementary material.



(a) GoogleLM



(b) ELMO



(c) BERT

Figure 2: RSA between different layers of each model given different context length in terms of number of previous sentences over the story words. In these plots, for example L1\_c3 means representation from layer 1, when the context length is 3 sentences including the current sentence. When  $c = 0$ , the model only sees the current words and when  $c = 1$  the model sees current sentence up to the target word. Here darker means more similar. The values are averaged over the four story blocks and the standard deviation of all the values across the four blocks are below 0.002.

by word on a screen in four continuous blocks.<sup>2</sup> Each word was displayed for 0.5 seconds and an fMRI scan was taken every 2 seconds. Figure 1 visualizes an example for the beginning of the chapter. More detailed statistical information about the stimuli can be found in Table 2.

**Brain Regions** The fMRI data contains activation values for approximately 40,000 voxels per scan, each reflecting the oxygen usage (the “BOLD response”) in approximately  $3mm^3$  of brain tissue. To obtain the brain representations, we flatten the 3D fMRI images into vectors thereby ignoring the spatial relationships between the voxels. We do this either for the whole brain, or for specific regions separately. Not all of the scanned voxels are related to language processing, but the changes in activity might be associated with other cognitive processes like, for example, the noise perception in the scanner. A common reduction method is to restrict the brain response to voxels that fall within a pre-selected set of regions. In our analysis, we only include the voxels from the top  $k$  regions that are most similar across different subjects given the same stimuli. We heuristically set the value of  $k$  to 16 based on the distribution of the similarity scores.<sup>3</sup>

**Delay** An important point to consider when dealing with fMRI data is the hemodynamic response delay: from the time neurons start firing, it takes 4 to 6 seconds until the Bold response reaches its peak (Buckner, 1998). This means that from the time a stimulus is presented to a subject, it takes approximately 5 seconds before we can observe its response in the fMRI scan of the brain. We account for this delay by varying the alignment between stimuli and scans. If we apply a delay of 0 seconds, scan 3 in the example would be applied to the sequence *boy he hated more*, Figure 1. With a delay of 2 seconds, it is aligned to the previous stimulus *he would meet a* and a delay of 4 would result in alignment with *Harry had never believed*.

<sup>2</sup>The story chapter is split into four almost equal length blocks, each reflecting approximately 12 minutes of measurements. Each block is presented to the participant in one continuous trial, and experimental blocks are separated by pauses for the subjects.

<sup>3</sup>We sort the brain regions based on their cross-subject similarities for different stimuli and pick a threshold value based when there is a relatively big jump in the similarity scores.

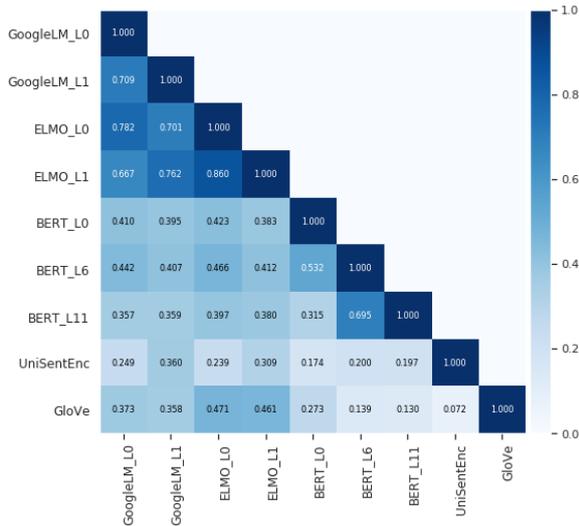


Figure 3: RSA across models

## 4 Analyzing Neural Language Models

In this section, we present the results of applying ReStA, Representational Stability Analysis, to three different language encoding models, GoogleLM, ELMO and BERT. We investigate what type of information is captured in the learned representations without making any explicit assumptions. Next, we apply standard RSA to, first, investigate the relations between different components of the language encoding models, and second to study the alignment of these components with the activity patterns in the human brain.<sup>4</sup>

### 4.1 Representational Stability Analysis

We define the *Representational Stability* as the similarity between the representations obtained from a model, when a single condition is changed, i.e. increase in context length. We use RSA to measure the similarity between the representational spaces. And to compute *RSA* we use *cosine* similarity to measure the intra-space similarities and use *Pearson* correlation to quantify the similarities across representational spaces.

In Figure 2 the representations of the different layers given different context lengths are compared for GoogleLM, ELMO and BERT. The values under the diagonal of these plots indicate the ReStA when the varying condition is context length. This is measured as  $RSA(L_{k-c_i}, L_{k-c_j})$ , where  $k$  is the layer id and  $c_i$  and  $c_j$  are differ-

<sup>4</sup>We made the code that reproduces all the experiments publicly available at <https://github.com/samiraabnar/Bridge>

ent conditions which in this case indicate different context lengths. We have depicted the trends of how the ReStA changes for different context length in Figures 4a and 4b.

**Effect of depth** As we can see in Figure 2 and more clearly in Figure 5, for the LSTM based models, we observe a higher degree of similarity between the two layers ( $\sim 0.75$  and  $\sim 0.80$ ) compared to BERT ( $\sim 0.35$ ). This can be partly explained by the higher number of layers in BERT, i.e the first and last layer are further apart. Moreover, the relation between the first and last layers is almost the same for all context lengths and for all these three models the two layers are most similar when provided with the same amount of context.

**Context sensitivity** Next, we analyse the sensitivity of different layers of each model to context length. In Figures 4a and 2, we see that for both LSTM based models, GoogleLM and ELMO, the first layer,  $L_0$ , is less sensitive to the changes in the context length compared to the last layer,  $L_1$ , i.e. the representations are not affected anymore by increasing the context length to more than 3 sentences. A hierarchical encoding mechanism, where the first layer is responsible for encoding the local context and the second(last) layer is encoding more global information, can justify these results.

We can see in Figure 4a, that the sensitivity to the context length is more significant in the Transformer based models compared to LSTM based models. In these models, the difference in the representations at different context lengths does not fade away as the context length increases but the rate of the changes becomes constant. As illustrated in Figures 4a and 2c we observe that in BERT, regardless of the current context length, adding more context leads to different representations. In addition, in this model, the representations from the first layer,  $L_0$  are more context-dependent than those from the last layer,  $L_{11}$ . Since in self-attention layers, there is a direct connection between the representations at different positions, the higher degree of sensitivity to context length is not surprising. This is evidence that, for computing the representations of each position in the input, the representations from all positions, no matter how far they are, are in fact taken into account. We speculate that the last layer of BERT is less sensitive to context could be that in higher

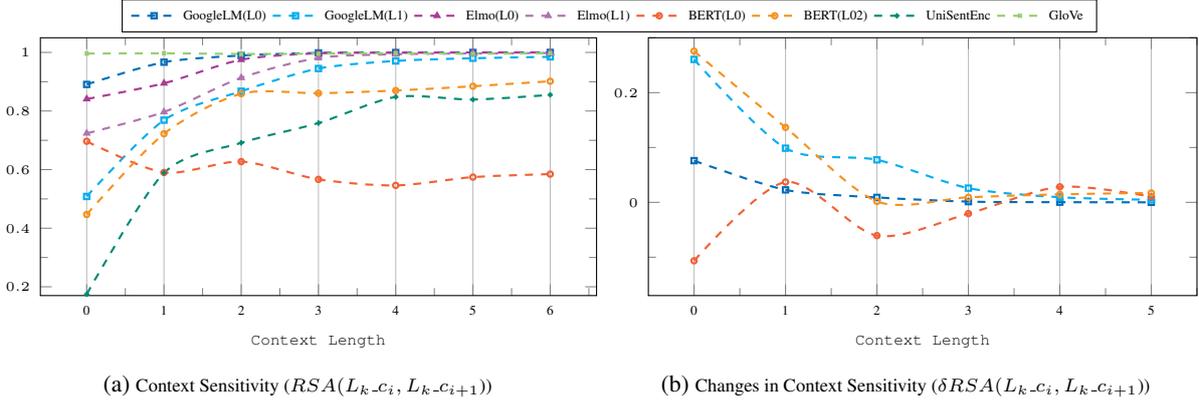


Figure 4: Changes in RSA by increasing context length. (a) Shows how the amount of difference in the representational spaces changes by increasing the context length. (b) Shows for all models that we study, regardless of whether and how much their representations change by increasing context length, the amount of difference becomes almost constant after context length of 3 sentences. Note that in (b), we have scaled the plot and removed some of the models to increase the readability.

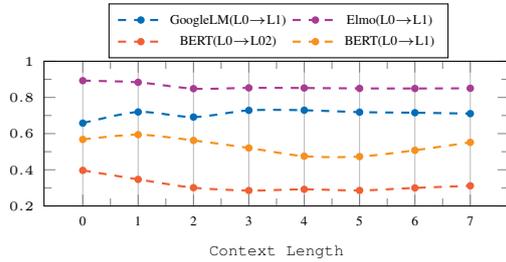


Figure 5: Layer similarities ( $RSA(L_{k-c_i}, L_{k+1-c_i})$ ). Here we show how increasing context length affects the similarity between different layers of the models.)

layers, the representations correspond to more abstract meanings, and the representational space becomes denser than the lower layers.

## 4.2 RSA across Models

In the second step, we study whether the computational models have learned inherently different representational spaces. According to representational similarity scores, among the models that we study, shown in Figure 3, UniSentEnc seems to learn very different representations from ELMO, GoogleLM and BERT. While BERT and UniSentEnc are both Transformer based models, the representational space of BERT is more similar to the representations from ELMO and GoogleLM that are LSTM based models. This can be due to the fact that ELMO, GoogleLM and BERT are trained with language modelling objectives, while UniSentEnc is trained on skip-thought and classification tasks and this could indicate the effect of the training objective on the representational spaces.

## 5 The Relation between the Models and the Activity Patterns in Human Brains

Figure 7 shows the similarity of different computational representation spaces with brain representations, with respect to different amounts of context provided to the models, averaged over all human subjects. Due to the hemodynamic response delay, we expect to see the peak in similarities after about 4s delay. As we can see in Figure 6, the highest RSA for all models is at  $Delay = 4s$ , the ranking of the models based on their similarities with brain representations is the same for all amounts of delay. Interestingly, the performances of these models on the NLP tasks are not correlated with their similarity with the brain representations (but note the overall low correlations). The representations learned by LSTM based models are most similar to the brain data, and for both ELMO and GoogleLM the representations from lower layers,  $L0$ , have higher similarity scores compared to the higher layers,  $L1$ . Interestingly, for UniSentEnc, BERT(L11) and also GoogleLM(L1), increasing the context length, which usually boosts the performance of language encoding models in language understanding tasks (Wang and Cho, 2016), leads to lower similarity with brain representations. It seems that the way these models integrate the context information, pushes the representation further away from the brain representations. This could mean: (1) These models are doing fairly well at encoding the local context, but not at a more global level. Alternatively, (2) The information about the more global aspects of the meaning is not encoded in the brain representations.

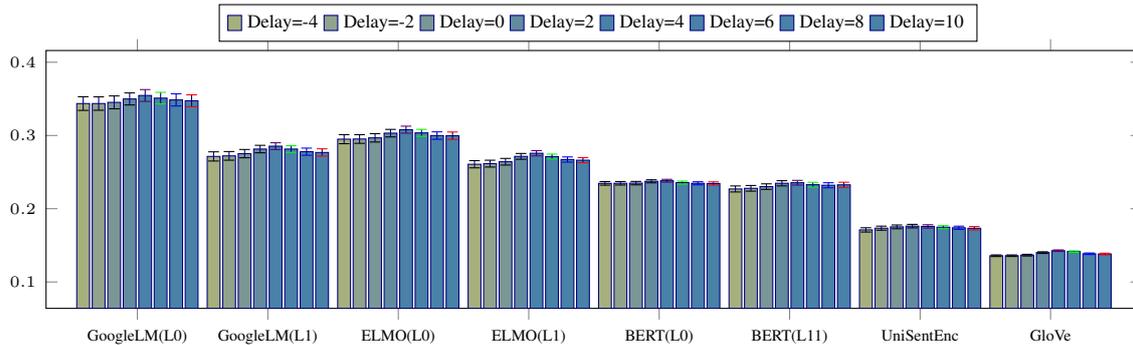


Figure 6: Representational similarity of the models and brains averaged over all subjects and the four blocks at different time delays after the human subjects have read the target words, when the context provided to the models is three sentences. Here the delay is increasing from left to right and the error bars indicate the standard deviation across different blocks.

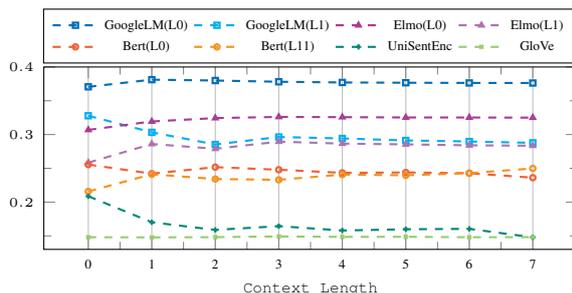


Figure 7: Similarity of the representations from different layers of different models, given different amount of context with brain representations, averaged over all subjects. Note that the average RSA of brains of different human subjects is about 0.55

**Different Segments of the Story** If during training the models are only trained on full sentences, it might be the case that the quality of their representations, when given complete sentences, is significantly better than when provided with incomplete sentences. On the other hand, the representation of sentences in the brain might also be more reliable when the full sentence is read. To take this into account, we look at the similarities of each of the models with brain representations, only at the steps in the story where an end of a sentence token is reached. Figure 8a presents the results. We see that in this case, the similarity of all the models with brain representations increases slightly, but this could be because of the reduced dimensionality of the similarity matrix, and we see that the general patterns stay similar.

In Figure 8b we observe that at the story segments where a name of a character is mentioned, the patterns of similarities change a bit, e.g. the last layer of BERT is less similar to the brain representations compared the first layer of BERT, when an intermediate amount of context is provided to the model. This finding is difficult to interpret, but

warrants further research.

**Different Regions of the Brain** We looked at the similarity scores of the computational representations with the representations at different regions of the brain. This is illustrated in Figure 9 for subject 4 as an example. We observe that the patterns of RSA of different models are very similar across different brain regions, i.e. the scores scale for all regions almost similarly across different models. Despite the low correlations between the models and the brain activation, we find that all the models are consistently best aligned with the regions in the Left Anterior Temporal Lobe (LATL). This region is known for semantic and sometimes syntactic processing of language (Westerlund and Pykkänen, 2014; Bemis and Pykkänen, 2011; Leffel et al., 2014). We also find some correlation with the Left Parietal Lobe, which is not known to be responsible for language processing. We also computed the average RSA between different brain regions for the eight subjects, both within and across subjects, and find that the different regions of a single brain are more similar ( $RSA = 0.4$ ) than the same regions of different brains ( $RSA = 0.12$ ). These are counter-intuitive findings that warrant further investigation. If brain functions involved in story comprehension are spatially localized and brains are organized similarly across individuals, we would expect the same regions from different subjects to be more similar than different regions from the same subject.

**Predictive Approach** Besides, RSA, we can use a predictive approach to see which regions of the brain are more predictable, given the representations from a computational model. In the predictive approach, we train a linear regression model

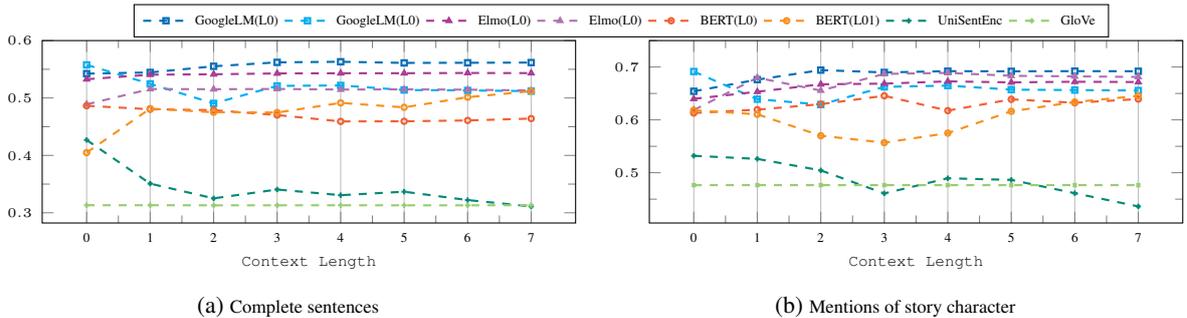


Figure 8: Similarity of the computational representations with brain representations at different segments of the story.

to predict the brain activity patterns at different steps of the story. This way, we can obtain more fine-grained insights into which parts of the model contribute more to which regions in the brain.

In Figure 10, we show the results of using representations obtained from GoogleLM( $L0$ ) to predict brain activity patterns of different subjects. Similar to the results we obtained from RSA, the effect of hemodynamic response delay is clearly visible here. One of the difficulties of employing a predictive approach is to train a regression model for such high dimensions and with so little data. Hence, if the performance of the prediction is low, it is hard to tell if it is because we are not able to train a good regression model or because there is no correlation between the two models. To overcome this challenge, one solution could be to first use RSA to reduce the search space and then employ predictive modelling to gain more fine-grained insights. We postpone further analysis with the predictive approach to future studies.

## 6 Discussion and Conclusion

In this paper, we employ a representational similarity metric to compare the representations from the language encoding models with the brain activity patterns, i.e. measure the alignment between the brain activation patterns and activations of the internal state of the models. The main advantage of RSA is that it treats both the brain and the model as a blackbox; it does not need to know how brains or models represent objects, words or sentences, but only how similar representations are to each other. For  $N$  stimuli considered, the analysis only compares  $\frac{1}{2}N(N-1)$  pairs of pairwise similarities (assuming similarities are symmetric), regardless of the dimensionality of two representational spaces. This bottleneck brings many advantages including computational efficiency, reuse of the similarity matrices in multiple comparisons, and

not having to worry about how to map representations of very different nature to each other. It also brings important limitations and inevitable information loss, e.g. standard RSA, assumes all features of the representational spaces to have equal contributions.

One of our contributions in this paper is the introduction of ReStA, which uses RSA to measure the stability of the representations from the models when an input condition such as context length is changed. Comparing the representational similarity of different layers of different models, we find that both architectural differences and different training objectives have a noticeable impact on the representations learned by the models and the way they change under different conditions. We see a clear difference in the sensitivity to context size between  $L0$  and  $L1$  in the LSTM based models. This means, in line with results from previous work using different methods (e.g., Giulianelli et al., 2018), that the  $L1$  component integrates information over time steps while  $L0$  does not.

Using brain data to evaluate the representations learned at different layers of each of the language encoding models, we find that layers of the LSTM based models achieve higher similarity score with brain data compared to single word representation models like GloVe and the Transformer based models. This observation could show that the learning biases of the LSTM based language models are closer to what happens in the human brain. Zooming into the results, we see that while changing the conditions of the inputs to the models has a significant impact on the representations they compute and their performance on NLP tasks (Khandelwal et al., 2018), these changes do not get reflected in their alignment with the brain representations.

Finally, evaluating computational models of language processing with brain imaging data for

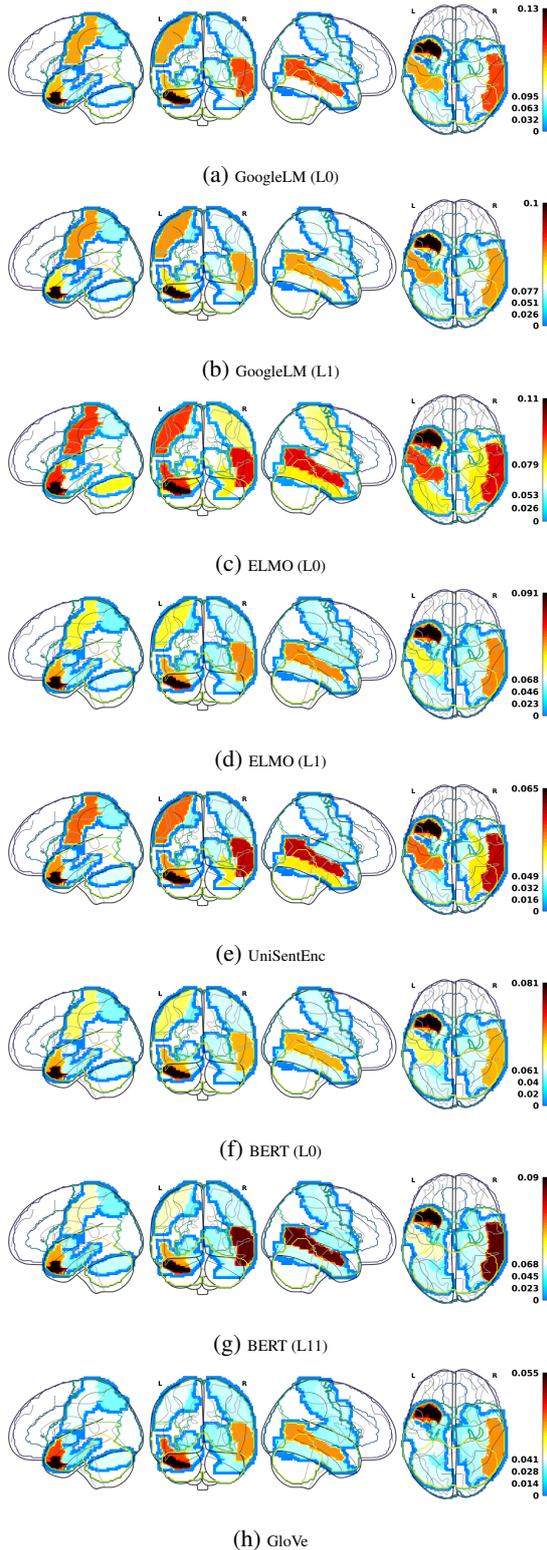


Figure 9: RSA of representations learned at different layers of different models with representations at different regions of Subject4’s brain which is chosen randomly (the code accompanying this paper can be used to generate the plots for the other subjects). In order to emphasize the difference if the similarity of each model with different brain regions, the color bar is scaled independently for each model. The darkest region for all models is the Left Anterior Temporal Lobe.

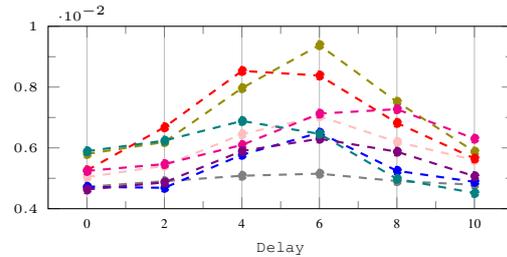


Figure 10: Predictive power of representations learned by Google LM (L0.Cinf) for brain representations in terms of explained variance (each subject in a different color).

a task such as “story reading” is hard, because of the inherent issues in the brain data and also the complexity of the task (Beinborn et al., 2019). Both the RSA framework and the predictive modelling approach make it possible to make a bridge between these black boxes, neural network models for language processing on the one hand and the human brain on the other. And while each of these approaches has its benefits and limitations, they might provide us with complementary information. Hence, it is invaluable to look at both of them.

In our experiments, we observe more similarities between representations learned by some architectures and brain representations. However, caution is required when interpreting these results, as the representational similarity between all models and the brain images remains very low. We plan to perform further analysis on various (bigger) datasets to get a better interpretation of what is happening in both the brain and these computational models.

## 7 Acknowledgement

We thank Dieuwke Hupkes, Arnold Kochari, the Language in Interaction BQ1 team, and the anonymous reviewers for useful comments on the research described here and earlier versions of this paper. The work presented here was funded by the Netherlands Organization for Scientific Research (NWO), through a Gravitation Grant 024.001.006 to the Language in Interaction Consortium.

## References

- Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem Zuidema. 2018. [Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity](#). In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 57–66. Association for Computational Linguistics.
- Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. 2014. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14.
- David Alvarez-Melis and Tommi Jaakkola. 2018. [Gromov-Wasserstein alignment of word embedding spaces](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, Brussels, Belgium. Association for Computational Linguistics.
- Lisa Beinborn, Samira Abnar, and Rochelle Choenni. 2019. Robust evaluation of language-brain encoding experiments. *International Journal of Computational Linguistics and Applications*, to appear.
- Douglas K Bemis and Liina Pylkkänen. 2011. Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *Journal of Neuroscience*, 31(8):2801–2814.
- Joachim Bingel, Maria Barrett, and Anders Sjøgaard. 2016. Extracting token-level signals of syntactic processing from fMRI - with an application to PoS induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 747–755. Association for Computational Linguistics.
- Randy L Buckner. 1998. Event-related fMRI and the hemodynamic response. *Human brain mapping*, 6(5-6):373–377.
- Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. [Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1091. Association for Computational Linguistics.
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. [Universal sentence encoder for english](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 169–174.
- Grzegorz Chrupała and Afra Alishahi. 2019. Correlating neural and symbolic representations of language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2019)*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2019)*.
- Dhanush Dharmaretnam and Alona Fyshe. 2018. The emergence of semantics in neural network representations of visual information. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2018)*.
- Alona Fyshe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. 2014. Interpretable semantic vectors from a joint model of brain-and text-based meaning. In *Proceedings of the conference. Association for Computational Linguistics. Meeting (ACL 2014)*, volume 2014, page 489. NIH Public Access.
- Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *1st BlackBoxNLP workshop at Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*.
- Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453.
- Shailee Jain and Alexander G. Huth. 2018. [Incorporating context into language encoding models for fMRI](#). In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems, NIPS’18*, pages 6629–6638, USA. Curran Associates Inc.
- Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *CoRR*.
- Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. [Sharp nearby, fuzzy far away: How neural language models use context](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 284–294.
- Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

- Aarre Laakso and Garrison Cottrell. 2000. Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical psychology*, 13(1):47–76.
- Timothy Leffel, Miriam Lauter, Masha Westerlund, and Liina Pykkänen. 2014. Restrictive vs. non-restrictive composition: a magnetoencephalography study. *Language, cognition and neuroscience*, 29(10):1191–1204.
- Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195.
- Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. [Selecting corpus-semantic models for neuro-linguistic decoding](#). In *Proceedings of the First Joint Conference on Lexical and Computational Semantics, SemEval '12*, pages 114–123, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP 2014)*, pages 1532–1543.
- Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018*, pages 2227–2237.
- Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016. Bridging lstm architecture and the neural dynamics during reading. In *Proceedings of International Joint Conferences on Artificial Intelligence Organization (IJCAI 2016)*.
- J. K. Rowling. 1998. *Harry Potter And the Sorcerer’s Stone*. Arthur A. Levine Books.
- Yu-Ping Ruan, Zhen-Hua Ling, and Yu Hu. 2016. Exploring semantic representation in brain activity using word embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 669–679.
- Anders Søgaard. 2016. [Evaluating word embeddings with fMRI and eye-tracking](#). In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 116–121. Association for Computational Linguistics.
- Liwei Wang, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, Yue Wu, Kun He, and John Hopcroft. 2018. Towards understanding learning representations: To what extent do different neural networks learn the same representation. In *Advances in Neural Information Processing Systems*, pages 9606–9615.
- Tian Wang and Kyunghyun Cho. 2016. [Larger-context language modelling with recurrent neural network](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*.
- Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014a. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *in press*.
- Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom M. Mitchell. 2014b. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP 2014)*.
- Masha Westerlund and Liina Pykkänen. 2014. The role of the left anterior temporal lobe in semantic composition vs. semantic memory. *Neuropsychologia*, 57:59–70.
- Haoyan Xu, Brian Murphy, and Alona Fyshe. 2016. [Brainbench: A brain-image test suite for distributional semantic models](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 2017–2021.

## 8 Supplementary Material

### 8.1 Preprocessing Brain Images

Besides the cognitive process of interest, other factors like the physiological processes in the bodies of the human subjects or technical features of the MRI-machine and scanning environment may influence the fMRI measurements. An important issue is therefore how to preprocess the data to filter out those irrelevant effects adequately.

**Detrending.** We normalise the brain activations in two steps: we scale the activation values by subtracting the per-voxel mean activation. We also experiment with a more elaborate preprocessing procedure, implemented in the `nilearn.signal.clean` Python library. Detrending is a popular strategy in cognitive neuroscience (Abraham et al., 2014), that removes the linear trend, applies a high pass filtering with 0.005 Hz, and standardises the vectors.

**Voxel selection.** To reduce the noise and remove the voxels which their activation is not related to the story reading task, we apply two steps for selecting the voxels. In the first step, we remove all the constant voxels. These are the brain regions in which the activation does not change at all during the scanning experiment. Next, we compare the similarity of different regions of the brain for all eight subjects and select those regions that their activations over the different segments of the story are most similar among the different subjects. To do this, we rank the regions based on the average of the similarity scores and then selected the top 16 regions. After applying this voxel selection strategy, we have approximately 10000 voxels for each subject.

In our experiments, we do not model the spatial dependency of the voxels. Thus, after the preprocessing steps, we flatten the 3D fMRI images into vectors with the size of the total number of the voxels.

## 8.2 Representational Similarity Across Different Layers of Different Models

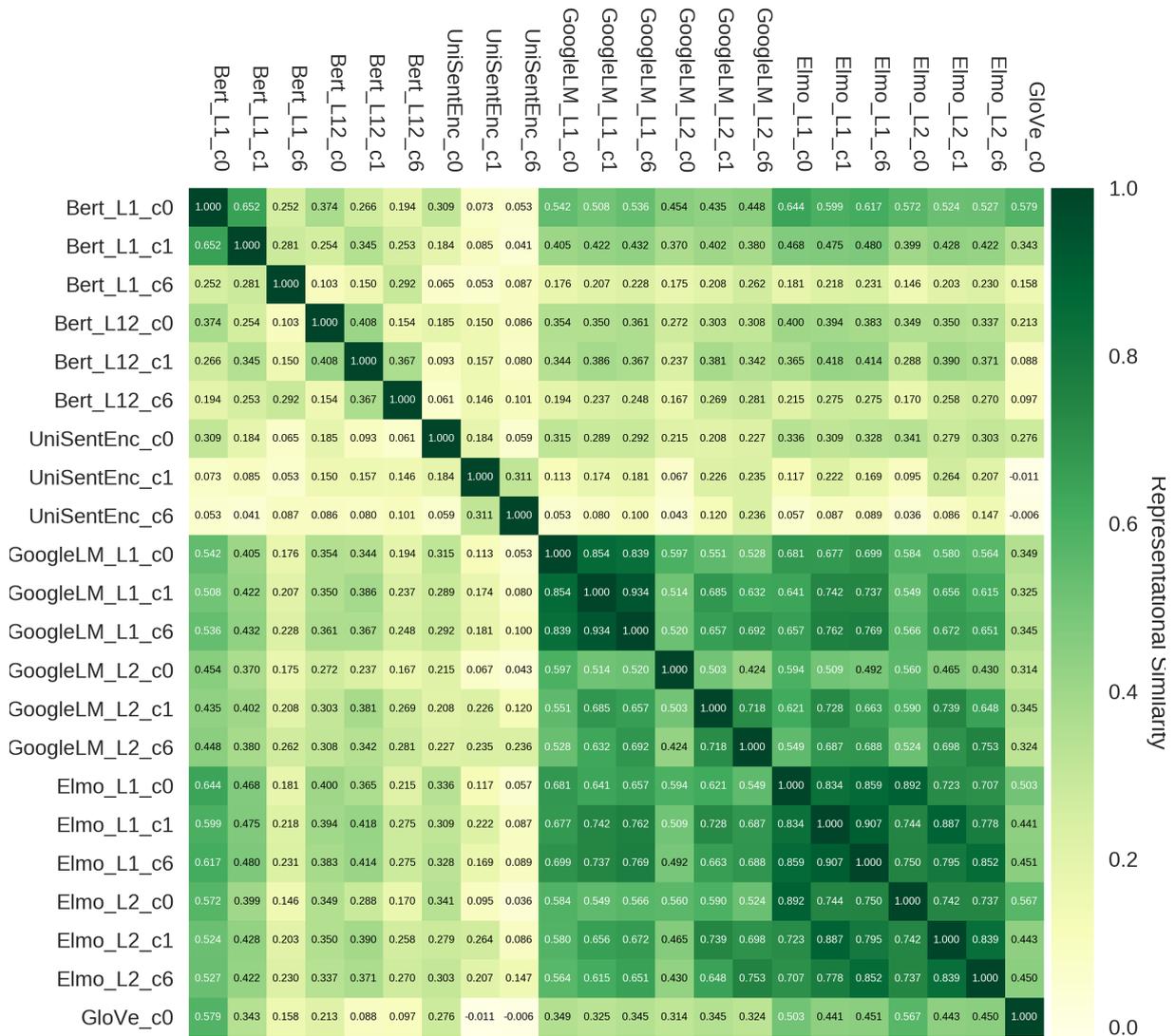


Figure 11: RSA of different layers of different models for different context length. In this plot, for example ELMO\_0\_c1 means representation from layer 1 of ELMO, when the context length is 1 sentences.