You can't learn what's not there: self supervised learning and the poverty of the stimulus

Csaba Veres[1] and Jennifer Sampson[2]

[1] University of Bergen, Norway csaba.veres@uib.no
[2] Equinor, London, U.K. jensam@equinor.com

Abstract. *Diathesis alternation* describes the property of language that individual verbs can be used in different subcategorization frames. However, seemingly similar verbs such as *drizzle* and *spray* can behave differently in terms of the alternations they can participate in (drizzle/spray water on the plant; *drizzle/spray the plant with water). By hypothesis, primary linguistic data is not sufficient to learn which verbs alternate and which do not. We tested two state-of-the-art machine learning models trained by self supervision, and found little evidence that they could learn the correct pattern of acceptability judgement in the locative alternation. This is consistent with a poverty of stimulus argument that primary linguistic data does not provide sufficient information to learn aspects of linguistic knowledge. The finding has important consequences for machine learning models trained by self supervision, since they depend on the evidence present in the raw training input.

1 Introduction

Language models which employ self supervised learning represent a major breakthrough for machine learning. These algorithms construct a large statistical model of language by performing tasks such as masked language modeling on a vast quantity of unlabelled text [7, 22]. The idea that linguistic knowledge can be acquired entirely from primary linguistic data (PLD) has been questioned for many years, with the poverty of stimulus (POS) argument. The term itself was introduced in [5], but has been part of Chomsky's arguments since at least 1965 [4]. The main problem raised by the POS argument is that the PLD does not contain the kinds of sentences that would help learners falsify (at least some of) the incorrect hypotheses about the grammar of their language [3, 6]. The consequences for machine learning are the same: if POS is correct, self supervised models will not have sufficient data for a complete understanding of linguistic structure.

Warstadt [23] developed the Corpus of Linguistic Acceptability (CoLA) to test the POS argument with machine learning models. They argued that if grammatical acceptability judgements can be learned to human level with no in-built language specific principles, then this argues against POS. Their results showed that state-of-the-art recurrent neural network models could not achieve human level performance, suggesting that grammatical knowledge can not be learned in its entirety from linguistic input alone. A similar conclusion was reached a year later with a BiLSTM model using the GLUE benchmark, which included CoLA [20]. However, with the advent of the transformer architecture [18] and ensuing implementations, performance improved

dramatically and the subsequent iteration of the benchmark, SuperGLUE, did not include the CoLA suite citing better than human performance by the XLNet-Large architecture [21, 24]

Does this result mark the end of the POS hypothesis? Veres and Sandblast [19] argue that it does not, because the CoLA does not pose a sufficiently strong test of the hypothesis. The corpus includes a wide variety of grammatical violations, but no attempt is made to show that any of them are potentially unlearnable in the suggested way. In fact [3] argue that "responsible nativists" try to account for acquired linguistic knowledge with the minimum language specific component of learning, so CoLA may not have focused on the critical test cases. Veres and Sandblast [19] propose a new benchmark which is composed of grammatical violations related to Baker's paradox [1], which are not learnable from linguistic data alone [15]. Instead they rely on knowledge of linking rules between lexico-semantic features and their syntactic expressions. Their preliminary results supported the POS argument, that learning requires knowledge about language which is not directly discernible from the primary linguistic data. This paper reports additional experiments to provide stronger evidence that self supervised models are not able to learn certain aspects of linguistic knowledge.

2 Learnability and Semantics

The learning problem involves verb subcategorisation frames and the possibilities for alternative frames involving the same verb, also called diathesis alternations. For example the verb *load* can appear in the following construction (examples taken from [16]).

(1) Hal is loading hay into the wagon.

In sentence (1) the grammatical subject (Hal) of the verb is the loader, the object is the contents being moved (the hay), and the object of the preposition *into* expresses the container into which the hay is being moved (the wagon). This is called the content-locative construction, or V-locatum-location, because the focus of the sentence is the locatum (hay). The same meaning can be expressed by sentence (2) where the object of the verb is now the container, changing the focus of the sentence. This is called the container-locative construction, or V-location-locatum.

(2) Hal loaded the wagon with hay.

A possible generalisation for the learner is that verbs appearing in content-locative constructions can also appear in container-locative constructions.

However the generalisation does not hold, as there are many other verbs which result in unacceptable sentences. Examples (3) and (4) show that *pour* does not accept the container-locative, and *fill* does not allow content-locative. There does not seem to be a clear way to distinguish the verbs that do, and the ones that don't allow the generalisation. In these examples *pour*, *fill*, and *load* are all verbs which describe someone moving something somewhere.

(3) a. Amy poured water into the glass.
b. *Amy poured the glass with water.

(4) a. *Bobby filled water into the glass.
b. Bobby filled the glass with water.

The fact that adult speakers of English can make these distinctions is a learnability paradox. Four conditions lead to the paradox: (a) language speakers generalise from observations, (b) they avoid some possible generalisations, (c) they are not corrected for erroneous generalisations, (d) there is no systematic difference between verbs that allow generalisation and those which do not. Clearly at least one of these statements cannot be correct.

Pinker argues that the fourth condition is where the solution to the paradox lies, and in fact systematic differences do exist. However the differences are described in terms of nonobvious descriptions of semantic structure in the form of broad- and narrow- range semantic rules. Broad range rules provide necessary conditions, and the narrow range rules provide sufficient conditions [15, 16].

A necessary condition for a verb to participate in the locative alternation is that it specifies both a type of motion and an end state. For example when someone *smears grease onto a bearing*, or *smears a bearing with grease*, then we know the kind of activity the person is engaged in and how the bearing will end up looking. On the other hand, the non alternating verb *fill* specifies only an end state. If I *fill the bottle with water* (not **fill water into the bottle*) then it is not clear how I filled it; what is clear is that the bottle is full. Conversely, if I *pour water into the bottle* (not **pour the bottle with water*) then the action I perform is more clear, but the end state of the bottle, less so. Note also that this contrast explains the subtle shift in meaning observed with alternating verbs. For example *The farmer loaded the cart with apples* suggests the cart is full, whereas *The farmer loaded apples into the cart* does not [11]. The V-location-locatum diathesis carries the semantic interpretation involving an end state.

The necessary conditions in themselves do not capture the full range of grammatical facts. For example, why is **I dripped the floor with water* not acceptable? In what way does it not entail an end state where the floor is covered with drops of water? If it did, then the construction should be grammatical under the present hypothesis. To explain these facts [15] further proposes a set of narrow range rules which provide fine-grained criteria which are sufficient to license the alternation for a given verb. The rules involve a range of language specific semantic properties which constrain the interpretation of concepts with respect to their expression, and particular argument structures are licensed by these semantic properties.

The broad and narrow range rules together are in fact *rules of construal* which are needed because cognition is too flexible to determine which syntactic device is most suited in expressing the communicative intent of the message. For example if someone in the real world *pours water into a glass*, are they affecting the water by causing it to move from one location to another (V-locatum-location) or are they affecting the glass by causing it to be less empty (V-location-locatum)? The broad range rule makes this determination for us. As far as language is concerned, *pour* is a verb that describes an action performed on the locatum rather than the state of the location. This principle is meant quite generally, such that the role of language is to funnel an infinitely flexible cognition into a more rigid and fixed system suitable for expression.

The narrow range rules provide specific constraints to determine the interpretation of narrow conflation classes. Returning to the example of *drip*, why does *I sprayed the plant with water* entail and end state but **I dripped the plant with water* does not? By hypothesis, the fine grained semantic description of *drip* verbs (which also includes *dribble, drizzle, dump, pour, ...*) is something like "a mass is enabled to move via the force of gravity." On the other hand *spray* verbs (which also includes *splash, splatter, sprinkle, squirt, ...*) are verbs where "force is imparted to a mass, causing ballistic motion in a specified spatial distribution along a trajectory" ([15], p.126). It is therefore a distinction between **enabling** and **causing** the motion of a mass, where the causation implies some element of control over the end state. "Dripping" does not entail an end state because we have no direct control over the end state.

This proposal is called the Grammatically Relevant Subsystem (GRS) approach, because the classification of verbs with respect to their subcategorization options is a matter for the specialised semantics embodied in the narrow range rules, rather than some more general classification problem. The semantic features are a part of the conceptual - linguistic linking system, and can not directly be inferred from the general properties of the observed linguistic input. Diathesis alternations are controlled by lexico-semantic facts that are not directly observable from the strings of the language, cannot be inferred from the statistical distribution of those strings, and should not be learned by systems that depend entirely on such distributions.

3 Related work

There is a growing body of research whose goal is to investigate the nature of knowledge acquired by machine learning models, beyond the commonly used NLP benchmark results. Many of these studies draw similar conclusions about the limitations of machine learning.

Bender et. al. [2] take a somewhat general view of the limits of machine learning, arguing that text corpora can only provide linguistic *form*, which is not sufficient to capture *meaning*, or more precisely, *communicative intent*. While this is not strictly speaking a POS argument for acquisition of knowledge *about* language, it is a reminder that exposure to written sentences is not sufficient to model the use of language for communication.

Kassner and Schütze [10] test for more specific aspects of linguistic knowledge. They investigate pretrained language models (PLMs) for evidence of specific factual knowledge. They conclude that PLMs have difficulty with learning about negation. Given the statement "The theory of relativity was *not* developed by [MASK]" they are just as likely to predict "Einstein" as if the statement was "The theory of relativity was developed by [MASK]." In addition, PLMs can be misled in a novel technique called *mispriming*, inspired by psycholinguistic studies, where a question framed as "Talk? Birds can [MASK]", can prompt the erroneous response "Birds can talk."

In another set of experiments designed to test linguistic capacities rather than performance on popular NLP tasks, [8] show that BERT [7] lacks knowledge of negation and it struggles with some difficult inference and role-based event prediction tasks.

Turning now to the question of learnability from PLD, [14] propose a hierarchical Bayesian framework which is able to model many aspects of learning verb constructions, including those involved in Baker's paradox. They showed that diathesis alternations could be predicted by distributional evidence alone. They used a hierarchical Bayesian model which regarded deviations from expected frequencies as a form of negative evidence for resolving Baker's paradox. The model uses a hierarchy of inductive constraints, or *overhypotheses*, based on the distributional evidence. The model learns the distribution of verb constructions across all verbs in a language, as well as the degree to which any individual verb tends to be alternating or non-alternating. This way it can learn prior probabilities that can be used to predict the alternation patterns of verbs in the corpus. One limitation of the study is that the model is built to detect the non occurrence of just the right sentences, that is, the lack of the ungrammatical alternation. But this is unnatural because it assumes that, of the potentially infinite non occurring sentences containing a particular verb, language learners are tuned to focus on just the right ones.

The critical point which emerges from prior work is that statistical models can potentially learn the relevant grammatical generalisations from PLD, but only if they include built in assumptions about expected distributions in text. Transformer models make no prior assumptions and therefore it is important to determine if they can learn the relevant generalisations from the input data alone.

4 Dataset

The preliminary studies of [19] showed a mixed set of results for the 24 different types of diathesis alternations selected from [11]. Amongst the poorest performers were the as-, locative-, reciprocal-, and fulfilling- alternations. The locative alternation we have been describing is one of the best understood, and it was used in the sentences in this study.

We constructed a set of 274 sentences in total, 137 alternating and 137 non alternating. The 137 alternating sentences were all grammatical, but half of the non alternating sentences were ungrammatical. Table 1 shows the conditions with a sample sentence in each.

5 Results

5.1 Acceptability experiments

We use the Hugging Face implementation of BERT (Bidirectional Encoder Representations from Transformers) [7]. The pre-trained model has been trained on vast amounts of general language data and can be fine-tuned by further training on downstream NLP tasks such as named entity recognition, classification, question answering, and acceptability judgement.

BERT is distinguished from other transformer-based networks by the input encoding it uses while training and the problems it was trained to solve during training: masked language modelling (MLM) and next sentence prediction (NSP).

	V locatum location	**V location locatum**
Alternating	The farmer had to load all the apples into the cart.	The farmer had to load the cart with all the apples.
With only	*The final step is to coat chocolate on the cake.	The final step is to coat the cake with chocolate.
Into/Onto/On only	Carla poured lemonade into the pitcher.	*Carla poured the pitcher with lemonade.

Table 1. Example sentences from the six different types in the experiment. The asterisk (*) denotes ungrammatical strings. The treatment conditions are named for the verb frame in which the example sentences are judged acceptable.

Since acceptability judgement is a form of classification, we used BERTForSequenceClassification classifier using bert-base pretrained model, fine tuned on the CoLa dataset. The validation accuracy was 0.70 with validation loss = 0.61.

The common metric for acceptability judgement is the Matthews correlation coefficient which measures the agreement between classification scores and human judgement. The measure is thought to be particularly meaningful because it takes into account true and false positives and negatives, unlike the F measure typically used in many other tasks [13].

Table 2 shows the Matthews correlation coefficient for the two sentence types, compared to the experimenter's judgement of grammatical acceptability.

	Matthews correlation
With only	0.27
Into/Onto only	0.05

Table 2. Matthews correlation coefficient for acceptability judgement obtained with BERT.

The results show almost no sensitivity to grammatical acceptability for Into/onto only sentences that are ungrammatical in the V location locatum construction. Table 3 shows the reason for this is low accuracy for ungrammatical constructions (e.g. **Carla poured the pitcher with lemonade*). On the other hand there is a weak positive correlation for with-only sentences, but accuracy for unacceptable sentences (e.g. **The final step is to coat chocolate onto the cake*) is still low.

Since there was a weak correlation in one condition We repeated the analysis using RoBERTa, a newer model based on BERT with a robustly optimized pretraining approach [12] which uses a much larger training set, and modifies the training regime by dropping the next sentence prediction task.

	Grammatical	Ungrammatical
With only	1.0	0.16
Into/Onto	0.9	0.14

Table 3. Accuracy of acceptability judgement obtained with BERT.

We used the RobertaForSequenceClassification classifier from Hugging Face based on the roberta-base pretrained model. The classifier was fine tuned on the CoLA task as before, obtaining a higher validation accuracy = 0.86 and loss = 0.43. We submitted our results to Kaggle for test validation and achieved a result of 0.62 [3]. Compare this to 0.678 for the Facebook implementation on gluebenchmark.com, where the current leader for this task is StructBERT from Alibaba with a score of 0.753 [22].

Table 3 shows the Matthews correlations. Surprisingly the with-only condition shows a slightly worse performance, but now the Into/Onto condition shows a stronger, moderate correlation.

	Matthews correlation
With only	0.17
Into/Onto only	0.4

Table 4. Matthews correlation coefficient for acceptability judgement obtained with RoBERTa.

The increased correlation in the Into/Onto condition is due to increased accuracy in the ungrammatical condition, as seen in table 5. The with-only accuracy is still low, as expected from the correlation score.

	Grammatical	Ungrammatical
With only	1.0	0.09
Into/Onto only	0.97	0.45

Table 5. Accuracy of acceptability judgement obtained with RoBERTa.

5.2 Embeddings

It is generally believed that embeddings capture aspects of word semantics, though the nature of the semantic properties is not well understood [17]. If verb alternations depend on subtle semantic distinctions, then the word embeddings should contain elements of such semantics.

Figure 1 shows a 2-dimensional principal components projection of the vector embeddings for the verbs in the experimental conditions. The with-only verbs are shown

[3] https://www.kaggle.com/c/cola-in-domain-open-evaluation/leaderboard

with a plus sign ”+” in the figure, and the Into/Onto-only verbs with the filled circles. Each verb appears more than once because embeddings are contextualised, and a given verb has a slightly different vector representation in different sentences. There is a very pronounced separation between the two sets of verbs, suggesting that something of the semantic difference was captured in the embedding space.

Closer inspection of the verbs, however, reveal a possible confound. It appears that the Into/Onto-only verbs in the cluster on the right of figure 1 appear with various liquids, while the with-only ones on the left can not. So, for example, pour/dribble/slop/slosh are actions one can perform with water but bandage/bind/decorate/dirty/bombard are not. This is just distributional semantics learned from the context in which words appear, where the distributional hypothesis [9] implies that words which cluster together are words which can be used interchangeably in relevant contexts. Sahlgren [17] calls this a *paradigmatic relation*. To control for the confound, we are currently collecting Into/onto-only sentences which do not include liquids. If we are correct then this should reduce classification accuracy to 0.

There is an alternative test we can perform with the current sentences, which is to see if verbs that allow alternation cluster differently to ones which do not. Figure 2. shows these verbs as upside down triangles. We can see that the alternating verbs are spread throughout the non alternating ones. This is important for Pinkers’s hypothesis since he writes: ”The exact differentiation of the nonalternating classes from one another is not crucial as long as the criteria distinguishing them from the alternating classes are clear” ([15], p.237). Clealry they are not distinguished in the verb embeddings. Further, alternating verbs which co-occur with liquids overlap with the non alternating verbs that co-occur with liquids. For example squirt/sprinkle/spray appear next to dribble/pour/spew. This strengthens the hypothesis that the semantics captured in RoBERTa is limited to distributional co-occurrence.

6 Discussion

We began by considering Baker’s paradox which concerns problems with the learning of syntactic diathesis alternations from primary linguistic data. The proposed solution involved a number of lexical semantic features that constrain the syntactic behaviour of individual verbs. We then asked if these features could be learned by modern machine learning architectures trained on massive text corpora. The results show that neither BERT nor RoBERTa were able to reliably differentiate the verbs on semantic grounds.

However, RoBERTa achieved moderate performance for recognizing the acceptability of Into/Onto-only verbs, and embeddings from both systems showed an appreciable separation between the two non alternating verb classes. We argued that this result was an artefact because the Into/onto-only verbs in our test sentences tended to have liquids as objects while the with-only verbs did not. We should then be able to abolish the model’s classification accuracy with a new test that included Into/onto-only verbs with non liquid objects, for example *”He coiled the chain around the pole” / ”*He coiled the pole with the chain.”*

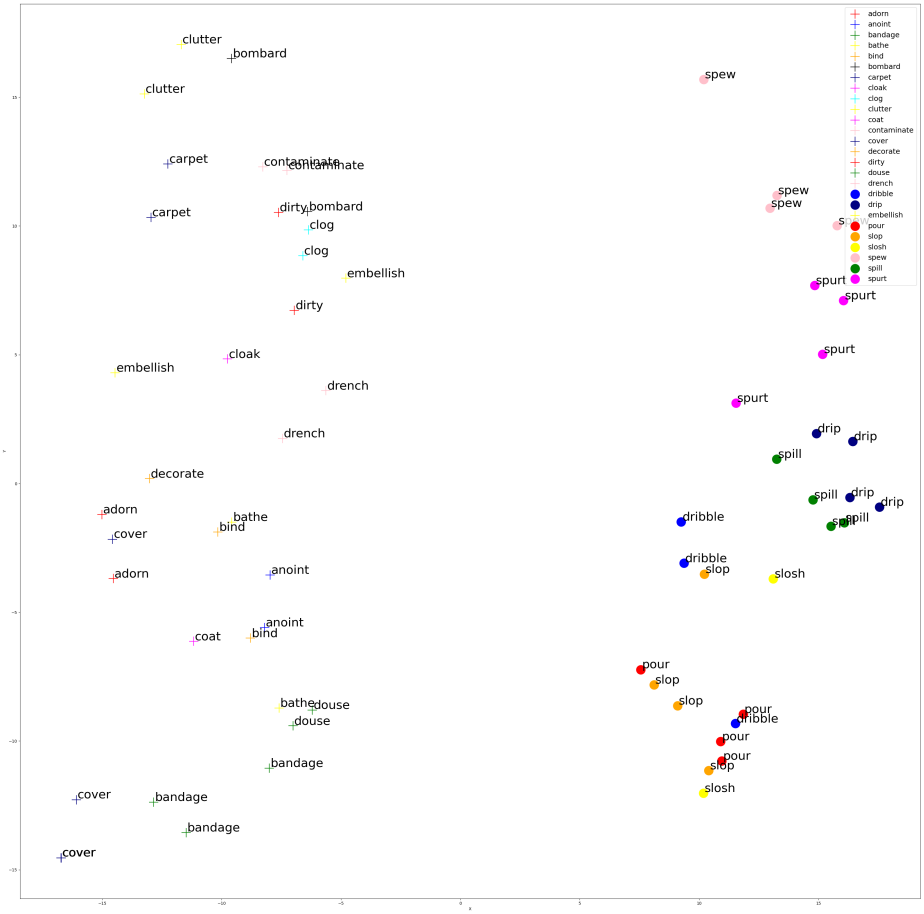

Fig. 1. 2-dimensional PCA projection of ”Into/Onto verbs” (filled circles) and ”With” verbs (”+”) signs

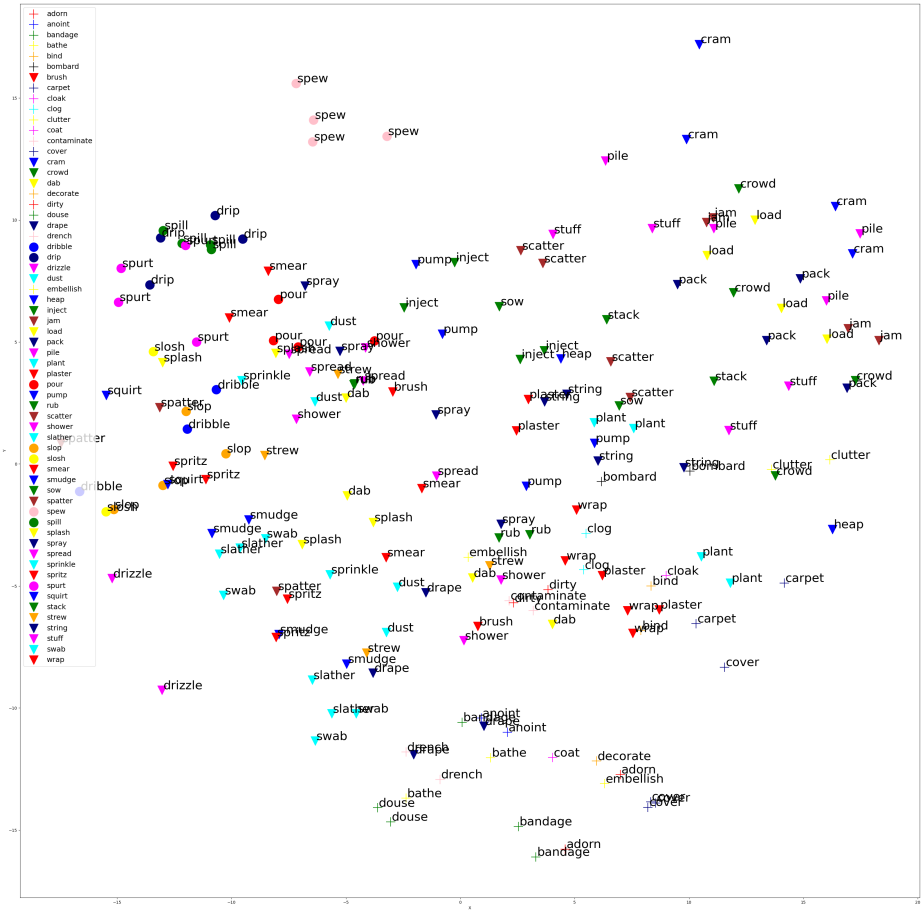

Fig. 2. 2-dimensional PCA projection of "Into/Onto verbs" (filled circles), "With" verbs ("+") signs, and alternating verbs (upside down triangle)

7 Conclusion

The results reported in this paper show that current state-of-the art machine learning systems cannot learn the necessary knowledge to be able to correctly judge the acceptability of the locative alternation, from text input alone. It is suggested that the poverty of the stimulus is a fundamental limitation for statistical learning from text corpora, and practitioners should be aware that their models could have unpredictable "blind spots".

References

[1] C. Baker. "Syntactic theory and the projection problem". In: *Linguistic Inquiry* 10 (1979).

[2] Emily M. Bender and Alexander Koller. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5185–5198. URL: https://www.aclweb.org/anthology/2020.acl-main.463.

[3] Robert C. Berwick et al. "Poverty of the Stimulus Revisited". In: *Cognitive Science* 35.7 (2011), pp. 1207–1242. ISSN: 1551-6709. DOI: 10.1111/j.1551-6709.2011.01189.x.

[4] Noam. Chmomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA., 1965.

[5] Noam. Chomsky. *Rules and Representations*. Columbia Classics in Philosophy. Columbia University Press, 1980. ISBN: 9780231048279. URL: https://books.google.no/books?id=KdYOYJwjFo0C.

[6] Fiona Cowie. "Innateness and Language". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2017. Metaphysics Research Lab, Stanford University, 2017.

[7] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *NAACL-HLT (1)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019, pp. 4171–4186. ISBN: 978-1-950737-13-0. URL: http://dblp.uni-trier.de/db/conf/naacl/naacl2019-1.html#DevlinCLT19.

[8] Allyson Ettinger. "What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models". In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 34–48. DOI: 10.1162/tacl_a_00298. eprint: https://doi.org/10.1162/tacl_a_00298. URL: https://doi.org/10.1162/tacl_a_00298.

[9] Zellig S. Harris. "Distributional Structure". In: *WORD* 10.2-3 (1954), pp. 146–162. ISSN: 0043-7956. DOI: 10.1080/00437956.1954.11659520. URL: http://dx.doi.org/10.1080/00437956.1954.11659520.

[10] Nora Kassner and Hinrich Schütze. "Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7811–7818. URL: https://www.aclweb.org/anthology/2020.acl-main.698.

[11] Beth Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago: The University of Chicago Press, 1993. ISBN: 0-226-47532-8.

[12] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019). arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692.

[13] Brian.W. Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405.2 (1975), pp. 442–451. ISSN: 0005-2795. DOI: https://doi.org/10.1016/0005-2795(75)90109-9. URL: http://www.sciencedirect.com/science/article/pii/0005279575901099.

[14] Amy Perfors, Joshua B. Tenenbaum, and Elizabeth Wonnacott. "Variability, negative evidence, and the acquisition of verb argument constructions". In: *Journal of Child Language* 37.3 (2010), pp. 607–642. DOI: 10.1017/S0305000910000012.

[15] Steven Pinker. *Learnability and Cognition: The Acquisition of Argument Structure (1989/2013)*. New Edition. Cambridge, MA: MIT Press, 2013.

[16] Steven Pinker. *The Stuff of Thought : Language as a Window Into Human Nature*. New York, NY: Viking, 2007.

[17] Magnus Sahlgren. "The distributional hypothesis." In: *Italian Journal of Linguistics* 20 (2008).

[18] Ashish Vaswani et al. "Attention is All You Need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010. ISBN: 9781510860964.

[19] Csaba Veres and Bjørn Helge Sandblåst. "A Machine Learning Benchmark with Meaning: Learnability and Verb Semantics". In: *AI 2019: Advances in Artificial Intelligence - 32nd Australasian Joint Conference, Adelaide, SA, Australia, December 2-5, 2019, Proceedings*. Ed. by Jixue Liu and James Bailey. Vol. 11919. Lecture Notes in Computer Science. Springer, 2019, pp. 369–380. DOI: 10.1007/978-3-030-35288-2_30. URL: https://doi.org/10.1007/978-3-030-35288-2%5C_30.

[20] Alex Wang et al. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: In the Proceedings of ICLR. 2019.

[21] Alex Wang et al. "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems". In: (2019).

[22] Wei Wang et al. *StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding*. 2019. arXiv: 1908.04577 [cs.CL].

[23] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. "Neural Network Acceptability Judgments". In: (2018).

[24] Zhilin Yang et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding". In: *CoRR* abs/1906.08237 (2019). arXiv: 1906.08237. URL: http://arxiv.org/abs/1906.08237.

Notes and Comments. This research was supported by the Project News Angler, which is funded by the Norwegian Research Council's IKTPLUSS programme as project 275872.