Data & Knowledge Engineering 147 (2023) 102208

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Self supervised learning and the poverty of the stimulus

Csaba Veres [a,*], Jennifer Sampson [b]

[a] University of Bergen, Bergen, Norway
[b] Equinor UK Ltd, London, UK

ARTICLE INFO

Keywords:
Classification
Learnability
Text mining
Machine Learning
NLP
Language

ABSTRACT

Diathesis alternations are the possible expressions of the arguments of verbs in different, systematically related subcategorization frames. Semantically similar verbs such as *spill* and *spray* can behave differently with respect to the alternations they can participate in. For example one can "spill/spray water on the plant", but while one can "spray the plant with water", it is odd to say "spill the plant with water". "Spray" is a verb which can alternate between syntactic frames while "spill" is not alternating. How human speakers learn the difference between such verbs is not clearly understood, because the primary linguistic data (PLD) they receive does not appear sufficient to infer the knowledge required for adult competence. More generally the poverty of the stimulus (POS) hypothesis states that the PLD is not sufficient for a learner to infer full adult competence of language. That is, learning relies on prior constraints introduced by the language faculty. We tested state-of-the-art machine learning models trained by self supervision, and found some evidence that they could in fact learn the correct pattern of acceptability judgement in the *locative alternation*. However, we argued that this was partially a result of fine-tuning which introduced *negative evidence* into the learning data, which facilitated *shortcut learning*. Large language models (LLMs) cannot learn some linguistic facts from normal language data, but they can compensate to some extent by learning spurious correlated features when negative feedback is introduced during the training cycle.

1. Introduction

Large Language Models (LLMs) trained with self supervised learning represent a major breakthrough for machine learning. These algorithms construct a comprehensive statistical model of language by performing tasks such as masked language modeling on a vast quantity of unlabelled text [1,2]. This reduces the reliance on manually labelled training data as specific downstream tasks can be performed by *fine tuning* the model with minimal additional annotated data, which can potentially be eliminated altogether on extremely large models that can demonstrate *few-shot* performance [3].

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 [4], but has been part of Chomsky's arguments since at least 1965 [5]. 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 [6,7]. 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 [8] 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,

* Corresponding author.
E-mail address: csaba.veres@uib.no (C. Veres).

https://doi.org/10.1016/j.datak.2023.102208
Received 26 November 2022; Received in revised form 26 July 2023; Accepted 31 July 2023
Available online 9 August 2023

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 cannot 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 [9]. However, with the advent of the transformer architecture [10] 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 [11,12]

Does this result mark the end of the POS hypothesis? Veres and Sandblast [13] argued 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 it is unclear how many examples are potentially unlearnable in the absence of language specific innate priors. As Berwick et al. [6] argue that "responsible nativists" try to account for acquired linguistic knowledge with the minimum language specific component of learning, any test of the POS hypothesis must specifically show that the phenomenon in question is not learnable from PLD. Veres and Sandblast [13] proposed a new benchmark which is composed of grammatical violations related to Baker's paradox [14] which are, by hypothesis, not learnable from PLD alone [15]. Learning is made possible by language specific linking rules between lexico-semantic features and their syntactic expressions. Their results supported the POS argument, that learning requires knowledge about language which is not directly available in the primary linguistic data. The present paper expands on their results, reporting additional experiments which provide stronger evidence that self supervised models are not able to learn certain aspects of linguistic knowledge.

2. Learnability and semantics

The learning problem investigated by Veres and Sandblast [13] concerns verb subcategorisation frames and the possibilities for alternative expressions of the arguments of verbs, also called *diathesis alternations* [16]. The sentence pairs below are examples of an acceptable (the (a) sentences) and an unacceptable (the (b) sentences) alternation in English. Note that unacceptable sentences are conventionally marked with an asterisk. Common examples include the *causative/inchoative* (1), *the dative* (2), *as alternation* (3), and *there-insertion* (4).

(1) a. The little boy broke the window./ The window broke.
b. Margaret cut the bread. / *The bread cut.
(2) a. I gave a book to Roy. / I gave Roy a book.
b. I donated a book to Roy. / *I donated Roy a book.
(3) a. The president appointed Smith press secretary. / The president appointed Smith as press secretary.
b. The captain named the ship Seafarer. / *The captain named the ship as Seafarer.
(4) a. A flowering plant is on the windowsill. / There is a flowering plant on the windowsill.
b. A lot of snow melted on the streets of Chicago. / *There melted a lot of snow on the streets of Chicago.

In this paper we will focus on the *locative*, in particular the *Spray/Load* alternation which denotes a transfer of a substance or set of objects (theme, content, or locatum) into or onto a container or surface (goal, container, or location) [15]. Note here that *theme* and *goal* refer to "the entity that directly receives the action of the verb" and "the direction towards which the action of the verb moves", respectively.[1]

For example the verb *load* can appear in the following constructions (examples taken from [17]).

(5) Hal loaded hay into the wagon.
(6) Hal loaded the wagon with hay.

In sentence (5) the grammatical subject of the verb (Hal) is the one doing the loading, the direct object is the content or locatum being moved (the hay), and the goal or container into which the hay is being moved (wagon) is the object of the preposition *into*. This is called the *content-locative* construction because the focus of the sentence is the content (hay). Sentence (6) swaps the object of the verb to the container, signaling a change in the focus of the action to the container. This is called the *container-locative* construction.

There are many verbs which can undergo this alternation and a tempting 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 if the transformation is applied. Examples (7) and (8) 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 do not allow the generalisation. In these examples *pour*, *fill*, and *load* are all verbs which describe someone moving something somewhere.

(7) a. Amy poured water into the glass.
b. *Amy poured the glass with water.
(8) a. *Bobby filled water into the glass.
b. Bobby filled the glass with water.

1 https://www.linguisticsnetwork.com/semantics-thematic-roles/

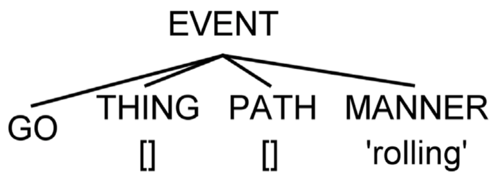

Fig. 1. Tree form representation of semantic core for the verb "roll" which is instantiated in sentence (5).

The fact that adult speakers of English can make these distinctions is a learnability paradox, often called Baker's paradox after it was discussed by Baker [14]. Four conditions lead to the paradox: (a) language speakers spontaneously generalise to newly learned verbs (b) they avoid some possible generalisations (c) they are not corrected for erroneous generalisations (d) there are no self evident systematic differences between verbs that allow generalisation and those which do not. Many attempts to resolve the paradox have proceeded to challenge at least one of these preconditions (e.g. see [15] for a summary). Pinker proposed a novel theory to show that there are, in fact, subtle semantic differences between verbs that do, and do not allow alternations, and it is the semantic conditions that license the expression of particular verb frames and their alternations [15]. The correct use of each verb is determined by its full meaning. Until the learner acquires the meaning of the verb, they tend to be conservative and use it only in the verb frame it was already encountered in. They occasionally use the verb in other constructions expressed by very closely related verbs, but this can result in errors before they have fully learned the verb, since the subtle dimensions of similarity have not yet been mastered. The mechanism responsible for the grammatical expression is therefore a highly specialised component of the language faculty which is responsible for linking conceptual structure to linguistic expressions [15,17].

Pinker's claim is that there is a level of semantic specification at which the almost infinite richness of high level conceptual representations are funnelled into a set of core semantic components that can be externalised in language. The core set of semantic components involve universally recurring ontological types, similar in scope to those proposed by Jackendoff in a series of influential publications [18,19], including Thing, Event, State, Action, Place, Path, Property, and Amount. There are also a number of operators which combine the semantic elements into larger constituents, such as GO, PATH, MANNER, BE, AT, ON, TO, INTO, etc. As a simple illustrative example, consider the sentence in (9)

(9) The ball rolled down the hill.

A simple tree form representation of the semantic constituent defined by the verb *rolled* is shown in Fig. 1. The empty parentheses are open arguments which are linked to constituents in the syntactic structure, in this case THING = "ball" and PATH = "down the hill". The semantic structure therefore determines which entities are expressed in the syntax. The empty parentheses are connected to the syntactic frames through *linking rules* whose only role is to establish the correspondence between the semantic and syntactic roles for individual expressions. Pinker argues that linking rules are universal across all languages and are an unlearned component of the language system [15] p. 268.

This machinery allows Pinker to re-imagine the process of verb frame alternations as transformations on conceptual constituents, which then have consequences for syntactic expression. In effect, Pinker's claim is that the locative alternation can be stated as a rule that takes a verb with a semantic core "X causes Y to move into/onto Z", and converts it into a new form with semantic core "X causes Z to change state by means of moving Y into/onto it". This involves a shift in perspective from an interpretation of *loading* as "moving a theme (e.g., hay) to a location (e.g., a wagon)", to one where *loading* is interpreted as "changing the state of a theme (the wagon) by means of moving something (the hay) into it".

The hypothesis also helps explain the widely studied "holistic/partitive effect" [16] p. 50, where sentences in the container-locative construction are interpreted as complete whereas the content-locative are not. For example in sentence (6) the interpretation is that the wagon is full, but this is not the case in (5), where the wagon might well have just one shovel's worth of hay. If the container-locative sentences imply a change of state then in this case the wagon went from empty to full.

Following the conceptual shift, linking rules accommodate the difference in argument structure. Under the first interpretation the directly effected entity (theme) was the thing moved (hay) which was linked to the direct object of the verb. In the second interpretation the effected entity is the thing that changes state (wagon) and hence it is linked to the direct object. Thus, the two different construals of the same event and the two different argument structures are both the result of this "gestalt switch": loading hay into a wagon is something that happens to hay; loading a wagon with hay is something that happens to a wagon. But how does this help explain which verbs can undergo the gestalt switch, and which cannot?

Since the interpretation shift in the locative alternation converts a movement into a state change, a necessary condition for a verb to participate in the 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 in the end. On the other hand, the non alternating verb *fill* specifies only an end state. If I *fill the bottle with water* then it is clear that the bottle becomes full, but it is not clear how I filled it (hence **fill water into the bottle*). Conversely, if I *pour water into the bottle* then the action I perform is clear, but the end state of the bottle, less so (hence the unacceptable **pour the bottle with water*).

The necessary conditions in themselves do not, however, 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? To explain these facts [15] further proposes a set of narrow range rules which provide fine-grained criteria to license the alternation for sets of verbs. Any particular verb has to satisfy both the necessary broad range and the sufficient narrow range rules in order to show the locative alternation.

The narrow range rules provide specific semantic features that determine the interpretation of narrow *conflation classes* which can or cannot alternate. Consider the example of *drip*: why is it that *I sprayed the plant with water* entails an end state but **I dripped the plant with water* does not? By hypothesis, the fine grained semantic description of *drip* and similar verbs (*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 a predictable end state because we have no direct control over the projection of the dripping liquid. Note that these are rules of *construal*, not a claim about physical reality. As Pinker notes, we cannot afford to debate whether *making someone laugh* is the same sort of causation as *making something fall* when we are speaking at 150 words per minute, even though we can clearly entertain such thoughts when we have the luxury of time. Language does not limit the flexibility of cognition, but when we express thoughts in language we must quickly select an expression that is best suited to that thought, especially in the spoken modality which is the evolved origin of language [20].

The broad and narrow range rules together are 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. 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, or are they affecting the glass by causing it to be less empty? The broad range rule makes this determination for us. When we talk, *pour* can only be a verb that describes an action performed rather than 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 conceptual classification problem. The semantic features are a part of the conceptual - linguistic linking system, and cannot directly be inferred from the statistical properties of the observed linguistic data. Diathesis alternations, on this view are controlled by lexico-semantic facts that are not directly discernible from the productions of the language, cannot be inferred from the statistical distribution of sentences, and therefore should not be learnable by systems that depend entirely on such distributions. This paper presents experiments to see if this hypothesis is supported by empirical facts.

3. Related work

Braine and Brooks [21] described the learnability problem as a problem in avoiding an *overgeneral* grammar. That is, given two grammars where one generates all the grammatical strings of a language and the second generates the grammatical stings as well as ungrammatical ones, how does the child chose the former grammar if all the sentences they hear from adult speakers are consistent with both possible grammars? The problem would be trivial if children received a reliable signal whenever they made an utterance which does not comply with the correct grammar. However there is very little evidence for such a signal in speech corpora, and even at times that the signal is available children tend to reject it [21]. The empirical facts appear to be that children undergo a period during which they produce overgeneral utterances, and that the frequency of errors decrease as they are exposed to more examples of the correct use of specific words. Learning which verbs do not alternate therefore proceeds at the level of individual lexical item, such that novel verbs are assumed to alternate until sufficient evidence builds up that they do not. Pinker's hypothesis is that errors result from an incomplete understanding of the meaning of the verb, which prevents them from being placed in the correct conflation class. Once the child (or adult) learns the correct meaning, the errors should stop.

Braine and Brooks proposed *canonical sentence schemas* which associate the form of the argument structure with the interpretation normally assigned to it. These schemas are based on commonalities of form and meaning extracted from exposure to many exemplars. When new verbs are learned they are initially assumed to participate in all of the meaning appropriate sentence schemas, which can result in ungrammatical utterances. For example in (10a.) the child uses *cover* in the ungrammatical content-locative construction, and in (10b) the verb *spill* is used inappropriately in the container-locative.

(10) a. *I'm gonna cover a screen over me.
b. *I don't want it because I spilled it of orange juice.

As children experience these verbs more often they notice that adults only use them in their canonical grammatical frames, the children abandon the ungrammatical sentences because of *entrenchment* and *preemtion*. Entrenchmet is the principle that if a child learns an argument structure of a verb with sufficient strength (resulting from repetition of the frame) they will assume that is the only frame. This will preempt any temptation to use alternative forms.

Braine and Brooks concede that their canonical sentence schemas are essentially equivalent to Pinker's broad range rules (meaning - form correspondence), but where they diverge is the way in which they explain the conditions under which verbs come to observe the narrow range constraints on frame alternation. Here, Pinker relies on the same mechanism of verb semantics and linking rules whereas Braine and Brooks rely on more general mechanisms like preemption. The former of the two theories postulates knowledge of rich lexical semantic facts and linking rules which are properties of the language system and are not inferred from observed patterns of use, while the latter relies on mechanisms which are closely tied to observed usage.

Brooks and Tomasello [22] ran a novel-word-learning experiment on children to test if they base their usage of novel verbs on Pinker's [15] narrow-range semantic classes, or if they use indirect negative evidence instead. Their results were mixed, showing evidence for narrow-range semantic classes, but only for older children above 4.5 years of age. The results suggest that it takes considerable experience with language use to figure out the appropriate semantic patterns involved in diathesis alternations.

This conclusion is further supported in experiments with Korean college students who were high-level learners of English [23]. The Korean language has different semantic verb conflation classes than English [24], and therefore it is interesting if the native speakers had managed to learn the appropriate narrow-range semantic classes for the English language. The results of a forced-choice picture-description task indicated that, while the Koreans did learn that sentences in the *container-locative (with)* construction have a holistic interpretation, they mistakenly attributed the holistic interpretation to ungrammatical non-alternating sentences. Native English speakers did not make this mistake.

Perfors et al. [25] proposed an alternative model for explaining how the correct pattern of verb use can be learned from positive distributional evidence alone. They proposed a hierarchical Bayesian framework which was able to model many aspects of learning verb constructions, including those involved in Baker's paradox. Their model regarded deviations from expected frequencies as a form of negative evidence. They implemented a hierarchy of inductive constraints, or *overhypotheses*, based on the distributional evidence. The Bayesian model could learn 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 learned prior probabilities that could be used to predict the alternation patterns of verbs in the corpus.

There is, however, an important logical problem with all of the arguments which depend on indirect negative evidence, which is that they rely on sentences that are **not** heard. In other words they assume that learners predict which sentences should be observed if a given hypothesis is true, and abandon that hypothesis if the sentences are not observed. The problem with this assumption is that human language is a generative system whose productions are by nature infinite, so how does the learner know which of a possibly infinite number of sentences that are not encountered, they should be looking out for? As Perfors et al. confess, "this knowledge has just been given to our model" [25] p. 634. Thus it is not the indirect negative evidence that is responsible for learning but the specific inductive biases that come with the proposed learning mechanism. In addition, Pinker [15] argues that the question of negative evidence is a non starter in considering verbs which have low frequencies of usage, such as **I ladled the floor with paint*. It would seem disingenuous to claim that the reason this sentence appears odd is because we once heard the sentence *I ladeled paint on the floor*, and have been waiting ever since in vain for the corresponding alternation.

Kann et al. [26] argued that artificial neural networks, "low-bias learners like ANNs" (p.287), could inform the debate about the need for innate priors if they could acquire specific grammatical knowledge from exposure to text. Their methodology was to test if word embeddings contained sufficient information to support a classification task. Specifically, they tested word and sentence level embeddings for their potential in training classifiers to discriminate between grammatical and ungrammatical constructions. Word level embeddings were used to train a multi-class classifier to learn all the legal frames a given verb could participate in, and a binary classifier to predict if an example sentence embedding containing a particular verb and syntactic frame was grammatically acceptable or not. Kann et al. constructed materials that contained five of the diathesis alternations studied by Levin [16], including the locative. Their results were a mixed bag, with major differences between the different frames, and lack of correspondence between the word and sentence embeddings. In the case of word embeddings for the locative verbs they reported poor performance in predicting word frames with the content-locatives but better performance with container-locatives (*with* sentences). However this did not transfer to the sentence embedding task where classifying the grammaticality of locatives obtained a Matthews correlation coefficient [27] of only 0.261.

A more extensive comparison of verb classes was performed by Veres and Sandblåst [13] who tested twenty-four different alternations studied by Levin [16] using the then state-of-the-art XLNet transformer based model [28] fine tuned on the CoLA dataset [8]. They also found a mixed bag of results in the grammatical acceptability task, with some alternations scoring perfect correlation while others performing at or below chance levels. For example the *Possessor-Attribute Factoring Alternations*, exemplified by the *I admired his honesty. / I admired him for his honesty.* vs. *I sensed his eagerness. / *I sensed him for his eagerness.* resulted in perfect acceptability judgements.

It would be remiss of us to move on without mentioning a few related streams of work who also aim to show the limits of learning from primary linguistic input alone. Bender et al. [29] took 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 language acquisition, it does show that linguistic input is not sufficient for adult linguistic competence, in the general sense.

Kassner and Schütze [30] tested for more specific aspects of linguistic knowledge. They investigated pretrained language models (PLMs) for their general knowledge and concluded that PLMs have difficulty with learning about negation. Given the prompt "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".

4. Experiment 1: Transformer models

The advent of the Transformer architecture [10] represents the biggest advance for natural language processing (NLP) in machine learning. The self-attention mechanism can encode long distance dependencies in sentence structures which are key to building a highly predictive *Language Model* through exposure to massive quantities of unlabelled text. The model can be further *fine tuned* for specific tasks using much smaller amounts of labelled data than was required for alternative architectures. Subsequent

Table 1
Example sentences from the six different sentence types in the experiment.

	Content-locative	Container-locative
Alternating	The farmer had to load apples into the cart.	The farmer had to load the cart with apples.
With only	[a]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.	[a]Carla poured the pitcher with lemonade.

[a]Denotes ungrammatical strings.

Alternating sentences are acceptable in both verb frames while "with only" sentences are only acceptable in the container-locative construction and "Into/Onto/On only" sentences are only acceptable in the content-locative construction.

implementations have proven increasingly successful at synthetic benchmarks like GLUE and SuperGLUE [9,11]. The Transformer architecture has proven to be quite general, showing success in many domains apart from natural language, including drug discovery from models of strings of amino acids [31], predicting protein structure from amino acid chains [32], and generating computer code [33]. These findings show that Transformer models are unlikely to contain language specific learning biases [34] which make them an ideal tool for testing the POS hypothesis.

4.1. Dataset

The preliminary studies of [13] showed a mixed set of results for the 24 different types of diathesis alternations selected from [16]. Amongst the poorest performers were the as-, locative-, reciprocal-, and fulfilling- alternations. The locative alternation we have been describing is one of the best documented, and it was used to construct 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.

4.2. Results

A common metric for acceptability judgement is the Matthews correlation coefficient which typically 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 [27]. Matthews correlation includes every quadrant of the confusion matrix whereas traditional F1 measure ignores the true negatives, which are important when judging grammatical acceptability since it is just as important to correctly classify unacceptable sentences as it is acceptable ones. For example if a model simply classifies every sentence as acceptable in a test set of 36 sentences where only half of the sentences are acceptable then the MCC = 0. On the other hand $F_1 = 0.66$ which is misleadingly high. Eq. (1) shows that the F measure ignores true negatives (TN) whilst MCC does not.

$$F_1 = \frac{2TP}{2TP + FN + FP} \tag{1}$$

$$MCC = \frac{TN * TP - FP * FN}{\sqrt{(TN + FN)(FP + TP)(TN + FP)(FN + TP)}} \tag{2}$$

Matthews Correlation Coefficient is a special case of Pearson Correlation Coefficient. Therefore, the interpretations for both of them are the same[2]

```
If r =
  +.70 or higher Very strong positive relationship
  +.40 to +.69 Strong positive relationship
  +.30 to +.39 Moderate positive relationship
  +.20 to +.29 weak positive relationship
  +.01 to +.19 No or negligible relationship
  -.01 to -.19 No or negligible relationship
  -.20 to -.29 weak negative relationship
  -.30 to -.39 Moderate negative relationship
  -.40 to -.69 Strong negative relationship
  -.70 or higher Very strong negative relationship
```

[2] https://leimao.github.io/blog/Matthews-Correlation-Coefficient/

Table 2
Matthews correlation coefficient for acceptability judgement obtained with human raters.

	Matthews correlation
With only	0.427
Into/Onto only	0.592

Table 3
Matthews correlation coefficient for acceptability judgement obtained with BERT.

	Matthews correlation
Alternating	1.0
With only	0.27
Into/Onto only	0.05

Table 4
Accuracy of acceptability judgement obtained with BERT.

	Grammatical	Ungrammatical
With only	1.0	0.16
Into/Onto only	0.9	0.14

4.2.1. Human evaluation

The materials were tested on human subjects to obtain a baseline estimate for the difficulty of the task. The crowd sourcing platform prolific.co was used as they claim to offer a more reliable research experience than competing platforms.

A total of 100 participants provided acceptability ratings on a six-point scale for the 137 non alternating sentences, where half the sentences were grammatical and half were not. Previous pilot studies showed that human participants were at ceiling in judging the grammaticality of alternating sentences, and since humans are limited in their attention span, we decided to omit the 137 alternating sentences. Results from participants whose response times were more than two standard deviations outside the mean were discarded from the study and replaced with new participants. The instructions were carefully worded to ensure that the participants understood the task, following advice from Fabian Bross' guide to acceptability judgements in linguistics [35]. Matthews correlation compared the vector of human ratings against a dummy vector of 1s for grammatical and 0s for ungrammatical sentences.

The results in Table 2 show strong positive relationship for both sentence types.

4.2.2. Machine learning models

We use the Hugging Face implementation of BERT (Bidirectional Encoder Representations from Transformers) [1]. The model is pre 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).

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

Table 3 shows the Matthews correlation coefficient for the three sentence types. The "Alternating" sentences were tested as a control to see how well the model can perform the task, specifically on sentences which have the locative construction. Ungrammatical versions of the sentences were constructed by randomly permuting the order of two adjacent words in the sentence. In the other two conditions each sentence had an acceptable and an unacceptable version, thereby acting as its own control.

The results show perfect performance for judging grammatical "Alternating" sentences but almost no sensitivity to grammatical acceptability for "Into/onto only" sentences that are ungrammatical in the container-locative construction. We performed a simple accuracy analysis to identify the source of the errors and as Table 4 shows, the reason is a high false positive rate with ungrammatical constructions such as **Carla poured the pitcher with lemonade* being rated as acceptable.

There is a weak positive correlation for "With only" sentences, but accuracy for unacceptable sentences such as **The final step is to coat chocolate onto the cake*, is still low. The increased MCC score appears to be due to a slightly higher accuracy in judging the grammatical sentences.

Since there was a weak positive correlation in at least one condition, we repeated the experiment using RoBERTa, a newer model based on BERT with a robustly optimised pretraining approach [36] which uses a much larger training set, and modifies the training regime by dropping the next sentence prediction task.

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

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

Table 5
Matthews correlation coefficient for acceptability judgement obtained with RoBERTa.

	Matthews correlation
With only	0.17
Into/Onto only	0.4

Table 6
Accuracy of acceptability judgement obtained with RoBERTa.

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

gluebenchmark.com, where the leader for this task at the time of running the experiment was StructBERT from Alibaba with a score of 0.753 [2].

Table 5 shows the Matthews correlations. Surprisingly the "With only" condition shows a slightly worse performance, but now the "Into/Onto only" condition shows a stronger correlation.

The increased correlation in the "Into/Onto only" condition is due to increased accuracy in the ungrammatical condition, as seen in Table 6. The "With only" accuracy is still low, as expected from the weak MCC score. Note also that the result contrasts with Kann et al. who found that word embeddings for "Into/onto only" verbs were less able to learn the frame classification problem than the "With only" verbs.

4.3. Embeddings

Following Kann et al.'s [26] hypothesis that word embeddings contain sufficient information to differentiate between possible verb frames, we performed an analysis on the embeddings computed by the BERT model. The embeddings were extracted from the pretrained model as the elementwise mean of the last four hidden layers.

The semantic content in word embeddings is derived from the context in which words find themselves, as explained in Zellig Harris' hypotheses about the distributional structure of language. Harris proposed that word meanings give rise to observable distributional patterns in language, such that two semantically unrelated words *A* and *C* would be less likely to be found in common linguistic contexts as two semantically related words *A* and *B* [37]. Modern machine learning techniques have made it computationally possible to embed very high dimensional distributional patterns in a much lower dimensional vector space, where the distances between any given vectors are related to the similarities of context in which the corresponding words are found in the training set. Semantic relatedness is therefore correlated with the calculated distance (e.g. cosine distance) between vectors. Sahlgren [38] calls this sort of co occurrence a *paradigmatic relation*. In our judgement the claim is **not** that the sentential context determines the meaning of a word, but that one can infer that two words which appear in similar contexts must have some similarity of meaning.

Fig. 2 shows a 2-dimensional principal components projection of the vector embeddings for the non alternating verbs. The "With only" verbs are shown with a plus sign "+" in the figure, and the "Into/Onto only" verbs with the filled circles. Each verb appears more than once in slightly different regions of vector space, in part because embeddings are contextualised and a given verb has a slightly different vector representation in different sentences. There is a pronounced separation between the two sets of verbs, suggesting that something of the semantic difference was captured in the embedding space, where the difference between the two verb classes is, by hypothesis, that the "Into/onto only" verbs describe a type of motion whereas the "With only" verbs describe an end state. It does appear that the group of verbs on the right of Fig. 1 do describe a manner of motion whereas the ones on the left do not.

Closer inspection of the verbs, however, reveals an alternative explanation for the two groups. It appears that the "Into/Onto only" verbs in the cluster on the right of Fig. 2 can typically appear with various liquids, while the "With only" ones on the left do not. So, for example, pour/dribble/slop/slosh are actions one can perform with water but bandage/bind/decorate/dirty/bombard are not. The difference between the verbs in the two clusters might simply reflect the distributional semantics learned from the context in which words appear.

If the semantic similarity reflected in the verb embeddings really is due only to their paradigmatic relatedness then the non alternating verbs which are also actions performed on liquids should have similar embeddings to alternating verbs, for example *spray* and *sprinkle*. This, in turn, should result in over generalisation of the "Into/Onto only" verbs where they are incorrectly used in the container-locative construction, because of their proximity to alternating verbs. Recall that language users tend to be conservative with newly learned verbs, using them only in their observed constructions or constructions of very closely related verbs. We are claiming that the vector similarity is creating a conflation class based on an imprecise semantic similarity, and therefore the container/locative frame is incorrectly extended to these verbs. In fact this is precisely the result we obtained with BERT, and the results reported by Kann et al. [26] who obtained an MCC of 0.645 in predicting licit verb frames from "With only" verb embeddings and a much lower 0.253 with "Into/Onto only" verbs. The former was in fact the strongest result of all the verb types they looked at, and the latter was close to the weakest. Fig. 3. shows alternating verbs as upside down triangles, revealing that the non alternating "Into/Onto only" verbs do in fact overlap alternating verbs which describe actions performed on liquids. For

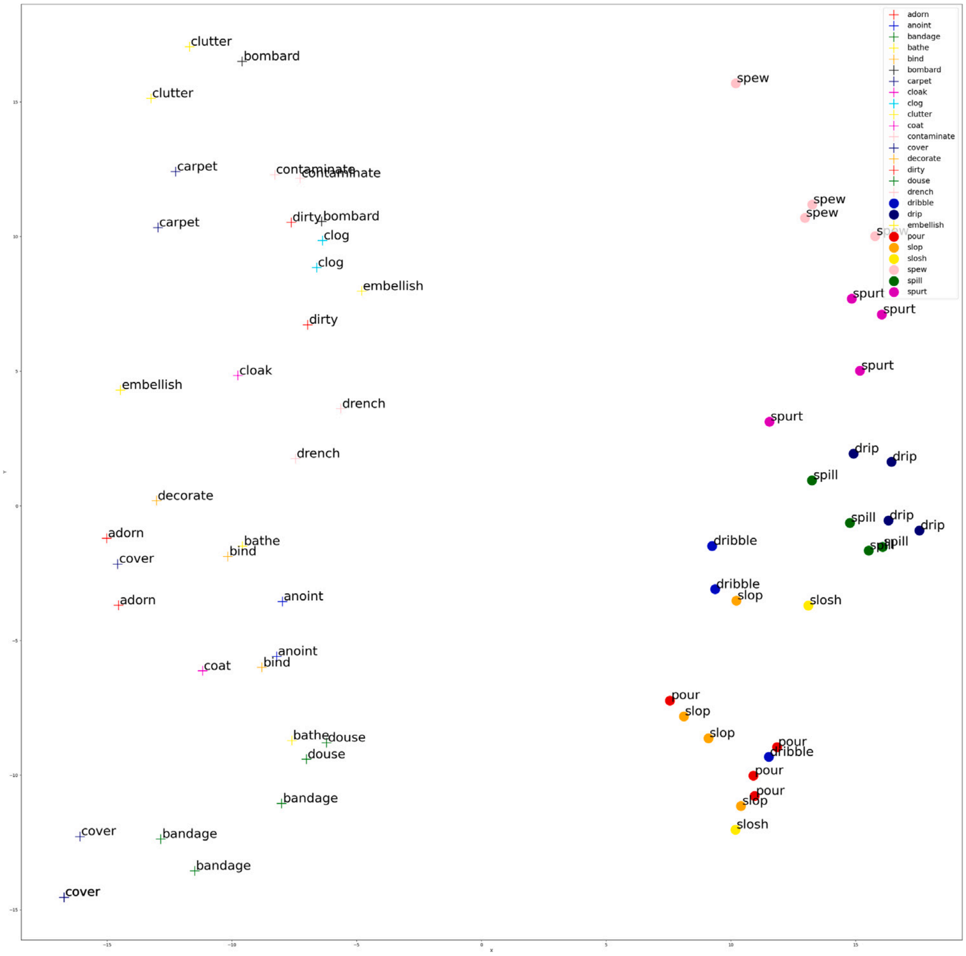

Fig. 2. 2-dimensional PCA projection of "Into/Onto verbs" (filled circles) and "With" verbs ("+"") signs. Note that the embeddings are computed in context, so each token will have a slightly different value on the projection.

example squirt/sprinkle/spray appear next to dribble/pour/spew. The overlap between alternating verbs and "With only" verbs was much less pronounced.

These results from the embedding visualisations suggests a more direct test of the hypothesis that the verb frame selectional preferences depend on simple paradigmatic relations. If the overgeneralisation of the "Into/Onto only" verbs is an artefact produced by their proximity to alternating verbs describing actions on liquids, then this should be reduced in those "Into/Onto only" verbs which are not describing actions on liquids. Our item set contained very few such verbs, even though there was no specific selection criterion to account for this. We therefore constructed a new set of items, "Into/onto solids", which had non alternating verbs that can only be used with solids. By eliminating the overlap we should reduce the number of false positives on ungrammatical non alternating verbs. Subjectively these sentences look to be far less acceptable than the previous examples, when they appear in illicit verb frames. To test our intuition, we performed a preliminary study on a small number of volunteers and obtained a MCC of 1.0, a perfect correlation, showing that people were very good at identifying illicit constructions as unacceptable.

This is interesting in light of Pinker's hypothesis because it suggests that the conceptualisation of the action may be more clear in the "Into/Onto solids" examples, and therefore result in fewer errors of (mis) construal. Consider the sentence pairs in 11 and 12. It seems that the ungrammatical versions of the "solid" sentences are more clearly unacceptable, as marked with the double asterisk. In fact it is very difficult to see how 11b. conveys the same message as 11a. The unacceptable sentences in 12 seem less strongly unacceptable.

(11)
a. Anyone can attach shelves to the wall.
b. **Anyone can attach the wall with shelves.
c. You should glue posters on the wall.
d. **You should glue the wall with posters.
e. The most popular party game is pin the tail on the donkey.
f. **The most popular party game is pin the donkey with the tail.

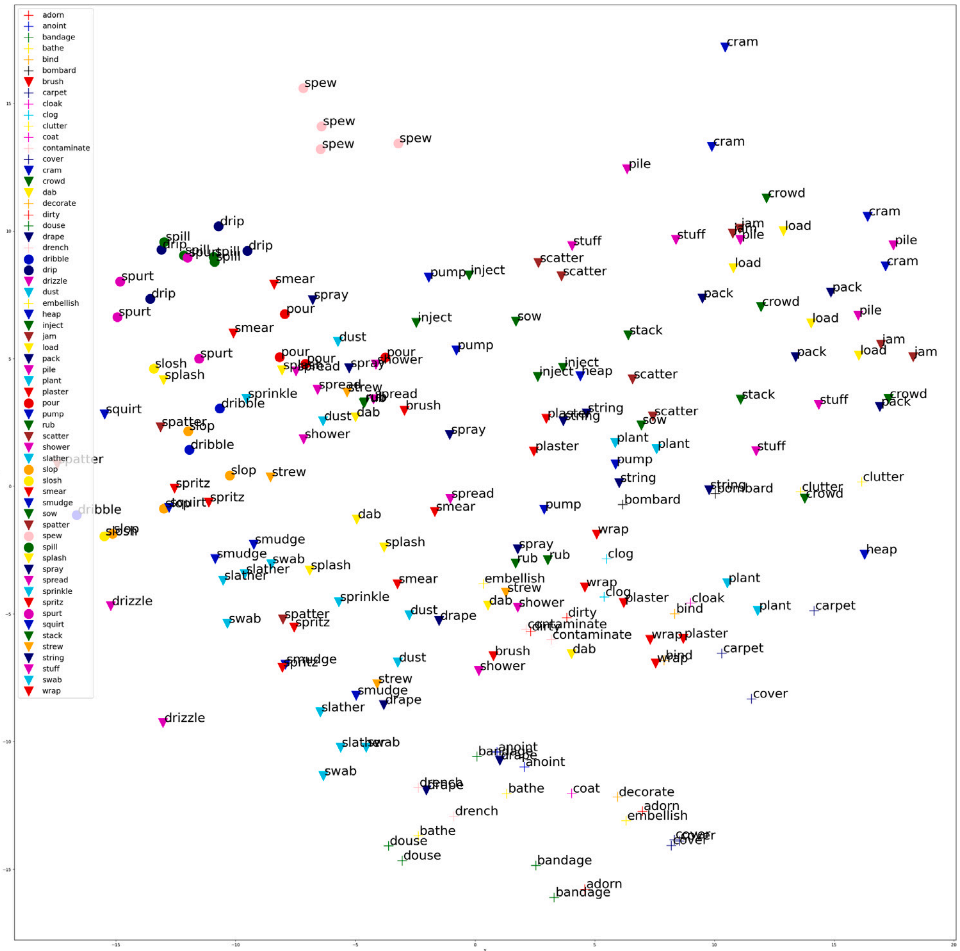

Fig. 3. 2-dimensional PCA projection of "Into/Onto verbs" (filled circles), "With" verbs ("+") signs, and alternating verbs (upside down triangle). Note that the embeddings are computed in context, so each token will have a slightly different value on the projection.

(12) a. John was silly enough to dribble paint onto the floor.
b. *John was silly enough to dribble the floor with paint.
c. If you spill coffee on the couch, wash the area with water and a mild detergent.
d. *If you spill the couch with coffee, wash the area with water and a mild detergent.
e. Tell Carla to pour lemonade into the pitcher.
f. *Tell Carla to pour the pitcher with lemonade.

Fig. 3 also shows that the "With only" verbs do not largely overlap with alternating verbs. If our current proposal that errors in locative verbs are in part caused by their overlap with alternating verbs is true, the prediction would be that the "Into/Onto solids" verbs should behave like the "With only" verbs resulting in better MCC scores. This hypothesis is tested in the following section.

5. Experiment 2: Comparing models

This experiment has three main objectives. The first is to further test the hypothesis that "Into/Onto only" verbs will perform less well than "With only" verbs because of their higher similarity to alternating verbs. Second, reducing the similarity of verb embeddings with alternating verbs will improve the results for "Into/Onto solids" verbs. The third is to test several different models for their performance with our test materials, especially looking at how performance changes with increasing model size. The general trend in machine learning is towards models with ever increasing number of trainable parameters, with a corresponding improvement in performance [39]. The following architectures were used to evaluate the "With only", "Into/Onto only" and "Into/Onto solids" sentences. All models were tested using the Hugging Face Inference API. The name of each model tested is provided in footnotes.

Table 7
Matthew's correlation for three categories of non alternating locatives and a seletion of large language models. For each model we show the number of trainable parameters, in millions.

Model	Parameters	Into/Onto only	With only	Into/Onto solids
distilbert-base	66	0.038	0.177	0.00
bert-base	110	0.13	0.21	0.00
bert-large	340	0.344	0.12	0.17
electra-base	110	0.11	0.3	0.4
electra-large	335	0.62	0.74	0.62
Human		0.592	0.427	1.0

DistilBERT[4] is based on the BERT architecture but is 40% smaller while retaining 97% of BERT's performance [40]. DistilBERT leverages knowledge *distillation* during the pre-training phase, a process where a smaller "student" model is trained to reproduce the behaviour of a larger "teacher", via an objective that includes the whole probability distribution of the output of the teacher. In practice DistilBERT-base has 66 million parameters compared with 110 million for BERT-base.[5] BERT-large[6] on the other hand has 340 million parameters [41].

Electra uses a pre-training method that is more sample-efficient than masked language modeling, called *replaced token detection*. In the conventional training regiment the model generates a candidate token to predict the original identity of a randomly masked token in the input. When training Electra, some of the input tokens are replaced with plausible alternatives sampled from a small generator network, and the model's task during training is to detect which token was replaced. As a result Electra performs discrimination rather than generation during training. The claimed advantage is that an Electra model will outperform a BERT model of similar size and training data because the Electra discrimination task is defined on all tokens of the input because tokens are never masked. Electra-base[7] has 110 million trainable parameters and Electra-large[8] has 335 million.

The results for the models are shown in Table 7, which broadly support the first hypothesis, since four out of five models show better performance for "With only" verbs. However the manipulation for reducing embedding overlap with the introduction of "Into/Onto solids" verbs did not achieve the intended result to improve their performance to the level of "With only" verbs, even though the human results very strongly supported the effect of the manipulation.

Perhaps the effect of the manipulation was dampened by some unknown properties of the language models. Recall that one of the main reasons for making the prediction was that Kann et al. [26] also found high performance with the "With only" verbs. However, that was with a multiclass classification task in which the possible verb frame selectional restrictions of each verb was predicted. The result did not transfer to a binary sentence acceptability task, where the locative performed worst out of their five tested constructions. Clearly the information in word embeddings is highly nuanced when it comes to the sentence level acceptability task.

In comparing the models it is apparent that the expected relative capability of the models is supported in this task. The DistilBERT-base model performed about as well as the larger BERT-base. In turn the 340 million parameter BERT-large improved on this results. The 110 million parameter Electra-base performed slightly better than the equal sized BERT-base, as was predicted. The 335 million parameter Electra-large outperformed the 340 million parameter BERT-large by a sizeable margin. In general larger models performed better, but it is possible to boost the performance of smaller models through specialised training.

There are two important puzzles which emerge from these results, and both concern the divergence of human and machine learning performance. First, the best machine learning model tested, electra-large, returned better-than-human results. Second, the "Into/Onto solids" manipulation did improve human performance but did not reliably affect the machine learning results. These are addressed in the following section.

5.1. GPT-3

How can we explain the better-than-human performance of electra-large? Does this mean the end of the POS hypothesis?

There is a fundamental theoretical difficulty in every experiment described thus far. The logical assumption behind the experiments, which was made explicit by Kann. et al. [26] is that, in order to train a discrimination with embeddings as input, the embeddings must contain information necessary for the discrimination. But it does not follow that the information is the same as what is used for learning about diathesis alternations in the course of human language acquisition. There could be correlated attributes which are sufficient to broadly discriminate classes of verbs but not precise enough to determine the appropriate use of subcategorization frames in real language performance. Thus the answer to the question "These empirical results raise a deeper scientific question: to what extent do the features learned by ANNs resemble the linguistic competence of humans?" (p. 287) might be, not a great deal. The problem is exasperated by the fact that the deep learning models were all fine-tuned on the CoLA dataset, which contains annotated examples of acceptable and unacceptable constructions including the locative, for example

4 distilbert-base-cased-CoLA.
5 textattack/bert-base-uncased-cola).
6 bert-large-uncased-CoLA B.
7 electra-base-avg-cola.
8 howey/electra-large-cola.

Table 8
Two GPT-3 models performing a grammatical acceptability task on locative constructions, as well as sentences from CoLA.

	Into/Onto only	With only	Into/Onto solids	CoLA
text-davinci-002	0.11	0.0	0.24	0.7
davinci-instruct-beta	0.0	−0.24	0.0	0.88

(13) a. 1 carla poured lemonade into the pitcher.
b. 0 * carla poured the pitcher with lemonade.

Sentence 13b. provides direct negative evidence which the model can use to learn the classification task using whatever correlated features it can find. We have no way of determining what the features are, or how closely they align with the linguistic features that actually determine the grammatical status of the verb frame alternations. We have already suggested one possible feature from paradigmatic relatedness, but clearly much more subtle correlated attributes can be learned by a sufficiently powerful model, especially when it receives direct negative evidence, resulting in "shortcut learning" [42]. The combination of shortcut learning and negative evidence has given these models curious behaviours where they can out perform humans in some conditions but under perform in others.

Since the models were trained with negative evidence, they cannot be used as a test bed for POS arguments. In light of our discoveries all previous results, including ours, must be treated with caution. There is, however, a more suitable test bed which we have not considered thus far. The most powerful state-of-the-art models do not necessarily require fine tuning and can demonstrate remarkable few-shot or zero-shot performance without the need for fine tuning on labelled data, for example GPT-3 [3] which scales all the way to 175 billion trainable parameters, and is trained on 300 billion tokens. Since GPT-3 does not receive any labelled data in the course of its training, it is an appropriate model for testing the POS hypothesis.

The model used for testing was the text-davinci-002 model from OpenAI[9] which at the time of writing was their newest and most capable model, designed for instruction-following tasks. This enables it to respond concisely and accurately, even in zero-shot scenarios, without the need for any examples given in the prompt. The davinci model was tested through the API with temperature set at 0 to prevent confabulation. Several values were tested but did not vary the results significantly. In addition, the davinci-instruct-beta model was also tested since the Instruct models are optimised further to follow instructions. This means they are better at producing accurate completions for prompts.

One difficulty with testing GPT-3 is in creating a good prompt to elicit the desired response [43]. We first created a prompt which elicited accurate acceptability judgements for known grammatical and ungrammatical sentences, which in our case were 16 randomly selected sentence pairs from the CoLA dataset, for a total of 32 sentences. Each sentence in a pair was roughly matched on their lexical items but across pairs they displayed a wide range of ungrammaticalities resulting from word order violations, or mismatch in grammatical case, or number. The sentences can be seen in Appendix A. The following prompt elicited very high positive relationship between GPT-3's classification results and the CoLA annotation (Table 8).

> You are doing a language proficiency test and receive this task. For each of the following sentences write down if they are grammatical or not grammatical. In each answer write yes if the sentence is grammatical and write no if the sentence is not grammatical.

Table 8 is quite clear that the 175 billion parameter GPT-3 performs significantly worse than the 335 million parameter model in Table 7. The relationship between model size and performance does not hold when fine tuning is eliminated, suggesting that the additional statistical knowledge in larger models is only useful for learning classification tasks presented for fine-tuning, but is not useful for grammatical acceptability judgements on the locative constructions. In fact, the source of the error was predominantly that the models judged the vast majority of sentences as grammatical. Simply put GPT-3 displayed *ungrammaticality blindness* to the majority of locative sentences in constructions that are deemed unacceptable by humans, showing that the primary linguistic data presented with the self-supervised learning paradigm was not sufficient to arrive at the level of competence required for the correct use of the locative.

5.1.1. ChatGPT

The latest and most well known interface to the GPT series of LLMs is ChatGPT, which has enjoyed public acclaim due to its ease of use, and human-like ability to answer questions and engage in conversational chats. ChatGPT exploits the statistical knowledge in LLMs and extends it by fine tuning with Reinforcement Learning from Human Feedback (RLHF), which makes it a useful tool for answering queries in many domains of application [44]. ChatGPT has sparked the public imagination about the future of AI, as exemplified by Ethank Mollick's article entitled "ChatGPT Is a Tipping Point for AI" in the Harvard Business Review where he describes it as a "... *very* big deal ..."[10]

A particularly useful feature is that ChatGPT has the unique characteristic of generating explanations for its decisions, which we hoped would help us to understand some of the factors which resulted in the GPT model errors. We presented examples of non

[9] https://beta.openai.com/playground?model=text-davinci-002
[10] https://hbr.org/2022/12/chatgpt-is-a-tipping-point-for-ai

alternating verbs for acceptability judgement in the ChatGPT interface,[11] Typical results are shown in the following dialogues. It should be noted that ChatGPT responses can vary widely depending on the way questions are asked, leading to an area of investigation called *prompt engineering*.[12] This investigation used a very basic prompt which was validated as shown in the preliminaries below, with two simple sentences. In addition it should also be noted that the performance can vary with time as the underlying GPT model evolves, with performance not always improving with successive versions [45]. This investigation used ChatGPT (March 14 version) [text-davinci-002-render-sha] https://chat.openai.com/?model=text-davinci-002-render-sha.

Q: Is the following sentence grammatically correct? "The cat is on the mat".

A: Yes, the sentence "The cat is on the mat" is grammatically correct. It is a simple and clear sentence that follows the basic subject-verb-object (SVO) structure. "The cat" is the subject. "is" is the verb (present tense of the verb "to be"). "on the mat" is a prepositional phrase that acts as the object, describing the location of the cat. Overall, the sentence conveys a straightforward statement about the cat's location on the mat.

Q: Is the following sentence grammatically correct? "The cat are on the mat".

A: No, the sentence "The cat are on the mat" is not grammatically correct. The subject "cat" is singular, so it should be paired with the singular form of the verb "to be", which is "is", rather than "are".

The task is performed accurately and with very useful commentary, showing that the prompt is able to elicit binary judgements about grammatical acceptability. An ungrammatical, non alternating sentence was tried next.

Q: Is the following sentence grammatically correct? "Amy poured the glass with water".

A: Yes, the sentence "Amy poured the glass with water". is grammatically correct. However, it might sound more natural if you say "Amy filled the glass with water". or "Amy poured water into the glass". These alternative phrasings are more commonly used in everyday language.

ChatGPT incorrectly classified the sentence as grammatical, but it also suggested some alternatives which are "more commonly used", showing that a slight problem with the sentence was detected. The first alternative is to substitute a different verb, *fill*, which is grammatical in the construction. However this alters the meaning of the sentence to the end state interpretation. The second suggestion is the grammatical content-locative version, which shows that the LLM has some knowledge of the alternative forms.

The result suggests that ChatGPT has acquired some knowledge of the grammatical alternation but has not learned that *pour* is a non alternating verb. However, the suggestion to use "filled the glass", which is does not express the same meaning, could indicate that the response is driven more by commonly used and semantically related expressions rather than the query about syntactic structures. We therefore decided to probe ChatGPT's knowledge of syntax with some seemingly similar but misleading examples. In English, the prepositional phrase beginning "*with* ..". can have interpretations other than the *container-locative* construction we have been considering. For example "She arrived *with* her boyfriend" means 'in the same place as', and "They opened the package *with* a knife" means 'using', or the 'instrumental case'[13] [46]. The previous query was modified by changing the verb to one which elicits the instrumental reading.

Q: Is the following sentence grammatically correct? "Amy cleaned the glass with water".

A: Yes, the sentence "Amy cleaned the glass with water". is grammatically correct. It indicates that Amy used water to clean the glass. However, it is worth noting that depending on the context and what exactly needs to be cleaned off the glass, water alone might not be sufficient. Water is commonly used to rinse or remove dirt, but for more thorough cleaning, additional cleaning agents may be required. So, the sentence is grammatically correct but may need additional information to be contextually accurate.

The answer is clearly consistent with the correct interpretation of the sentence. The answer shows that ChatGPT has another surprising capability, in that it is able to incorporate previous queries and answers in a chat session for answering subsequent queries. Sophisticated tools to manage the information flow in chat sessions have since emerged,[14]. Given the two queries, the chat session now had a number of different possible interpretations for the *with* phrase. The initial query was then repeated to determine if the interpretation prompted by the intermediate query altered the answer.

Q: Is the following sentence grammatically correct? "Amy poured the glass with water".

A: Yes, the sentence "Amy poured the glass with water". is grammatically correct. In this sentence, "Amy" is the subject, "poured" is the verb, and "the glass with water" is the object. It indicates that Amy poured water into the glass. While this sentence is grammatically correct, as mentioned earlier, a more common and natural way to express the same idea would be to say "Amy poured water into the glass".

Indeed, the answer is now quite different, and reveals the extent of naïvety with which ChatGPT answers queries about grammar. ChatGPT understood "the glass with water" as an Noun Phrase (NP) ("the glass") modified by the Prepositional Phrase (PP) ("with water"), which serves as the object of the Verb Phrase (VP) ("pour"). In this example ChatGPT attached the PP to the wrong constituent and therefore accepted a highly implausible and ungrammatical interpretation of the sentence (i.e. "Which glass did Amy pour"? "Amy poured the glass with water".) The memory for previous examples also influenced the answer, once again suggesting "Amy poured water into the glass" as a better way to express *the same idea*. However, the interpretation where "the glass with water" is the object, is **not** the same idea as we just saw!

As a brief aside, the problem of attachment is illustrated by the classic example of structural ambiguity in sentence (14) a., where the two possible interpretations are determined by the position in the syntactic tree where the PP is attached, and therefore which

[11] https://chat.openai.com/

[12] https://en.wikipedia.org/wiki/Prompt_engineering

[13] https://dictionary.cambridge.org/grammar/british-grammar/with

[14] https://python.langchain.com/docs/get_started/introduction.html https://github.com/logspace-ai/langflow

constituent it modifies, as illustrated in (14) b. and c.[15] The failure of ChatGPT to respond appropriately to problems of structural attachment deeply undermine its ability to authentically parse natural language sentences.

(14) a. I saw someone with a telescope.
b. I was using a telescope, and I saw someone. (PP modifies VP)
c. I saw someone, and that person had a telescope. (PP modifies NP)

In conclusion, ChatGPT adds nothing to the underlying language models in answering questions about the locative alternation. In fact it complicates the investigation by introducing memory for chat sessions and by being highly sensitive to the query format.

6. Discussion

We began by considering Baker's paradox which concerns the problem of learning syntactic diathesis alternations from primary linguistic data. The favoured solution from Pinker [15] involved 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 showed that smaller language models were unable to learn the task but larger models were more capable.

However, an analysis of verb embeddings revealed a potential confound, namely that verbs in putative semantic conflation classes could also be related paradigmatically. This, and potentially other correlated features may have enabled the models to learn the classification task during fine tuning with CoLA sentences to the extent demonstrated in the experiments. In order to test if language models which are only exposed to primary linguistic data without fine tuning were able to reliably judge the acceptability of locative constructions we performed a final experiment using the 175 billion parameter GPT-3. The results indicated no evidence that GPT-3 learned to distinguish grammatical and ungrammatical sentences in the locative construction during its massive training period. The result is particularly impressive since it goes completely against the trend that larger and larger models will be able to solve problems which smaller ones cannot. In our experiments even the 110 million parameter BERT model with fine tuning outperformed the 175 billion parameter GPT-3.

The presence of confounding explanatory variables was identified through an analysis of the embedding spaces. Word embeddings encode information that both alternations are possible for individual verbs, from primary linguistic data alone. During fine tuning with positive and negative exemplars, a sub set of verbs are trained to not allow one form of the alternation or other. If these verbs are in a distinct region of vector space then the classification becomes more precise, but if they overlap with other verbs that continue to be observed in both frames, the learning becomes less strong. This leads to the prediction that the "Into/Onto only" verbs which are acceptable only in the content-locative construction would show lower scores than the "With-only" verbs since the former overlap to a greater extent with the alternating verbs. This was in fact shown in Experiment 1 with BERT but not RoBERTa, and was the result reported by Kann et al. [26], on their word embeddings. The trend was also found in four out of five models compared in Experiment 2.

The implication is that, while embeddings are indeed able to encode aspects of the frame selection preferences of verbs, they do not encode the linguistically important properties which define the conflation classes. Verb similarity in embedding space is determined in large part through paradigmatic relations which are more dependent on the kinds of objects that are acted on, rather than the properties of the actions described by the words. If the two are at least partially correlated then the model will partially learn a classification on the correlated features. This highlights a common problem with machine learning, that the algorithms learn from observed patterns in linguistic productions rather than the causal factors giving rise to the productions [47]. In general this can result in "shortcut learning" where the network learns to perform a task using irrelevant but spuriously correlated features [42].

A generative system will causally output volumes of productions with highly correlated constituent structure. The question is, how reliably can the productions be used to infer the properties of the generative system which might involve observable and latent variables. The POS hypothesis claims that it is not always possible to infer the variables, and that domain specific constraints on hypotheses may be needed in order for the productions to be useful in inferring specific causal variables. For example Perfors et al. [25] proposed a Bayesian model which could learn from positive evidence alone, but only because the model had built in assumptions about which kind of **missing** evidence counted as indirect negative evidence. In the absence of such assumptions, we showed that even the most powerful state-of-the-art *language model* was not able to infer the causal variables responsible for explaining the patterns of acceptability in the locative alternation.

Schölkopf et al. [47] argue that current machine learning techniques are unable to infer causal processes in general, which limits their ability to perform on novel tasks because the ability to infer causation is key to *out-of-distribution* generalisation and intelligent thought in general [47,48]. They introduce principles by which future machine learning models could learn causal links between variables from data as well as the suitable units which admit the inferred causal models, thereby combining causal modelling and representation learning. In order to learn the correct diathesis alternations such models would need to infer causal variables for the verb conflation classes [15], or hierarchical *overhypotheses* [25], from the distributional evidence alone, without the benefit of language specific constraints. Whether or not this is possible remains an empirical question, where the predictions from the POS position are clear.

To speculate on the importance of the current results we suggest that in current machine learning approaches which do not contain causal structural models, it becomes impossible to predict if and when they will fail at a given task. Without a theory of the

[15] https://ecampusontario.pressbooks.pub/essentialsoflinguistics2/chapter/structural-ambiguity/

domain it is not possible to know if errors should be expected from the training data. The experiments presented in this paper form a case study where a very specific limitation of language models was identified through a systematic process driven by a rigorous theoretical analysis of the problem domain, in this case natural language. It is our recommendation that any task of significance to society should be analysed with comparable enthusiasm to ascertain if the causal variables required to solve the task were likely to be learnable from the training data alone. For example Ruis et al. [49] argue that large language models fail to make context dependent *implicatures* such as realising that the answer "I wore gloves" to the question "Did you leave fingerprints?" actually means "no". ON our account this should be expected, as the answer requires an understanding of causal variables which cannot be learned from training data. If a machine learning model **does** show proficiency in the task then it should be treated with suspicion and the problem domain should be assessed for spurious correlations among non causal variables which may be exploited by the machine learning algorithm.

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 hence statistical learning cannot provide an explanatory theory of language competence. Nevertheless, it is possible for sufficiently powerful models to learn ways to correctly answer some queries in some circumstances. Such examples of correct performance highlights the danger of illusory success by "shortcut learning" [42] since models can learn classification tasks that are correlated with, but different from the ones we assume they are learning.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A

16 sentence pairs from COLA.
1. They drank the pub dry.
1a. They drank the pub.
2. Harry coughed himself into a fit.
2a. Harry coughed us into a fit.
3. John wrote books.
3a. John write books.
4. I wonder who could solve the problem in this way.
4a. How do you wonder who could solve this problem.
5. Fruit dropped from the tree.
5a. Fruit dropped from the tree from the clouds.
6. John learned French perfectly immediately.
6a. John perfectly learned French immediately.
7. While Truman doesn't want to visit every city, he does Barcelona.
7a. While Rusty might leave in order to please Mag, he won't his father.
8. The knowledge of the problem is quite thorough.
8a. The problem's knowledge is quite thorough.
9. He will put the chair between some table and some sofa.
9a. What table will he put the chair between some table and?
10. They can cry.
10a. They can happy.
11. John suddenly got off the bus.
11a. John suddenly got the bus off.
12. John deposited some money in the checking account on Friday and Mary did the same thing on Monday.
12a. John deposited some money in the checking account and Mary did the same thing in the savings account.
13. The monkeys kept forgetting their lines.
13a. The monkeys kept forgot their lines.
14. Harry says that Sally dislikes him.
14a. Harry says that Sally dislikes himself.

15. Not speaking English is a disadvantage.
15a. Speaking not English is a disadvantage.
16. Who do you believe invited Sara?
16a. Who do you believe that invited Sara?

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Jill Burstein, Christy Doran, Thamar Solorio (Eds.), NAACL-HLT (1), Association for Computational Linguistics, ISBN: 978-1-950737-13-0, 2019, pp. 4171–4186, URL http://dblp.uni-trier.de/db/conf/naacl/naacl2019-1.html#DevlinCLT19.

[2] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, Luo Si, StructBERT: Incorporating language structures into pre-training for deep language understanding, 2019, arXiv:1908.04577.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei, Language models are few-shot learners, 2020, http://dx.doi.org/10.48550/ARXIV.2005.14165, URL https://arxiv.org/abs/2005.14165.

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

[5] Noam Chomsky, Aspects of the Theory of Syntax, MIT Press, Cambridge, MA., 1965.

[6] Robert C. Berwick, Paul Pietroski, Beracah Yankama, Noam Chomsky, Poverty of the stimulus revisited, Cogn. Sci. (ISSN: 1551-6709) 35 (7) (2011) 1207–1242, http://dx.doi.org/10.1111/j.1551-6709.2011.01189.x.

[7] Fiona Cowie, Innateness and language, in: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Fall 2017, Metaphysics Research Lab, Stanford University, 2017.

[8] Alex Warstadt, Amanpreet Singh, Samuel R. Bowman, Neural network acceptability judgments, 2018.

[9] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, The Proceedings of ICLR, 2019.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, Illia Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS '17, Curran Associates Inc., Red Hook, NY, USA, ISBN: 9781510860964, 2017, pp. 6000–6010.

[11] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman, SuperGLUE: A stickier benchmark for general-purpose language understanding systems, 2019.

[12] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, Quoc V. Le, XLNet: Generalized autoregressive pretraining for language understanding, 2019, CoRR arXiv:1906.08237, URL http://arxiv.org/abs/1906.08237.

[13] Csaba Veres, Bjørn Helge Sandblåst, A machine learning benchmark with meaning: Learnability and verb semantics, in: Jixue Liu, James Bailey (Eds.), AI 2019: Advances in Artificial Intelligence - 32nd Australasian Joint Conference, Adelaide, SA, Australia, December 2-5, 2019, Proceedings, in: Lecture Notes in Computer Science, Vol. 11919, Springer, 2019, pp. 369–380, http://dx.doi.org/10.1007/978-3-030-35288-2_30.

[14] C. Baker, Syntactic theory and the projection problem, Linguistic Inquiry 10 (1979).

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

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

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

[18] Ray Jackendoff, Semantics and Cognition / Ray Jackendoff, MIT Press Cambridge, Mass, ISBN: 0262100274, 1983, p. xiii, 283.

[19] Ray Jackendoff, Semantic structures / ray jackendoff, MIT Press Cambridge, Mass, 1990, p. xiv, 322, ISBN: 0262100436, 026260020.

[20] Steven Pinker, Paul Bloom, Natural language and natural selection, Behav. Brain Sci. 13 (4) (1990) 707–727, http://dx.doi.org/10.1017/s0140525x00081061.

[21] Martin Braine, Patricia Brooke, Verb argument structure and the problem of avoiding an overgeneral grammar, in: Michael Tomasello, William E. Merriman (Eds.), Beyond Names for Things: Young Children's Acquisition of Verbs, Lawrence Erlbaum, Hillsdale, NJ, 1995, pp. 353–376.

[22] Patricia J. Brooks, Michael Tomasello, How children constrain their argument structure constructions, Language 75 (4) (1999) 720–738, ISSN: 00978507, 15350665, URL http://www.jstor.org/stable/417731.

[23] Robert Bley-Vroman, Hye-Ri Joo, The acquisition and interpretation of english locative constructions by native speakers of Korean, Stud. Second Lang. Acquis. 23 (2) (2001) 207–219, ISSN: 02722631, 14701545, URL http://www.jstor.org/stable/44485736.

[24] Soonja Choi, Mélissa Bowerman, Learning to express motion events in english and Korean: The influence of language-specific lexicalization patterns, Cognition 41 (1991) 83–121.

[25] Amy Perfors, Joshua B. Tenenbaum, Elizabeth Wonnacott, Variability, negative evidence, and the acquisition of verb argument constructions, J. Child Lang. 37 (3) (2010) 607–642, http://dx.doi.org/10.1017/S0305000910000012.

[26] Katharina Kann, Alex Warstadt, Adina Williams, Samuel R. Bowman, Verb argument structure alternations in word and sentence embeddings, in: Proceedings of the Society for Computation in Linguistics (SCiL) 2019, 2019, pp. 287–297, URL https://www.aclweb.org/anthology/W19-0129.pdf.

[27] Brian W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta (BBA) (ISSN: 0005-2795) 405 (2) (1975) 442–451, http://dx.doi.org/10.1016/0005-2795(75)90109-9, URL http://www.sciencedirect.com/science/article/pii/0005279575901099.

[28] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, Quoc V. Le, XLNet: Generalized autoregressive pretraining for language understanding, 2019, CoRR arXiv:1906.08237, URL http://arxiv.org/abs/1906.08237.

[29] Emily M. Bender, 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, Association for Computational Linguistics, 2020, pp. 5185–5198, Online, URL https://www.aclweb.org/anthology/2020.acl-main.463.

[30] Nora Kassner, 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, Association for Computational Linguistics, 2020, pp. 7811–7818, Online, URL https://www.aclweb.org/anthology/2020.acl-main.698.

[31] Daria Grechishnikova, Transformer neural network for protein-specific de novo drug generation as a machine translation problem, Sci. Rep. 11 (1) (2021) 321, http://dx.doi.org/10.1038/s41598-020-79682-4.

[32] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, Demis Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature (ISSN: 0028-0836) 596 (7873) (2021) 583–589, http://dx.doi.org/10.1038/s41586-021-03819-2.

[33] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, Oriol Vinyals, Competition-level code generation with AlphaCode, 2022, http://dx.doi.org/10.48550/arxiv.2203.07814, arXiv arXiv:2203.07814.

[34] Csaba Veres, Large language models are not models of natural language: They are corpus models, IEEE Access (ISSN: 2169-3536) 10 (2022) 61970–61979, http://dx.doi.org/10.1109/access.2022.3182505.

[35] Fabian Bross, Acceptability Ratings in Linguistics: A Practical Guide to Grammaticality Judgments, Data Collection, and Statistical Analysis. Version 1.02, URL www.fabianbross.de/acceptabilityratings.pdf.

[36] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, CoRR arXiv:1907.11692, URL http://arxiv.org/abs/1907.11692.

[37] Zellig S. Harris, Distributional structure, WORD (ISSN: 0043-7956) 10 (2–3) (1954) 146–162, http://dx.doi.org/10.1080/00437956.1954.11659520.

[38] Magnus Sahlgren, The distributional hypothesis, Italian J. Linguist. 20 (2008).

[39] Md. Zahangir Alom, Tarek Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, Mst Nasrin, Mahmudul Hasan, Brian Van Essen, Abdul Awwal, Vijayan Asari, A state-of-the-art survey on deep learning theory and architectures, Electronics 8 (2019) 292, http://dx.doi.org/10.3390/electronics8030292.

[40] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, 2019, http://dx.doi.org/10.48550/ARXIV.1910.01108, URL https://arxiv.org/abs/1910.01108.

[41] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018.

[42] Melanie Mitchell, How do we know how smart AI systems are? Science 381 (6654) (2023) adj5957, http://dx.doi.org/10.1126/science.adj5957, arXiv:https://www.science.org/doi/pdf/10.1126/science.adj5957, URL https://www.science.org/doi/abs/10.1126/science.adj5957.

[43] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, Graham Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021, http://dx.doi.org/10.48550/arxiv.2107.13586, arXiv:2107.13586.

[44] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, Bao Ge, Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models, 2023, arXiv:2304.01852.

[45] Lingjiao Chen, Matei Zaharia, James Zou, How is ChatGPT's behavior changing over time?, 2023, arXiv:2307.09009.

[46] Wikipedia contributors, Instrumental case — Wikipedia, the free encyclopedia, 2023, https://en.wikipedia.org/w/index.php?title=Instrumental_case&oldid=1140152022, [Online; accessed 21-July-2023].

[47] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, Yoshua Bengio, Towards causal representation learning, 2021, http://dx.doi.org/10.48550/ARXIV.2102.11107, URL https://arxiv.org/abs/2102.11107.

[48] Judea Pearl, Dana Mackenzie, The Book of Why. The New Science of Cause and Effect, Basic Books, New York, ISBN: 978-0-465-09760-9, 2018.

[49] Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, Edward Grefenstette, Large language models are not zero-shot communicators, 2022, arXiv:2210.14986.

Csaba Veres was born in Budapest in 1964. Csaba received the Ph.D. degree in cognitive science from the University of Arizona, Tucson. His thesis was in the field of Psycholinguistics, where he studied the role of meaning on sentence parsing and representation. He subsequently began work as a computer and information scientist at Melbourne University in Australia.

He is currently Full Professor at the Department of Information Science and Media Studies at the University of Bergen, Norway. His areas of expertise include NLP, machine learning, and semantic web technologies. He has experience as an Academic and Practitioner. He founded a Norwegian company called LexiTags, and consulted as Head of AI with the London based educational technology startup, Flooved. He has published a wide assortment of original research articles, and has had popular linked data apps on the Apple store, called MapXplore and AuotoMind. He has also held positions as a Research Scientist at the Australian Defence Science and Technology Organisation, and as a Senior Lecturer in the Department of Information Systems, Melbourne University.

Jennifer received a Ph.D. in computer science from the Norwegian University of Science and Technology, (NTNU) in Trondheim, Norway. Her thesis combined semantic technologies with information visualisation pioneering the use of whole ontology visualisation techniques for alignments. She has a Master of Business Studies in Information Systems from Massey University, New Zealand. Jennifer currently works for Equinor UK, in London, as an NLP practitioner and lead data scientist.