[Proceedings of the Society for Computation in Linguistics](https://scholarworks.umass.edu/scil)

Volume 2 Article 30

2019

Verb Argument Structure Alternations in Word and Sentence Embeddings

Katharina Kann
New York University, kann@nyu.edu

Alex Warstadt
New York University, warstadt@nyu.edu

Adina Williams
New York University, adinawilliams@nyu.edu

Samuel R. Bowman
New York University, bowman@nyu.edu

Verb Argument Structure Alternations in Word and Sentence Embeddings

Katharina Kann*, **Alex Warstadt***, **Adina Williams*** **and Samuel R. Bowman**
New York University, USA
{kann, warstadt, adinawilliams, bowman}@nyu.edu

Abstract

Verbs occur in different syntactic environments, or frames. We investigate whether artificial neural networks encode grammatical distinctions necessary for inferring the idiosyncratic frame-selectional properties of verbs. We introduce five datasets, collectively called FAVA, containing in aggregate nearly 10k sentences labeled for grammatical acceptability, illustrating different verbal argument structure alternations. We then test whether models can distinguish acceptable English verb–frame combinations from unacceptable ones using a sentence embedding alone. For converging evidence, we further construct LaVA, a corresponding word-level dataset, and investigate whether the same syntactic features can be extracted from word embeddings. Our models perform reliable classifications for some verbal alternations but not others, suggesting that while these representations do encode fine-grained lexical information, it is incomplete or can be hard to extract. Further, differences between the word- and sentence-level models show that some information present in word embeddings is not passed on to the downstream sentence embeddings.

1 Introduction

Artificial neural networks (ANNs) are powerful computational models that are able to implicitly learn syntactic and semantic features necessary for a variety of natural language tasks. These empirical results raise a deeper scientific question: to what extent do the features learned by ANNs resemble the linguistic competence of humans?

Studying the linguistic competence of ANNs, in addition to its intrinsic value for model evaluation, can help resolve outstanding scientific questions in linguistics about the role of prior grammatical bias in human language acquisition. Chomsky (1965) suggests that the acquisition of rich grammatical distinctions is facilitated by an innate universal grammar (UG), which imparts specific grammatical knowledge to the learner. This proposal crucially depends on the *poverty of the stimulus* argument, which holds that the acquisition of certain linguistic features by purely domain-general data-driven learning should not be possible (Clark and Lappin, 2011). Studying the ability of low-bias learners like ANNs to acquire specific grammatical knowledge can provide evidence relevant to this argument.

In this work, we evaluate ANNs' treatment of verbs; verbs contribute to the overall meaning of sentences by encoding information about how entities are related to, and participate in, events. Concretely, we investigate if ANNs acquire the specific grammatical distinctions necessary for inferring the frame-selectional properties of verbs. Cross-linguistically, the lexical entry of a verb is associated with a set of syntactic contexts or *syntactic frames* in which it can appear. This information is lexically idiosyncratic, i.e., even verbs that are intuitively very similar in meaning may vary as to which syntactic frames they can appear in:

(1) a. Sharon **sprayed** water on the plants.
b. Sharon **sprayed** the plants with water.
c. Carla **poured** lemonade into the pitcher.
d. *Carla **poured** the pitcher with lemonade.[1]

Certain verbs, e.g., *spray*, select multiple related frames and are therefore known as *alternating verbs*. In contrast, other semantically similar verbs, e.g., *pour*, select only a single frame and are thus not alternating. Information about whether a given verb alternates (as well as which frames it

*The first three authors contributed equally and are listed in alphabetical order.

[1] In this paper, stars mark ungrammatical sentences.

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 287-297.
New York City, New York, January 3-6, 2019

Verb Frame	Example Sentences		
Caus. Inch.	Jessica **dropped** the vase. The vase **dropped**.	Jessica blew the bubble. *The bubble blew.	
Dative-Prep. Dative-2-Obj.	Liz **gave** a gift to the boy. Liz **gave** the boy a gift.	Liz administered a test to the kid. *Liz administered the kid a test.	*Liz charged $50 to Jon. Liz charged Jon $50.
Spr.-Lo.-*with* Spr.-Lo.-Loc.	Sue **loaded** the truck with wood. Sue **loaded** wood onto the truck.	Sue coated the deck with paint. *Sue coated paint on the deck.	*Sue swept the bin with sand. Sue swept sand into the bin.
no-*there* *there*	Fear **remained** in my mind. There **remained** fear in my mind.	A girl focused on the quiz. *There focused on the quiz a girl.	
U.-Obj.-Refl. U.-Obj.-No-Refl.	Ada **clapped** her hands. Ada **clapped.**	Ada permed her hair. *Ada permed.	*Ada exercised herself. Ada exercised.

Table 1: Examples from each verb frame in the dataset. Bolded verbs evoke both verb frames; other verbs evoke only one. Transitive verb frames include: Causative, SPRAY–LOAD *with*, SPRAY–LOAD locative, UNDERSTOOD-OBJECT reflexive. Intransitive verb frames include: Inchoative, no-*there* (with locative adjunct), *there* (with locative adjunct), and UNDERSTOOD-OBJECT no-reflexive. 2-obj. class includes a ditransitive frame and a prepositional dative frame.

can appear in) has been described and classified in several verb lexica (Grishman et al., 1994; Baker et al., 1998; Fillmore et al., 2003; Kipper-Schuler, 2005; Kipper-Schuler et al., 2006). Knowledge about verb frames and their alternations is part of a human speaker's linguistic competence, and as such, should potentially be learned by ANNs.

We present two datasets and two experiments that compare ANNs' knowledge of verb frame alternations at the word level and the sentence level, respectively. First, we ask if a verb's word embedding can be used to predict which frames that verb can licitly appear in. We construct a dataset of verbs, the **L**exic**a**l **V**erb–frame **A**lternations dataset (LaVA), based on Levin (1993), and train a multi-class classifier to identify the licit syntactic frames associated with a verb from its word embedding alone (if successful, the classifier should be able to determine, e.g., that *sprayed* alternates and can appear in sentences with *with*-alternants like (1-b), but that *poured* cannot (1-d)).

Second, we ask whether sentence embeddings encode the frame-selectional properties of their main verb. The main verb's frame-selectional properties have consequences for grammaticality at the sentence level; to give an example, (1-d) is not grammatical, because *poured* cannot participate in this frame alternation. To exploit this, we semi-automatically generate sentences in such a way to ensure that the main verb's frame alternation information is the only information determining the (un)grammaticality of the sentence. For a portion of the sentences, the main verb can participate in a given verb frame alternation, and for another portion it cannot; if the main verb cannot participate in the alternation, then one of the sentences in the pair will be ungrammatical. Using this dataset, the **F**rames and **A**lternations of **V**erbs **A**cceptability dataset (FAVA), we train a binary classifier to judge the acceptability of sentences containing verbs in various syntactic contexts using the sentence embeddings alone.

We find that verb frame information is extractable from both word embeddings and sentence embeddings, but that these two complementary methods differ in performance. The LaVA and FAVA datasets are available under `https://nyu-mll.github.io/CoLA` for future research and model evaluation.

2 Verb Frame Alternations

The lexical meaning of each verb includes a description of an event and how entities participate in it (Fillmore, 1966; Fillmore et al., 2003), and this information is present for the various syntactic frames associated with each verb. To determine whether our ANNs encode this information, we select five verb frame alternations from Levin (1993); the verb frames which comprise each alternation vary either in the number of arguments they can take, in the order in which the arguments appear, or in both. Examples are given in Table 1, and statistics are provided in Tables 2 and 3.

To give an example, in (1-a), there is an event of *spraying* in which *Sharon* is the main actor (often referred to as *agent*), *the plants* is the entity affected by the event (i.e., the *patient*), and *water* is the entity used in the event (i.e., the *instrument* or *theme*). In (1-a), the verb frame of *spray* has three

Levin class	CAUS.–INCH.		DATIVE		SPRAY–LOAD		*there*-INSERTION		UNDERSTOOD-OBJECT	
	Inch.	Caus.	Prep.	2-Obj.	*with*	Loc.	no-*there*	*there*	Refl.	No-Refl.
Positive	70	120	63	72	90	81	50	145	11	81
Negative	140	(0)	356	405	220	229	185	(0)	466	396
Total	210	120	419	477	310	310	235	145	477	477

Table 2: Overview of the lexical dataset. "Positive" refers to the number of verbs that evoke each frame (i.e., will yield a grammatical sentence) and "negative" refers to the number of verbs which do not evoke those frames (i.e., will yield an ungrammatical sentence). Causative and *there* sentence frames have no negative examples (i.e., every verb participating in the alternation can instantiate these frames).

roles, and they come in a specific order: the *agent* is the subject, the *instrument* is the object, and the *patient* or *location* is part of a prepositional phrase adjoined to the verb. Participants (e.g., *Sharon* and *water*) that are provided by the verb are called *arguments* of the verb; the other argument *the plants* is within a prepositional phrase and is therefore not provided by the verb.

Whether a verb can introduce a given number of arguments can affect its sentence-level grammaticality and is therefore of interest here. Verbs can be *intransitive*, taking only one argument (e.g., *dropped* in *the vase dropped.*), *transitive*, taking two arguments (e.g., *dropped* in *Jessica dropped the vase*), or *ditransitive*, taking three arguments (e.g., *gave*, in *Liz gave the boy a gift*).

Two different verb frames may be related by the addition or deletion of an argument (e.g., CAUSATIVE-INCHOATIVE), or by realizing the same arguments in a different syntactic configuration (e.g., SPRAY-LOAD; (1-a) and (1-b)). When several verbs with similar argument structures can productively appear in such related verb frames, this is called an *argument structure alternation*. Examples are listed in Table 2.

For some alternations, there are examples of verbs that participate in both frames (e.g., are positive examples for both dative and double object frames), only the first frame (e.g., are positive examples for the dative frame, but negative examples for the double object one), or only the second frame (e.g., are positive examples for the double object frame and negative ones for the dative frame). However, full empirical coverage is not always possible for every alternation. In our corpora, two of our alternations (CAUSATIVE–INCHOATIVE and *there*-INSERTION) are sparse; some of their frames cannot be provided with negative examples. We discuss this issue in more detail in Section 3.1).

3 Datasets

In this section, we describe in detail our word-level dataset, which we call the **L**exic**a**l **V**erb–frame **A**lternations dataset (LaVA); and the corresponding sentence-level dataset, which we call the **F**rames and **A**lternations of **V**erbs **A**cceptability dataset (FAVA). Five argument structure alternations are chosen and verbs that evoke at least one frame of the alternation are included in our lexical corpus. These verbs are subsequently used to semi-automatically create a sentence acceptability corpus for our second experiment. We describe our selected argument structure alternations in the remainder of this section and introduce our corpora.

3.1 LaVA—The Lexical Corpus

We construct LaVA from 515 verbs manually mined from five of the largest syntactic verb frame alternations provided by Levin (1993): CAUSATIVE–INCHOATIVE, DATIVE, SPRAY–LOAD, *there*-INSERTION, and UNDERSTOOD-OBJECT. Each alternation consists of two different syntactic frames. Our dataset lists whether each verb participates in each frame (wherever available, see the subsection on sparsity below); the alternations and their verb frames are described in the following.

CAUSATIVE–INCHOATIVE Alternation The CAUSATIVE–INCHOATIVE (Sundén, 1916; Fillmore, 1966; Hale and Keyser, 1986, 2002) dataset is an expanded version of the CAUSATIVE–INCHOATIVE dataset from Warstadt et al. (2018), and it contrasts verbs which can evoke both causative and inchoative frames, like *drop* in Table 1, with verbs that can evoke only the causative frame, like *blow*. Importantly, the causative frame is *transitive*—taking two syntactic arguments—and the inchoative frame is *intransitive*—taking only one. In the causative frame, the subject (e.g., *Jessica*) causes the object (e.g., *the vase*) to undergo a change of state (e.g., to be *dropped*), but, in the inchoative frame, the argument which undergoes a change of state is the subject.

DATIVE Alternation The DATIVE (Bresnan, 1980; Marantz, 1984; Larson, 1988) dataset consists of verbs that indicate transfer of possession; both frames evoked by these verbs take three arguments, but the two frames differ in the order of arguments. In the prepositional dative frame, the *theme* is the syntactic object of the verb, and the *recipient* is within a prepositional phrase; in the dative double object frame, there is no prepositional phrase, and both the *theme* and the *recipient* appear after the verb. Table 1 provides examples from the three sets of verbs: one set of verbs evokes both the prepositional dative frame and the double object frame (e.g., *give*), another set only evokes the prepositional dative frame and not the double object frame (e.g., *administered*), and the last set of verbs only evokes the double object frame, but not the prepositional dative frame (e.g., *charged*).

SPRAY–LOAD Alternation The SPRAY–LOAD (Tenny, 1987; Levin and Hovav, 1995; Arad, 2006) dataset includes transitive verb frames that relate to putting objects in places or covering things with other objects as described in Section 2.

***There*-INSERTION Alternation** The *there*-INSERTION (Poutsma, 1904; Milsark, 1974; Szabolcsi, 1986) dataset contains intransitive verbs that can evoke a frame in which the subject of the sentence (e.g., *fear*) follows the verb (as in *There remained fear in my mind.*, despite the fact that it would usually appear before the verb in other frames; for these sentences the subject position is filled with a dummy word, *there*. The *there* frame requires a prepositional phrase adjunct—e.g., *There remained fear *(in my mind)*—but the no-*there* frame does not—e.g., *Fear remained (in my mind)*. Verbs that evoke both frames are verbs of existence, spatial configuration, meandering movement, manner of motion, appearance, and inherently directed motion.

UNDERSTOOD-OBJECT Alternation The UNDERSTOOD-OBJECT (Rice, 1988; Levin, 1993) dataset contains verb frames that vary in transitivity and describe conventionalized movements of body parts. In the transitive UNDERSTOOD-OBJECT reflexive frame, the body part is the object of the verb (e.g., *Ada clapped her hands.*). In the intransitive UNDERSTOOD-OBJECT no-reflexive frame, the affected *theme* participant (e.g., the body part, or *hands*) is recoverable from the verb (e.g., *clapped*) even though the frame does not require the *theme* (i.e., we know that Ada is clapping her hands and not something else when we interpret the object-less sentence *Ada clapped*).

Sparsity Due to the nature of verb argument structure alternations, in some cases no negative examples can be obtained. For instance, there are no English verbs that can appear in the inchoative, but not the causative (see the first two columns of Table 2). This means that, for the CAUSATIVE–INCHOATIVE alternation, verbs can either evoke both causative and inchoative frames (i.e., be positive examples for both frames) or just the causative frame (i.e., be a positive example for causative and a negative example for inchoative). Similarly, there are verbs that can appear in only no-*there*, but no verbs that can only appear in the *there* frame. This leads to sparsity of annotations. As a result, word-level classifications for these frames are trivial.

Another factor that contributes to data sparsity is that our lexical corpus relies on verbs that Levin (1993) provides as positive (i.e., *grammatical*) or negative (i.e., *ungrammatical*) examples; it does not provide grammaticality judgments for each verb in every frame. In some cases, this is for a linguistic reason: CAUSATIVE–INCHOATIVE alternation verbs can take at most two arguments, and thus do not appear in frames requiring 3 arguments like the prepositional dative or double object frames. In other cases, there is no obvious reason for a particular verb to not appear in another frame, but the annotations in Levin (1993) do not provide that verb–frame combination. In many of these cases, we augment Levin's judgments with our own, also semi-automatically, in attempts to alleviate this issue. However, despite these efforts, the resulting dataset is still sparse, i.e., it does not list whether every verb is a positive or negative example for every frame.

3.2 FAVA—Acceptability Judgments Corpus

FAVA is a set of nearly 10k sentences with acceptability judgments. It is constructed semi-

Levin Class	Sentences	% Positive
CAUSATIVE–INCHOATIVE	1168	78.9
DATIVE	644	70.2
SPRAY–LOAD	5127	58.6
there-INSERTION	718	77.0
UNDERSTOOD-OBJECT	705	54.2

Table 3: Sentence counts for our acceptability corpus. "% Positive" is the percentage of sentences that count as acceptable, i.e., as positive examples.

automatically from the verbs in the lexical corpus; Table 3 provides a brief overview.

Two of the authors, both trained as linguists, manually construct lexical sets consisting of verbs with similar frame-selectional properties that are paired with semantically plausible nouns (and prepositions, where needed). These lexical sets are used to automatically generate sentences with different syntactic frames. For example, the lexical set in (2) is used to generate 18 minimal pairs of sentences as in (3) (one pair for each combination of verb, patient, location, and preposition).

(2) verbs = {hung, draped}
patients = {the blanket, the towel, the cloth}
locations = {the bed, the armchair, the couch}
prepositions = {over}

(3) a. Betty draped the blanket over the couch.
b. *Betty draped the couch with the blanket.

A similar, semi-automatic sentence creation method focusing only on the passive alternation (and non-argument structure syntactic reorderings using negation and relative clauses) was employed by Ettinger et al. (2016) and Warstadt et al. (2018).

Using this method, we construct five sentence-level datasets highlighting different verb alternations (CAUSATIVE–INCHOATIVE,[2] DATIVE, SPRAY–LOAD, *there*-INSERTION, UNDERSTOOD-OBJECT) that are chosen so that sentences could be generated with the maximum of variability in the choice of verbs. We split our data into training, development, and test sets by binning lexical sets into training and evaluation bins randomly, in equal proportions. The evaluation set is then split 80/20 into test and development set. Splitting by lexical bin rather than by sentence prevents models from finding a trivial solution to classification by learning to recognize specific verbs and verbal arguments from the training set in the evaluation or test set.

[2]The CAUSATIVE–INCHOATIVE dataset presented here is an expanded version of an analysis dataset in Warstadt et al. (2018).

4 Pre-Trained Representations

Embeddings, i.e., vector representations of linguistic objects like characters, words, or sentences, encode helpful information for downstream applications (Mikolov et al., 2013). In particular, they can be used to leverage knowledge from one task for another and have been shown to improve performance on a diverse set of tasks. Embeddings are usually low-dimensional; common sizes differ between 100 and 300. Our experiments make use of three types of word and sentence embeddings, which we will describe in the following.

Word Embeddings For our word-level experiments, we use two different embeddings which differ in the way of their creation. First, we use 300-dimensional GloVe embeddings trained on 6B tokens (Pennington et al., 2014).[3] GloVe embeddings are used frequently in natural language processing (NLP), so evaluating them for knowledge of verb frames will be relevant for their application to and future research on tasks requiring rich syntactic features. Second, we use embeddings trained on the smaller 100M token British National Corpus[4] (BNC), optimizing a language modeling objective. The language model (LM) is a (single-directional) LSTM trained by Warstadt et al. (2018) using PyTorch and optimized using Adam (Kingma and Ba, 2015). The BNC data is tokenized using NLTK (Bird and Loper, 2004) and words outside the 100k most frequent words in the BNC are replaced with `<unk>`.

Our peripheral interest in how humans learn lexical frame-selectional properties motivates us to investigate these LM-trained word embeddings. We reduce the potential differences between human learners and our models by considering embeddings that are trained on an amount of data similar to what humans are exposed to during language acquisition. For this reason, most publicly available, pre-trained word vectors are a rather unnatural fit, since these embeddings are usually trained on several orders of magnitude more data than humans see in a lifetime.[5]

[3]`http://nlp.stanford.edu/data/glove.6B.zip`

[4]`http://www.natcorp.ox.ac.uk`

[5]If we extrapolate from data gathered by Hart and Risley

Sentence Embeddings We further produce sentence embeddings with the help of an existing sentence encoder. Namely, we employ the sentence encoder trained by Warstadt et al. (2018) which performs best in their downstream acceptability classification task. The encoder is trained on a real/fake discrimination task. This is a binary classification task in which a model learns to distinguish *naturally occurring* sentences in the BNC from *fake* sentences. Fake sentences themselves are either generated by a LM or by permuting naturally occurring sentences. The real/fake dataset consists of about 12M sentences, including about 6M sentences from the BNC, about 3M million LM-generated sentences, and 3M permuted sentences. The data is tokenized and unknown words replaced in the same way as in the LM training data. A development set is used for early stopping. 20 real/fake encoders are trained for 7 days or until the completion of 4 training epochs without improvement in Matthews correlation coefficient on the development set.

The architecture of the real/fake encoder is shown in Figure 1. A bidirectional long-short term memory network (LSTM, Hochreiter and Schmidhuber, 1997) reads the words of a sentence. A fixed-length sentence embedding is then produced by a max-pooling operation over the concatenations of the forward and backward hidden states at each time-step. This encoding serves as input to a sigmoid output layer, which outputs a binary prediction. The input to the encoder are ELMo-style (Peters et al., 2018) contextualized word embeddings from a trained LM. As in ELMo, the representation for a word w_i is a linear combination of the hidden states h_i^j for each layer j in an LSTM LM, though we depart from that paper by using only a forward LM.

As argued in Warstadt et al. (2018), this sentence encoder is a reasonable model for a human learner because it is not exposed to any knowledge of language that could not plausibly be part of the input to a human learner. Its training data consists of the same 100 million tokens used to train the word embeddings, augmented with another 100 million generated tokens in the *fake* data.

(1992), we can estimate that children are exposed to about 100 million tokens, on average, by age 5.

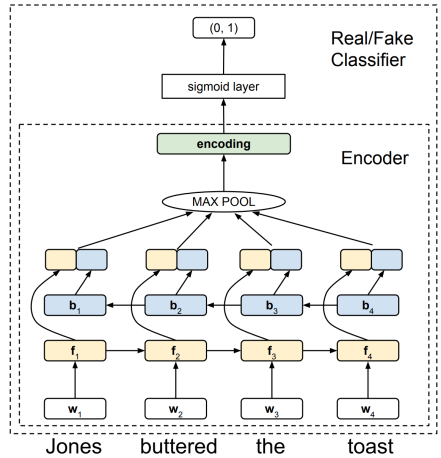

Figure 1: Real/fake model. w_i = word embeddings, f_i = forward LSTM hidden state, b_i = backward LSTM hidden state. Figure from Warstadt et al. (2018).

5 Experiment 1: From Word Embeddings to Argument Structures

In our first experiment, we aim at classifying acceptable syntactic frames, given embeddings for each of the verbs.

5.1 Model

Architecture We cast the identification of syntactic frames in which a verb can appear as a multi-label classification problem. We train one classifier per alternation, and the classes to be predicted correspond to the participating frames (cf. Table 1), i.e., each classifier predicts values for 2 different classes.

Employing a multi-layer perceptron (MLP) with a single hidden layer, the probability of a syntactic frame s being acceptable for a given verb is modeled as:

$$p(s) = \sigma(W_2(f(W_1 x)) \quad (1)$$

Here, x is the input, i.e., a word embedding representing a given verb, W_1 and W_2 are weight matrices, σ denotes the sigmoid function, and the activation function f is a rectified linear unit (ReLU).

Hyperparameters and Training Regime We employ the same hyperparameters for all word-level classifiers. In particular, we use 30-dimensional hidden states; note that the size of the embedding vectors is defined by the type of embeddings we use. During the final classification, we use a threshold of 0.7 to map the model's predictions to binary outputs.

For training, we use the Adam (Kingma and Ba, 2015) optimizer. All ANNs are trained for 15 epochs, but we apply the best performing model on the test set. Further, we use 4-fold cross-validation: the set of verbs is split into 4 equally sized parts out of which 2 are chosen to be the training set, 1 functions as the development set and 1 as the test set.

5.2 Metrics

We report both accuracy and Matthews correlation coefficient (MCC, Matthews, 1975) for this and the following experiment (cf. Section 6), but primarily rely on MCC for evaluation following Warstadt et al. (2018). MCC is a special case of Pearson's *r* for binary classification. It measures correlation between two binary distributions in the range from -1 to 1, with any two unrelated distributions having a score of 0, regardless of class imbalance. As such, this metric is more robust to unbalanced classification than traditional metrics like F1 or accuracy, both of which favor classifiers with a majority class bias.

5.3 Results

Table 4 shows our results. Our first observation is that, overall, accuracies for GloVe and CoLA-style embeddings are comparable for all classes. This suggests that they both contain similar information about verbs and syntactic frames, and is in line with the fact that both embeddings are based on co-occurrences of words.

Second, we find that, for GloVe embeddings, the MLP performs on par with the majority baseline for some verb frames, namely causative and *there*, as well as DATIVE prep. and DATIVE 2-Obj.; a look at the model predictions reveals that it indeed predicts the majority class for all examples. In this case, MCC will be zero, which is indicative of situations where the model predictions are no better than random. We will not further analyze these cases, since the results likely indicate that our lexical dataset does not contain enough examples for the model to learn from, and, thus, do not tell us anything meaningful about the embeddings. We would like to note that methods which explicitly account for skewed datasets might help for DATIVE prep. and DATIVE 2-Obj., but we leave an investigation of such methods for future work.

Finally, we obtain a weak (0.1–0.5) to moderate (0.5–0.7) MCC for both embedding methods and all other classes (with the MLP's accuracy also often being higher than that of the majority baseline). This indicates that information about the evoked syntactic frames can indeed be extracted from verb embeddings. Relatively good performance (>0.45) is found for the inchoative frame (both embeddings), the DATIVE 2-Obj. frame (CoLA), the *with* frame (both embeddings), and *no*-there frame (both embeddings). Since our classification method (an MLP) is rather simple, our results can be considered a lower-bound on performance, thus showing that verb-frame information is rather obvious in our investigated embeddings.

6 Experiment 2: From Acceptability to Acceptable Argument Structures

Linguists are able to arrive at a classification of a verb according to its syntactic frames by interrogating whether sentences with a given verb and frame are acceptable. Analogously, we can observe whether a verb's frame-selectional properties can be extracted from a sentence embedding by training an acceptability classifier to distinguish sentences with acceptable from sentences with unacceptable verb-frame combinations. If a classifier is able to reliably classify all minimal pairs of several verbs with different frame-selectional properties from a sentence embedding alone, we can infer that the sentence embedding contains enough information to distinguish both the frame-selectional properties of the verbs and the relevant syntactic frames.

Model Our acceptability classifier is again an MLP with a single hidden layer. We model the probability that a that sentence S is acceptable as:

$$p(S) = \sigma(W_2(tanh(W_1 x))) \qquad (2)$$

Here x is the input, a sentence embedding obtained from the real/fake sentence encoder described in Section 4, W_1 and W_2 are weight matrices, σ denotes the sigmoid function, and $tanh$ is the hyperbolic tangent activation function. We use a threshold of 0.5 to map the model's predictions to binary outputs.

Training Details To select hyperparameters, we train 20 acceptability classifiers on each of the five datasets, and an additional 20 classifiers on a dataset produced by aggregating all the datasets. We repeat all experiments augmenting each dataset with the more than 10k sentences

		CAUSATIVE–INCHOATIVE		DATIVE		SPRAY–LOAD		*there*-INSERTION		UNDERSTOOD-OBJECT	
		Inch.	(Caus.)	Prep.	2-Obj.	*with*	Loc.	no-*there*	(*there*)	Refl.	Non-Refl.
CoLA: Majority BL	Acc.	66.7	(100.0)	85.0	84.9	71.0	73.9	78.7	(100.0)	97.7	83.0
CoLA: MLP	MCC	**0.555**	0.0	0.32	**0.482**	**0.645**	0.253	**0.459**	0.0	0.0	0.219
	Acc.	81.0	(100.0)	86.6	88.3	85.8	72.9	84.3	(100.0)	97.7	79.0
GloVe: Majority BL	Acc.	66.8	(100.0)	85.0	85.3	71.0	74.6	79.1	(100.0)	97.6	81.5
GloVe: MLP	MCC	**0.672**	0.0	0.0	0.0	**0.585**	0.145	**0.536**	0.0	0.0	0.3
	Acc.	85.5	(100.0)	85.0	85.3	83.9	73.4	85.8	(100.0)	97.6	73.2

Table 4: Results from Experiment 1 for CoLA-style embeddings (top) and GloVe embeddings (bottom); "Majority BL" denotes the majority baseline. Bolded MCC values represent reasonably strong correlations (above 0.45). Results for the majority baselines differ due to different words not having a vector representation within the respective embeddings. The corpus does not contain negative examples for caus. and *there* frames (parenthetical); these results cannot be interpreted and are only included for completeness.

from the corpus of linguistic acceptability (CoLA) built by Warstadt et al. (2018). Hyperparameters are chosen by random search within the following ranges: hidden size $\in [20, 100]$, learning rate $\in [10^2, 10^5]$, and dropout rate $\in \{0.2, 0.5\}$. All models are trained using early stopping with a patience of 20 epochs.

6.1 Results

Table 5 shows results for acceptability classification on the verb–frame datasets. These results lead us to conclude that the sentence encoder we test does reliably encode some fine-grained lexical information, but fails to do so in all cases. Our models are able to perform reliable acceptability classifications on several of the alternations featured in FAVA, achieving a moderate correlation (0.5–0.7) in 5 out of 12 experiments, and a strong correlation (>0.7) in one experiment. Most classifiers achieve a correlation above 0.3.

Across all verb classes, augmenting the training data with CoLA examples lowers MCC. However, when evaluating on the aggregate dataset augmenting the training data with CoLA improves MCC. One explanation for this might be that the distribution from which the test set is drawn does not resemble the training distribution: for instance, in the CAUSATIVE–INCHOATIVE with CoLA set, training examples illustrating the relevant alternation are outnumbered about 20:1 by CoLA examples that illustrate mostly unrelated syntactically or semantically complicated phenomena.

On the other hand, augmenting the combined dataset with sentences from CoLA helps. Performing well on the combined dataset requires an acceptability classifier with knowledge of several unrelated phenomena, so it is not surprising that augmenting the verb-alternation sentences with domain-general CoLA data improves performance.

The easiest phenomenon by a wide margin for acceptability classifiers was the UNDERSTOOD-OBJECT alternation. One explanation for this fact might be that the semantic relatedness of verbs like *blink* and objects like *her eyes* makes it easier to recognize from the sentence embedding whether their co-occurrence is expected or anomalous; for example, *eye* is the most common collocate for *blink*, *hand* is the most common one for *clap*, and *tooth* is in the top five most common collocates for *chip* (Davies, 2008, 2009).[6]

The next easiest alternations for our models to learn are CAUSATIVE–INCHOATIVE and *there*-INSERTION, both of which have at least one intransitive verb frame (both frames are intransitive in the case of *there*-INSERTION, but in one frame there is a locative adjunct). One common denominator among these three easiest alternations for the acceptability model is that they all involve verbs appearing in an intransitive frame (in the case of *there*-INSERTION a locative adjunct is present as well). By contrast, the DATIVE and SPRAY–LOAD alternations both involve verbs that take multiple arguments, appearing with up to three arguments (or possibly two arguments and a locative adjunct) in all frames. Intransitive verb frames are the simplest syntactic frames possible, and it might be expected that they are easiest to recognize.

Qualitatively, we do not find that the amount of training examples in the dataset was correlated with performance. By way of illustration, the SPRAY–LOAD alternation accounts for over half

[6] https://corpus.byu.edu/coca/

		Comb.	CAUSATIVE–INCHOATIVE	DATIVE	SPRAY–LOAD	*there*-INSERTION	UNDERSTOOD-OBJECT
w/o CoLA	MCC	0.290	**0.603**	0.413	0.323	**0.528**	**0.753**
	Acc.	64.6	85.4	76.0	66.2	72.9	87.4
w/ CoLA	MCC	0.361	**0.464**	0.329	0.261	**0.523**	**0.638**
	Acc.	68.7	81.2	59.0	63.4	72.5	81.8
Majority BL	MCC	0.0	0.0	0.0	0.0	0.0	0.0
	Acc.	66.6	77.6	82.1	60.3	77.5	53.7

Table 5: Results from Experiment 2. "w/o CoLA" are models trained on datasets not augmented with CoLA; "w/ CoLA" are models trained on augmented datasets; "Comb." refers to an aggregate dataset. Bolded MCC values represent moderate correlations (above 0.45).

of all the generated data, yet it was by far the hardest individual alternation for our models to learn.

7 Related Work

This investigation is part of a growing body of work which seeks to investigate the linguistic competence of ANNs. For instance, a study by Linzen et al. (2016) tested the ability of ANNs to identify mismatches in subject–verb agreement, even in the presence of intervening "distractor" nouns. Similarly, Ettinger et al. (2016) investigated whether sentence embeddings contain grammatical information, e.g., about the syntactic scope of negation.

Further previous studies on which types of information are contained in embeddings include Bjerva and Augenstein (2018), which asked whether certain phonological, morphological and syntactic information can be extracted from language embeddings. Malaviya et al. (2017) predicted features from language embeddings which were trained as part of an ANN for machine translation. Finally, Östling and Tiedemann (2017) learned language embeddings via multilingual language modeling and used them to reconstruct genealogical trees. However, we are interested in *word* or *sentence* embeddings. Extracting information from word embeddings is a common task in natural language processing. While most NLP research is application-oriented and directly or indirectly focuses on obtaining embeddings which contain as much knowledge about the task at hand as possible (e.g., by varying the training corpus or embedding method), we are interested in the question how much information is trivially contained in selected popular embeddings.

Also worth mentioning here is a lexical resource named VerbNet (Kipper-Schuler, 2005; Kipper-Schuler et al., 2006). This database contains verbs which were classified according to their semantic and syntactic properties, including their Levin classes.[7] VerbNet has been used in various NLP applications, e.g., semantic role labeling (Giuglea and Moschitti, 2006), word sense disambiguation (Brown et al., 2011), information extraction (Mausam et al., 2012), or investigation of human language acquisition (Korhonen, 2010). While this resource is very extensive, it only provides a few example sentences (generally only one or two per frame) for each verb. Since we want to investigate if argument structure information is present in sentence embeddings, we create a larger corpus.

8 Conclusions

We present complementary word-level and sentence-level datasets, LaVA and FAVA, covering five verb-alternations. We train classifiers on verb embeddings to distinguish which syntactic frames a verb can evoke and which it cannot. We further train acceptability classifiers with sentence embeddings as input for sentences which do or do not contain acceptable verb–frame combinations. We conclude that information about verb-argument structure alternations is present in both word-level and sentence-level embeddings. However, some frames seem to be easier to judge than others, and for only few frames a strong correlation can be obtained between model predictions and our gold annotations. There is considerable opportunity for future work which generalizes these experiments to other sentence encoders, verb alternations, and lexical properties.

Acknowledgments

This project has benefited from financial support to SB and KK from Samsung Research, and to SB from Google.

[7] To be exact, the set of classes was extended to a superset of the original Levin classes.

References

Maya Arad. 2006. *The Spray-Load Alternation*. Wiley Online Library.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*.

Steven Bird and Edward Loper. 2004. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Joan Bresnan. 1980. Polyadicity: Part i of a theory of lexical rules and representations. *Lexical Grammar*, pages 97–121.

Susan Windisch Brown, Dmitriy Dligach, and Martha Palmer. 2011. VerbNet class assignment as a WSD task. In *Proceedings of the 9th International Conference on Computational Semantics*.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Alexander Clark and Shalom Lappin. 2011. *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons.

Mark Davies. 2008. *The corpus of contemporary American English*. BYE, Brigham Young University.

Mark Davies. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*.

Charles J Fillmore. 1966. A proposal concerning english prepositions. *Monograph Series on Languages and Linguistics*, 19:19–34.

Charles J Fillmore, Christopher R Johnson, and Miriam RL Petruck. 2003. Background to framenet. *International Journal of Lexicography*, 16(3):235–250.

Ana-Maria Giuglea and Alessandro Moschitti. 2006. Semantic role labeling via framenet, verbnet and propbank. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

Ralph Grishman, Catherine Macleod, and Adam Meyers. 1994. COMLEX syntax: Building a computational lexicon. In *Proceedings of the 15th conference on Computational Linguistics*.

Ken Hale and Jay Keyser. 1986. Some transitivity alternations in English. *Anuario del Seminario de Filología Vasca "Julio de Urquijo"*, 20(3):605–638.

Kenneth Locke Hale and Samuel Jay Keyser. 2002. *Prolegomenon to a theory of argument structure*, volume 39. MIT press.

Betty Hart and Todd R Risley. 1992. American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental Psychology*, 28(6):1096.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 2015 International Conference on Learning Representations*.

Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, The University of Pennsylvania.

Karin Kipper-Schuler, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Anna Korhonen. 2010. Automatic lexical classification: Bridging research and practice. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1924):3621–3632.

Richard K Larson. 1988. On the double object construction. *Linguistic Inquiry*, 19(3):335–391.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Beth Levin and Malka Rappaport Hovav. 1995. *Unaccusativity: At the syntax-lexical semantics interface*, volume 26. MIT press.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Alec Marantz. 1984. On the nature of grammatical relations. *Linguistic Inquiry Monographs*, 10.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Gary L Milsark. 1974. *Existential sentences in English*. Ph.D. thesis, Massachusetts Institute of Technology.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Hendrik Poutsma. 1904. *A Grammar of Late Modern English: the elements of the sentence*, volume 1. P. Noordhoff.

Sally Rice. 1988. Unlikely lexical entries. In *Proceedings of the 14th Annual Meeting of the Berkeley Linguistics Society*.

Karl Fritiof Sundén. 1916. *Essay I. The Predicational Categories in English: Essay II. A Category of Predicational Change in English*, volume 1. At the University Press, E. Berling.

Anna Szabolcsi. 1986. Indefinites in complex predicates. *Theoretical Linguistic Research*, 2:47–83.

Carol Lee Tenny. 1987. *Grammaticalizing aspect and affectedness*. Ph.D. thesis, Massachusetts Institute of Technology.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.