What Transformers Might Know About the Physical World: T5 and the Origins of Knowledge

Haohan Shi (hshi420@uchicago.edu)
Division of the Social Sciences, 1126 E 59th St
Chicago, IL 60637

Phillip Wolff (pwolff@emory.edu)
Department of Psychology, 36 Eagle Row
Atlanta, GA 30322

Abstract

Features of the physical world may be acquired from the statistical properties of language. Here we investigate how the Transformer Language Model T5 is able to gain knowledge of the visual world without being able to see or feel. In a series of four studies, we show that T5 possesses an implicit understanding of the relative sizes of animals, their weights, and their shapes, but not their colors, that aligns well with that of humans. As the size of the models was increased from 60M to 11B parameters, we found that the fit to human judgments improved dramatically, suggesting that the difference between humans and these learning systems might ultimately disappear as the parameter sizes grow even larger. The results imply that knowledge of the perceptual world—and much of semantic memory—might be acquired in dis-embodied learning systems using real-time inferential processes.

Keywords: knowledge representation; transformers; T5; natural language processing; semantic memory; perception

Approaches to the Origins of Knowledge

There are essentially three accounts of the origin of knowledge (Anderson, 1989). According to *nativism*, concepts and beliefs are hard-wired into the brain (Samuels, 2002). According to *Empiricism*, knowledge is acquired from the storage of perceptual experience. A third possibility, *Rationalism*, holds that knowledge is acquired from reasoning. The mental processes operations implied by these accounts are clearly not mutually exclusive. Knowledge of the physical world can emerge from perceptual experience constrained by innate processing capabilities and extended through reasoning. However, even when the accounts are combined, the acquisition of knowledge would seem to ultimately depend on direct perceptual experience. Recent evidence suggests that this requirement may not always be necessary.

Kim, Elli, and Bedny (2019) investigated congenitally blind individuals' knowledge of animals' visual properties. Surprisingly, blind individuals' judgments about the relative size, height, shape, and skin texture of a wide range of animals largely agreed with sighted individuals. A possible explanation for the findings is that the bind people relied on sighted people's language about animals to infer their perceptual properties. This possibility predicts that agreement between blind and sighted people should be highest for perceptual properties that are relatively easy to verbalize, like color. As it turns out, it was for this very perceptual property—color— that agreement between the sighted and non-sighted participants was lowest, suggesting that the relatively high agreement between blind and sighted individuals was not simply due to remembering comments of sighted people.

Kim et al. raise another possible explanation for the relatively high agreement between blind and sighted individuals. In particular, blind people might use knowledge of an animal's taxonomic category—which they learned through language—to draw inferences about an animal's perceptual properties. The explanation is supported by findings showing that children can use knowledge of natural kinds to draw inferences about invisible properties, such as animals' insides (Gelman & Wellman, 1991; Keil; 1989). While this hypothesis seems entirely reasonable, it leaves unexplained the difference in performance between properties like shape and color. An inference account still depends on language. If language is used to acquire knowledge of taxonomic categories, then why would it not be used for answering questions about color? Perhaps, as suggested by Kim et al. and claimed by John Locke (1924), blind individuals know more about properties such as shape than color because blind individuals have perceptual experience with shape through the sense of touch. Such an account could help explain how blind individuals could acquire specific features of the perceptual world, such as the shapes, sizes, weights, and textures of certain kinds of objects, but not all. As demonstrated in Kim et al.'s experiment, blind individuals have surprisingly accurate knowledge of the sizes and weights of animals they have never touched, like flamingos and killer whales.

Here we investigate a third possibility we call the *Linguistic Perception Hypothesis*. The hypothesis holds that people can acquire knowledge about distant, unseen, or unfelt features of the world from patterns in language. We test this hypothesis using a remarkable new class of language learning models called transformers (Vaswani et al., 2017). Transformers can acquire knowledge not only syntax (Ganesh, Sagot & Saddah, 2019; Goldberg, 2019; Hewitt & Manning, 2019; Peters et al., 2018; Tenney, et al. 2019) but also of commonsense (Petroni et al., 2019; Da & Kasai, 2019). Perhaps such models may be able to acquire knowledge of the physical world through exposure to

millions of words, explaining how knowledge of the physical world might be acquired without direct perceptual experience.

Transformers

A transformer architecture is a type of statistical learning system that includes multiple layers of feed-forward networks preceded by an attentional mechanism that determines the degree to which the "meaning" of a word in a sequence of words is "colored" by the meaning of other words in the sequence. Training is driven by ambiguity resolution: words in the input sequence are masked, and the network attempts to predict the word or set of words behind the mask. Training is computationally intensive and can take several weeks to complete. Fortunately, previously trained networks can be freely downloaded.

This research used the recently introduced Text-To-Text Transfer Transformer (T5) (Raffel, et al. 2020). T5 is, in a certain sense, an old model because it uses the transformer architecture first proposed in Vaswani et al. (2017). In its original formulation, a transformer has two main parts: a sequence of encoders and decoders. The idea behind an encoder and decoder is made most transparent in language translation problems. The encoders are used to read in and "comprehend" the sentences from one language (e.g., French), while the decoders "generate" text in the second language (e.g., English). One of the many innovations in T5 is that it extends this basic idea to a range of language tasks: translation, grammaticality judgments, sentence similarity assessment, summarization, and question answering. It can also complete a sentence, as shown in Figure 1.

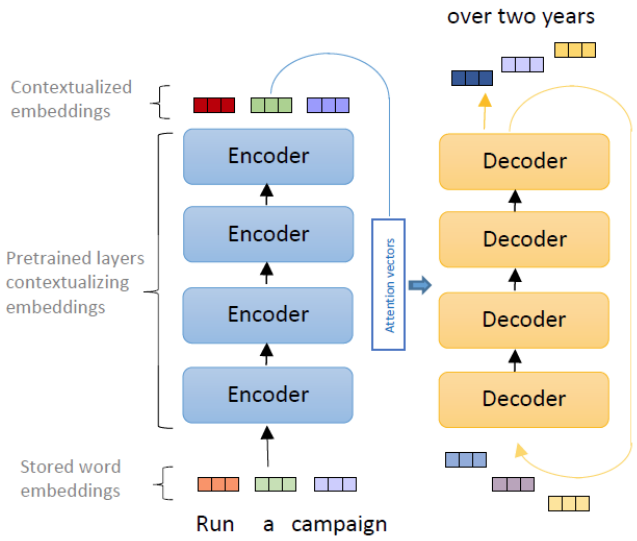

Figure 1: Two main parts of a transformer: Stacks of encoders and decoders that contextualize word embeddings. Words are processed in parallel in the encoder and sequentially in the decoder.

The processing begins with the encoder and involves accessing embeddings for each input word (or word tokens). A word embedding is essentially a point in a semantic space. Over the layers, the embeddings are modified by the encoders depending on the input's other words. For example, in the phrase *Run a campaign*, contextualization allows the meaning of the word *run* to take on the meaning suited to campaigns, as opposed to, for example, miles. The ability to contextualize word embeddings is one of the key ways these models differ from those producing static, context-free embeddings, such as Word2vec (Mikolov, et al., 2013) or Glove (Pennington, Socher, & Manning, 2014). The contextualization process occurs in real-time and is bidirectional, entailing that the process uses words on both sides of the target word. Contextualization depends on self-attentional mechanisms present in each encoder (and decoder). Self-attention reflects the degree to which a word is linked to other words in a sentence. The strength of this connection determines the impact each word will have on a word embedding. Several self-attention weight matrices are learned for each encoder. The self-attention matrices direct the encoder to weigh connections between verbs and their complements, pronouns and their referents, and ambiguous nouns and other nouns. Having multiple self-attention matrices—or heads—allows the encoder to capture the full range of dependencies in a sentence.

The encoders map an input sequence into a continuous abstract representation. The decoders then take that continuous representation and generate words in a step-by-step manner, using the previous step's output as input on the current step. Also, the decoders' output is constrained by attentional vectors formed from the output of the top encoder. The inclusion of both encoders and decoders is one of the ways T5 differs from several other recently transformer-based models, such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), which include only the encoder part of the transformer. T5 model comes in several sizes, as specified in Table 1.

Table 1 shows the five T5 models, along with BERT-Large and RoBERTa-Large for comparison.

Table 1. Model size variants

Model	Parameters	# layers	# heads
T5-Small	60M	6	8
T5-Base	220M	12	12
BERT-Large	336M	24	16
RoBERTa-Large	355M	24	12
T5-Large	770M	24	16
T5-3B	3B	24	32
T5-11B	11B	24	128

T5-small is significantly smaller than BERT-Large and RoBERTa-Large. T5-Large is approximately the same size as these two other networks. T5-11B is quite large, containing 11 billion parameters, and requires approximately 40GB of memory on a GPU. The model can also be run on a CPU on a system having 120GB of RAM.

A second major innovation of T5 is the manner in which it is trained. In BERT and RoBERTa, the target for an individual mask is associated with a single word piece. In T5, the target for a mask can be several words. For example, given the original sentence is *An elephant is larger than a goat*, T5 might be presented with the string *An elephant is <X> a goat*, with the target being several words, namely *<X> larger than <Y>*. Multiword targets are made possible through the use of the stack of decoders.

All versions of T5 were trained on a cleaned version of the common crawl called the Colossal Cleaned Common Crawl (C4). The training corpus is over two times larger than Wikipedia. The largest version of the model, T5-11B, achieved state-of-the-art performance results on the GLUE, SuperGLUE, SQUAD, and benchmarks, which involve natural language processing tasks such as sentiment analysis, question answering, grammaticality judgments, paraphrase detection, selection of plausible causes, and results, textual entailment detection, intended meaning detection, and reading comprehension with commonsense reasoning (Raffel, et al. 2020). As already stated, the reason why T5 succeeds in these tasks is likely due to its ability to acquire knowledge of both language and the world.

The ability to capture at least certain aspects of world knowledge offers a potential test of the *Linguistic Perception Hypothesis*. Suppose we find that a model like T5 is unable to learn visual perceptual properties of the world. In that case, it will provide modest support to the proposal that a learning system devoid of any perceptual senses is unable to capture perceptual properties of the world. On the other hand, if T5 can capture the physical properties, it would suggest that perceptual senses are not a prerequisite for perceptual knowledge and that properties of the physical world can be acquired from language alone, contra Lock.

We investigated the *Linguistic Perception Hypothesis* in a series of studies examining the T5's knowledge of animal size, weight, shape, and color. It is certainly possible that T5 could show limited evidence of such knowledge but far below human judgments. Under these circumstances, it would be good to know if the limit is a fundamental property of these statistical systems or merely a function of the size of the system. To address this question, T5's knowledge was investigated for different model sizes. If the knowledge increases with model size, it would suggest that any limits in its understanding might be a simple matter of the model's capacity rather than an inherent limitation of the architecture.

Study 1: T5's Knowledge of Animal Size

In Study 1 we attempted to replicate judgments of humans about the relative size of animals reported in Kim, Elli, and Bedny (2019). In Kim et al., blind and sighted participants were presented with index cards with animals' names on them. For the bind participants, the names were printed in Braille. Their task was to order the cards from smallest to largest. The relative ordering of the blind and sighted individuals was nearly identical. In the current study, we assessed T5's knowledge of size by generating cross-entropy loss scores to statements such as *A cat is smaller than a bear* and compared these scores to those in which the order of the animals was reversed, e.g., *A bear is smaller than a cat*. To the extent that T5 knows the relative size of animals, then loss scores should be lower for orderings that agree with humans' relative ordering of the sizes. To help establish T5's generalizability, comparative size statements were expressed using three different comparative size adjectives: *smaller*, *larger*, and *bigger*. Also, we investigated the impact of providing a small amount of linguistic context on T5's judgments. In the context condition, the key sentence was preceded with a few sentences introducing the topic of animals. Specifically, T5 was presented with the following sentences: *Animals live outside. They breathe and drink water. They also differ in size*, and then presented with the comparative statement *A cat is smaller than a bear*. The context was included because prior pilot work with another transformer, XLNet (Yang, et al., 2020), suggested that the performance of these networks may be improved when context was provided to "warm them up.".

Methods

Materials The list of animals (n = 15) was the same same as those investigated in (Kim, Elli, & Bedny, 2019): namely and in the order of relative size, *mosquito, bee, butterfly, toad, pigeon, raven, cat, koala, turkey, sheep, donkey, cow, bear, rhino,* and *elephant*.

Procedure Comparative sentences (n= 210) were generated with different pairs of animals and three kinds of degree adjectives (*smaller, larger, bigger*), e.g., *A cat is smaller than a bear*. T5 was credited with understanding the relative size difference of two animals when the cross-entropy loss was lower for the correct ordering. Correct understanding was coded with a 1 and incorrect understanding with a 0. In this and the following analyses, we used the HuggingFace implementation of T5 (Wolf, et al., 2020).

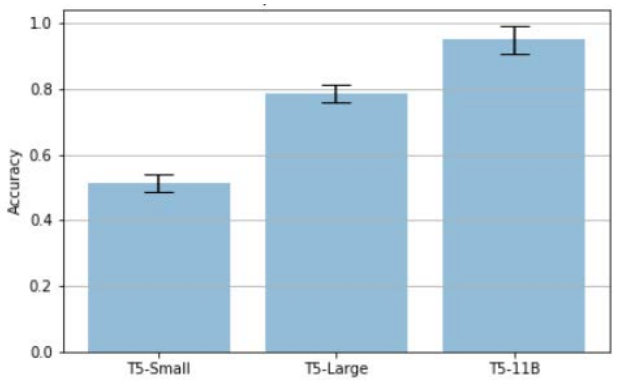

Figure 2: Accuracy of T5 models on animal size with context sentences. Error bars are 95% confidence intervals.

Results and Discussion

The results indicate that T5 has some knowledge of the relative sizes of animals. Treating T5-11B as a classifier, the model was able to predict human responses with 94% accuracy (Precision = 1; Recall/Sensitivity = 0.92; F_1 score = 0.96), which differed from chance by binomial test, p = 9.13e-69. T5-large was also able to predict human judgments, but at a lower level of 78% accuracy (Precision = .86; Recall/Sensitivity = 0.8; F_1 score = 0.83), which differed from chance by binomial test, p = 5.16e-25. T5-small, on the other hand, had an accuracy level of 51% (Precision = .67; Recall/Sensitivity = 0.53; F_1 score = 0.59), which did not differ from chance, p = .652. A main effect of model-type confirmed difference in performance across the models, $F(1,312)=4257$, $p < .0001$. Interesting, accuracy was higher when there was a preceding context than not, $F(1,312)=128$, $p < .0001$, suggesting that such systems may benefit from short priming of the topic. There was also an effect of adjective type, $F(2,624)=7.51$, $p = .001$. However, this effect occurred only in the absence of context, with accuracy being higher for the *small* than *large* and *big*. Overall, we conclude that the results were largely the same across adjectives, especially when there was a preceding context, suggesting that the knowledge is not tied to a particular linguistic expression. Crucially, knowledge of the relative size of animals is very robust in T5-11B, implying that networks with transformer architectures may be able to approach blind individuals' understanding of this dimension of experience.

Study 2: T5's Knowledge of Animal Weight

The design and implementation of Study 2 were analogous to Study 1, except that instead of size, we investigated T5's knowledge of weight, a perceptual property that presumably depends on haptics but is likely to be informed by relative size. Weight was not examined in Kim, Elli, and Bedny (2019), but we could use relative size as an indicator of the correct responses.

Methods

Materials The analyses used the same list of 15 animals as in Study 1.

Procedure We ranked the animals in the list from the lightest to the heaviest based on their size. The texts were generated in the same way as in Study 1 except that we replaced the adjectives with "weighs more" and "weighs less." A context preceding each text, specifically: *Animals live outside. They breathe and drink water. They also differ in weight.* Afterward, T5 was presented with sentences like *A raven weighs more than a bee*. Accuracy scores were determined by comparing the cross-entropy scores, as in Study 1.

Results and Discussion

The results indicated that the larger versions of T5 have world knowledge about the relative weight of animals. The overall mean accuracy for the different models is shown in Figure 3.

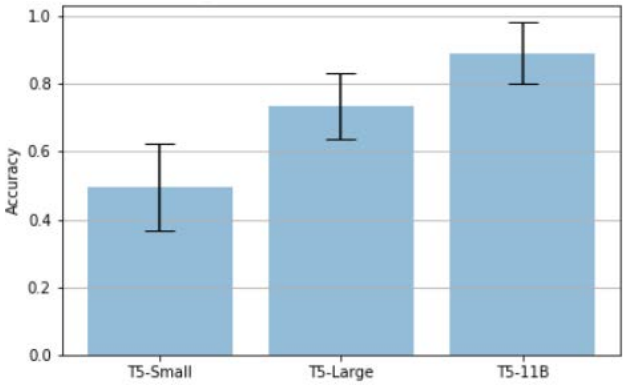

Figure 3: Accuracy of T5 models on animal weight with context sentences. Error bars are 95% confidence intervals.

The results were highly similar to those of size. T5-11b could predict human judgments with 89% accuracy (Precision = .97; Recall/Sensitivity = 0.81; F_1 score = 0.88), which differed from chance by binomial test, p = 9.03e-12. Similarly, T5-large could also predict human judgments, but not quite as well as T5-11b, at 73% accuracy (Precision = .78; Recall/Sensitivity = 0.65; F_1 score = 0.71), a level that differed from chance by binomial test, p = 3.47e-33. The accuracy level of T5-small, 49% (Precision = .50; Recall/Sensitivity = 0.61; F_1 score = 0.54), did not differ from chance by binomial test, p = .945.

Weight represents a type of force. The results raise the possibility that large transformer models like T5-11B may learn visual information about the world and invisible quantities such as forces, which suggests they may have the prerequisites for causal knowledge (Wolff & Shepard, 2013).

Study 3: T5's Knowledge of Animal Shape

In Study 3 we investigated T5's knowledge of animal shape. Unlike size and weight, which can be reduced to a single dimension and referred to by an adjective, shape is inherently multidimensional and more difficult to describe. Instead of focusing on a single dimension, T5 was presented with sentences describing each animal with respect to a range of shape-related adjectives like *long*, *short*, *thin*, and *thick*. For example, T5 evaluated the acceptability of sentences like *A giraffe is long* or *A sloth is thick*. The result was an animal-by-shape-dimension matrix of cross-entropy scores. The dimensionality of the matrix was reduced and submitted to *k*-means clustering. In Kim et al. (2019), people's sorts of animal names were used to place 30 animals into one of 8 categories with respect to shape. We created 8 clusters from the animal-by-shape-dimension matrix and evaluated the degree to which these clusters agreed with those produced by participants in Kim et al..

Methods

Materials The analyses used the list of 30 animals that were used in the animal shape card sorting task (Kim, Elli, & Bedny, 2019). Their names are shown in Figure 4.

Procedure We introduced 13 different adjectives describing animal shape (*long, short, straight, curved, thin, thick, circular, square, vertical, horizontal, rounded, angular, pointed*). The texts were generated in a pattern that contained a noun (animal) and an adjective (shape), e.g., *The shape of a giraffe is vertical*. With 30 animals, 30 texts were generated for each adjective. A context preceded each text: specifically, *Animals live outside. They breathe and drink water. They also differ in shape*. We used T5 to predict the adjectives and took the cross-entropy loss as a raw score for each text. The dimensionality of the resulting cross-entropy matrix was reduced to three dimensions using the IVIS dimensionality reduction framework (Szubert et al., 2019), which does a better job of preserving both local and global structure than linear projection methods such as Principal Components Analysis (PCA) and other non-linear dimensionality reduction methods such as t-SNE (Maaten & Hinton, 2008). The reduced space was partitioned into 8 clusters using the k-means++ algorithm available in scikit-learn (Pedregosa, et al., 2011).

Results and Discussion

The results indicate that T5-11B possesses some knowledge of animal shapes. Agreement between the humand and T5 was assessed by identifying the most common cluster label generated by T5 for each human category, counting the number of times that label occurred within each human category, and dividing by the total number of animals. This measure indicated that human clusters overlapped 70% with clusters generated from T5-11b, 60% with T5-large, and 57% with T5-small. The overlap between T5-11b and humans differed from the chance agreement of .399 (derived from simulation), by binomial test $p = .001$, as did T5-large and T5-small at $p = .021$ and $p = .047$, respectively. A text plot of one of the solutions derived from T5 is shown in Figure 4.

Figure 4: Text plot of animals clustered with respect to shape as determined by T5.

Study 4: T5's Knowledge in Animal Color

In this last study, we investigate T5's knowledge of animal colors. Unlike shape, languages like English make it relatively easy to refer to color. However, animals are not uniform in color, as exemplified by zebras, killer whales, and giraffes. Given the multidimensional nature of animal colors, Kim et al. (2019) measured knowledge of animal colors in terms of eight categories, as they had done with shape. As a consequence, we measured knowledge of animal colors in the same way we did in Study 3. T5 evaluated sentences such as *A gorilla is yellow*, or *A gorilla is black*. The resulting matrix of cross-entropies was reduced in dimensionality and partitioned into eight clusters. Finally, we measured the degree of overlap between the clusters formed by humans in Kim et al. (2019) and those produced from different versions of T5.

Methods

Materials The analyses used the same list of 30 animals as in Study 3.

Procedure We introduced six different adjectives that can be used to describe the color of the animals in the list (*black, white, brown, grey, yellow, blue*). The texts were generated in the same way as in Study 3, except that the adjectives describing shapes were replaced with adjectives describing colors. With 30 animals, 30 texts were generated for each adjective. We used T5 to predict the adjectives and took cross-entropy loss as a raw score for each text.

Figure 5: Text plot of animals clustered with respect to color as determined by T5.

Result and Discussion

T5 appears to possess less knowledge about animal color than animal size, weight, and shape. Agreement between humans and T5-11b was 60%, which differed from chance (.399) by binomial test, $p = .021$, but not as strongly as observed for the other perceptual dimensions. The levels of agreement for T5-large and T5-small were 43% and 33%, neither of which differed from chance by binomial test, $p = .417$ and $p = .293$. A text plot of the solution produced by T5 with respect to color is shown in Figure 5. Several of the clusters are intuitive, such as clusters containing black animals (*bats, panthers, crows,* and *gorillas*) and white animals (*swan, polar bear*), but several other clusters are less clear, such as the cluster containing *hippos, sloths,* and *pigeons*. We conclude that T5 has limited implicit knowledge of animal colors.

General Discussion

In this research, we investigated a model organism's perceptual knowledge with extensive experience with language but no direct or indirect experience with the physical world. Remarkably, this wholly disembodied organism demonstrated high sensitivity to visual and haptic features of the world. The results suggest that the information needed to learn perceptual features of animals is implicitly present in the statistics of language. The results imply that the model's success is not due to the simple retrieval of statements about animals' perceptual characteristics. If performance was merely a matter of memory retrieval, the systems' awareness of the colors of animals should have been much stronger than was observed. Interestingly, the results largely mirror the results found in Kim, Elli, and Bedny (2019), with the agreement between sighted and blind participants highest for size, shape, and then relatively poor for color.

The results provide support for the Linguistic Perception Hypothesis, which holds that people can acquire knowledge about distant, unseen, and unfelt features of the world through an analysis of the statistical properties of language. The remote linguistic perception might contribute significantly to people's semantic memory. Semantic memory concerns our knowledge about facts, ideas, beliefs, and intuitions (McRae & Jones, 2013). By definition, semantic memory is not episodic: the origins of this knowledge cannot be connected to a particular experience or event. The origins of such knowledge remain largely unknown. Abstraction over main instances could, perhaps, account for some of our lack of knowledge of the origins of semantic knowledge (e.g., that roses are red). However, the knowledge in question is often unlikely to have been acquired from even a single episode (e.g., knowing that string can be thrown further if it is rolled up in a ball than left straight). The results of this study begin to show how the origins of certain kinds of knowledge can be derived from wholly implicit bits of knowledge, most likely beneath our conscious awareness.

References

Anderson, J. R. (1989). A theory of the origins of human knowledge. *Artificial Intelligence, 40,* 313-351.

Da, J., & Kasai, J. (2019). Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing, pp 1-12.

Gelman. S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the nonobvious. Cognition, 38, 213-244.

Ganesh, J., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

Goldberg, Y. (2019). Assessing BERT's syntactic abilities. Computing Research Repository, arXiv:1901.05287.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics.

Keil, F. C. (1989). Concepts, kinds, and conceptual development. MIT Press: Cambridge, MA.

Kim, J. S., Elli, G. V., & Bedny, M. (2019). Knowledge of animal appearance among sighted and blind adults. *Proceedings of the National Academy of Sciences, 116*(23), 11213-11222.

Locke, J. (1924). An essay concerning human understanding. Nature 114, 462.

McRae, K., Jones, M. (2013). Semantic memory. In D. Reisberg (ed.), The Oxford Handbook of Cognitive Psychology. New York, NY: Oxford University Press.

Mikolov, T., Chen, K., Corrado, D., & Dean, J. (2013). Efficient estimation of word representations in vector space. In International Conference on Learning Representations (ICLR) 2013. Retrieved from https://sites.google.com/site/representationlearning2013/workshop-proceedings

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830.

Peters, M. E., Neuman, M., Zettlemoyer, L., & Yih, W. (2018). Dissecting contextual word embeddings: Architecture and representation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 – November 4, 1499-1509.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on EMNLP (pp. 1532–1543). New York, NY: Association for Computational Linguistics.

Petroni, F., Rocktaschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language, pages 2463 – 2473.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv:1910.10683[cs.LG]

Samuels, R. (2002). Nativism in cognitive science, *Mind & Language, 17*, 233-265.

Szubert, B., J. E. Cole, C. Monaco and I. Drozdov (2019). Structure-preserving visualisation of high dimensional single-cell datasets. Sci Rep 9(1): 8914.

Tenny, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S., Das, D., & Pavlick, E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. In International Conference on Learning Representations.

Vaswani, A., Shazeer, N., Parmar, N., Uszhoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). arXiv:1706.03762v5 [cs.CL]

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579-2605.

Wolf, T., et al. (2020). HuggingFace's transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771v5.

Wolff, P., & Shepard, J. (2013). Causation, touch, and the perception of force. P*sychology of Learning and Motivation, 58*, 167 – 202.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q.V. (2020). XLNet: Generalized Autoregressive Pretraining for Language Understanding. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.