Tupleised co-occurrence measures vs LLM word embeddings for corpus linguistics: The case of English light verb construction detection

Ryan Ka Yau Lai
University of California, Santa Barbara
kayaulai@ucsb.edu

Abstract

This paper examines how word embeddings from large language models (LLMs) can be leveraged for corpus-linguistic studies of co-occurrence. Specifically, I examine whether Phrase-BERT (Wang et al. 2021) representations contain information about co-occurrence properties of English verbs and nouns, such as token frequency, attraction, productivity and dispersion, and if so, how Phrase-BERT can be used alongside such measures in corpus-linguistic analyses. I find that (a) Phrase-BERT representations partially encode information from co-occurrence statistics, (b) Phrase-BERT by itself predicts quite well whether a verb-noun combination is a light verb construction, but predictions are further improved by corpus statistics and semantic information, (c) Phrase-BERT's predictions as to whether something is an LVC can be partially explained through corpus statistics.

1 Introduction

Co-occurrence is at the heart of both corpus and computational linguistics. Both fields are interested in exploring forms that regularly co-occur with each other to form *collocations* or *multi-word expressions*. Both began studying co-occurrence with similar methods: counting co-occurrence between pairs of forms, computing statistics for measuring the salience of co-occurrence, and choosing the highest-scoring pairs (Dras & Johnson 1996, Evert 2005, Tan et al. 2006 etc.).

Yet the two traditions have parted ways. Modern computational linguistics treats the extraction of multi-word expressions as a sequence labelling problem (e.g. Waszczuk et al. 2019, Taslimipoor & Rohanian 2018): Given a sequence of tokens in a corpus, how can we label the beginning and end of multi-word expressions? The methodology has moved beyond statistics to using pre-trained large language models (LLMs), which calculate the probabilities of strings of tokens using very large corpora.

Meanwhile, corpus linguistics has further developed the traditional method. Rather than a single co-occurrence statistic (such as PMI or G^2), recent work suggests that co-occurrence properties are better captured by suites of statistics that operationalise different aspects of distribution with different psycholinguistic interpretations (e.g. Gries 2022a, 2024, van Hoey 2023). This movement towards multi-dimensionality is called *tuplelisation*: it involves gathering combinations, or *tuples*, of corpus statistics. Crucial to this development is the realisation that correlation between statistics comprising the tuples should be minimised, and the introduction of tools to do so (Gries 2022b, 2022c).

Nevertheless, the versatility and accuracy of black-box language models remain attractive for corpus linguists. For example, while a linguist cannot obtain accurate co-occurrence statistics for a pair of words involving a word that did not occur in the corpus, this is unproblematic if we use word embeddings (vector-space representations) based on LLMs: word vectors are trained on much larger corpora and, in their modern incarnations, can handle unseen words, since word embeddings are creating by combining embeddings of subwords: fragments of words determined by a tokeniser.

Given the strengths of LLM word embeddings, one may ask how to integrate them into the corpus linguist's workflow without sacrificing the linguistic interpretability desired in theoretical corpus-linguistic work, and how it make it work alongside traditional corpus-linguistic methods. Extensive work has demonstrated that LLM word embeddings encode all types of linguistic information, from word classes (Belinkov et al. 2018) to agreement and anaphora (Lin et al. 2019), named entities and semantic roles (Tenney et al. 2019), syntactic structures (Jawahar et al. 2019) and, crucially for this paper, constructional information (Tayyar Madabushi et al. 2020), including filler-slot attraction (Thrush et al. 2020). This suggests that LLM behaviour can be pinned down to aid corpus-based investigations of language use, including co-occurrence.

This paper tackles this question through the case study of association between verbs and their objects in English, particular as regards the identification of *light verb constructions*, combinations of a semantically light verb with a semantically heavy lexical noun, as such constructions are particularly relevant to corpus-based lexicography and constructicography. LLM-based word embeddings are taken from Phrase-BERT (Wang et al. 2021).

Specific research questions of this paper are:

1. To what extent do Phrase-BERT embeddings of verb-object sequences encode co-occurrence information between the verb and the head noun of the object?
2. Do tupleised co-occurrence statistics encode any information useful for identifying light verb constructions *not* already present in Phrase-BERT?
3. Can tupleised co-occurrence statistics, along with semantic and syntactic information, be used to interpret *how* Phrase-BERT predicts whether a verb-object sequence is a light verb construction?

Section 2 gives the background information to this paper. Section 3 describes the nature of the datasets used. Section 4 shows that Phrase-BERT embeddings can partially predict tupleised co-occurrence statistics calculated from the British National Corpus (BNC; Leech 1992). Section 5 examines the detection of light verb constructions. It demonstrates that corpus statistics are still useful when used alongside Phrase-BERT embeddings for LVC detection. It also shows how tupleised corpus statistics can help interpret the behaviour of a Phrase-BERT-based model of light verb detection.

2 Background

2.1 Covarying collexeme analysis

The linguistic phenomenon studied in this paper is combinations of verbs and objects within a specific construction type in English: active, transitive clauses. Thus, it can be regarded as a covarying collexeme analysis (Stefanowitsch & Gries 2005): We are looking at the co-occurrence of items within two constructional slots of a construction.

2.2 Tupleised co-occurrence statistics

The corpus statistics used in this paper are mostly based on Gries (2022a). Most of the measures are calculated using values from the following contingency table, where *n* stands for noun (i.e. the object), *v* stands for verb, and ¬ means ‘not’:

	n	$\neg n$	Totals
v	$f(n,v)$	$f(\neg n,v)$	$f(v)$
$\neg v$	$f(n,\neg v)$	$f(\neg n,\neg v)$	$f(\neg v)$
Totals	$f(n)$	$f(\neg n)$	N

For example, if *n* is ‘look’ and *v* is ‘take’, then $f(n,v)$ is the number of tokens of verb-object combinations with *take* as verb and *look* as object; $f(\neg n,v)$ is the number of tokens of verb-object combinations where the verb is *take* and the object is not *look*; $f(\neg v)$ is the number of verb-object combinations where the verb is not *take*; and so on. From these numbers, estimated probabilities can be calculated: For example, $p(\neg n,v) = f(\neg n,v)/N$ is the estimated probability that a verb-object combination has *take* as verb and an object other than *look*, and $p(v|n) = f(v|n)/f(n)$ is the estimated probability that the verb is *take* given that the object is *look*.

Eight corpus statistics will be considered in this paper. Firstly, *token frequency* is simply $f(n,v)$.

The second and third statistics are measures of *unidirectional association*, i.e. how much is the noun attracted to the verb, and the verb to the noun? For the attraction of the verb to the noun, this is calculated using the Kullback-Leibler divergence (KLD) between the distribution of the verb given the noun and the unconditional distribution of the verb. The more dissimilar these two distributions are, the more highly the verb is attracted to or repelled from the noun:

$$KLD(v|n) = p(v|n)\log_2\frac{p(v|n)}{p(v)} + p(\neg v|n)\log_2\frac{p(\neg v|n)}{p(\neg v)}$$

Following Gries (2022a), this value is then normalised to fall between 0 and 1, with 0 being the lowest attraction and 1 being the highest attraction by applying the exponential function to -1 times the KLD and then subtracting the result from 1. In cases of repulsion, i.e. $p(v|n) < p(v)$, a negative sign is added in front of the negative KLD, so the final quantity ranges from -1 to 1. The formula for this value is as follows:

$$KLD_{norm}(v|n) = \operatorname{sgn}\big(p(v|n) - p(v)\big) \times (1 - e^{-KLD_{v\to n}})$$

The attraction of the noun to the verb is calculated similarly, just with *n* and *v* swapped in the formula. For example, in the construction *play truant*, *play* is highly attracted to *truant* (high verb-to-noun attraction), but *truant* is not highly attracted to *play* (low noun-to-verb attraction), since if we know the noun is *truant*, the verb much more likely to be *play* than most other nouns; but if we know the verb is *play*, it is very hard to guess the noun is *truant*.

The next four statistics all measure productivity: The degree to which verbs can combine with a variety of nouns, and vice versa. The fourth and fifth statistics are the type frequencies: the number of noun types that accompany each verb, denoted tf_v, and the number of verb types that co-occur with each noun, denoted tf_n. I take the logged values of both, i.e. $\log(tf_v)$ and $\log(tf_n)$.

The sixth and seventh statistics are entropy, which measures how unpredictable the noun is given the verb, and vice versa. Unlike type frequency, this measure also takes into account the relative prevalence of different collocates. For example, if one noun co-occurs with a single verb 99% of the time and 99 other verbs the remaining 1% of the time, its entropy would be nearly 0 even though the type frequency is 100. Unlike the conventional formula for entropy, the entropy used in this paper is normalised, following Gries (2022a), such that it cannot exceed 1. For the entropy of the verb given the noun, the entropy is normalised by the frequency of the noun:

$$H_{norm}(v|n) = \frac{-\sum_v p(v|n)\log_2 p(v|n)}{\log_2 f(n)}$$

The entropy of the verb given the noun is similarly calculated by swapping *v* and *n* in the formula.

The eighth and final statistic is DP_{nofreq} (Gries 2022c). This calculates how evenly distributed the verb-object combination is within the corpus. The first step in calculating this value is to get the raw dispersion statistic *DP*. To do this, we first calculate the proportion of instances of a verb-object combination, say *take* + *look*, that comes from each document in the corpus. We then calculate the proportion of verb-object combinations in general that comes from each document in the corpus. We then find the Manhattan distance between the two vectors of proportions. Next, we estimate the minimum and maximum values of DP given the token frequency of *take* + *look*. Finally, we calculate DP_{nofreq} by calculating its position within the range of possible values: the minimum value is 0, the maximum value is 1, and if the DP value is halfway between the minimum and maximum, then DP_{norm} is .5, and so on. Details of calculation are in Gries (2022c).

2.3 Light verb constructions

The particular application of corpus statistics and Phrase-BERT in this paper will be focused on the identification of light verb constructions (LVCs). A light verb construction is a grammatical construction consisting of a semantically light verb that contributes little to no predicational information and a lexical item, generally a nominal, which contributes the bulk of the information about the event or state being described. In English, a typical light verb construction consists of a verb followed by an indefinite object such as *take a peek* or *do backflips*. This paper will consider exclusively those LVCs that contain a noun.

Light verb constructions are studied in both corpus linguistics and NLP. They are a type of multi-word expression of great interest in both applied and theoretical linguistics: They are a common source of L2 errors because of their idiosyncratic properties (e.g. which verbs are paired with which nominals) (García Salido 2016), and their cognitive representation is a constant topic of interest, e.g. in English, they have the form of verb-object constructions, yet in some ways function like intransitive predicates (e.g. Wittenberg & Piñango 2011). It also has applications in NLP tasks like event extraction and information retrieval (Vincze et al. 2013), since the noun in an LVC should be treated as part of the predicate, rather than a participant in the event. Thus, extracting LVCs from corpora has many applications, such as for compiling computer- and/or human-readable glossaries of LVCs within a domain, for studying the grammatical properties of LVCs in L1 and L2 production, etc.

2.4 Phrase-BERT

As mentioned above, this paper uses Phrase-BERT (Wang et al. 2021) to classify constructions as LVCs. The main advantage of Phrase-BERT is that unlike most BERT-based approaches to calculating phrasal similarity, it is trained on collections of paraphrases such that phrases with similar meaning but no words in common will have similar embeddings, whereas words with overlapping words but very different meanings will have different embeddings. Thus, Phrase-BERT does not rely heavily on lexical overlap between phrases, and can better capture similarity between phrases that do not necessarily share words. As LVCs are a highly abstract category mostly characterised by how meaning is distributed in different parts of the construction, using Phrase-BERT can potentially make it easier to detect LVCs even if their component words do not appear in LVCs in the training data, and avoid mistakenly classifying non-LVCs as LVCs just because they share words with LVCs. This may be especially useful for detecting LVCs in L2 production, which may have less lexical overlap with LVCs in L1 data, but still share the semantic properties of LVCs.

2.5 Related work

To date, LLMs' most common uses in corpus linguistics are (a) using word embeddings to measure semantic similarity, which predates LLMs (Desagulier 2019, Tiun et al. 2020, etc.) and (b) using outputs generated from LLMs for automatic annotation (e.g. Weissweiler et al. 2024, Yu et al. 2024). Though this paper also uses LLMs to produce annotations, it uses word embeddings originating from LLM representations as predictors, rather than using LLM-generated output directly.

Concerning co-occurrence specifically, Uchida (2024) found that ChatGPT produces a collocation list that has 42.8% overlap with the list of collocations in the Corpus of Contemporary American English (COCA) created by selecting all collocations with mutual information (a bidirectional association measure) over 1, suggesting that ChatGPT's weights may encode some knowledge about co-occurrence of words (though the collocations may have also come from memorising collocation lists and dictionaries in the training data, rather than actually analysing co-occurrence between words).

In computational linguistics, Kanclerz & Piasecki (2022) has reintegrated statistical measures into MWE labelling; their approach, however, only uses bidirectional association measures to create lists of non-MWEs for negative training data. Thus, their co-occurrence statistics are not tupleised, and word embeddings and co-occurrence statistics are used at two different stages of their system for different purposes; they were not directly compared. To my knowledge, no work has attempted to compare word embeddings from LLMs to tupleised co-occurrence statistics.

3 Data

Three data sources were used for this study. Firstly, I took the verb-object constructions from the British National Corpus annotated by Tu & Roth (2011). This dataset includes the verbs *make*, *get*, *do*, *have*, *take*, *give*; around half were annotated as LVCs and half as non-LVCs. Secondly, I took annotations of OntoNotes 4.0 (Weischedel et al. 2011) from the latest version of PropBank (Bonial et al. 2014), which annotates for LVCs and other verb-object combinations. These two datasets were combined; to make the two comparable, the surrounding context of the LVCs, i.e. words before the verb or after the object, were discarded. Instances where the noun precedes the verb were also ignored. Dependency parses of the LVCs were used to extract the presence of dependencies like articles (*a*, *an*, etc.). An LLM-based disambiguation model (Wahle et al. 2021) was used to find the WordNet synset corresponding to the noun. The lexical file of the synset was then used as a semantic feature, dividing the nouns into categories like 'artifact', 'cognition', 'process', 'substance', 'animal', etc., similar to one of the features in Tu & Roth (2011). This dataset will be referred to as the LVC dataset.

For calculation of corpus statistics related to verb-noun constructions, the entire BNC was parsed using spaCy (Honnibal & Montani 2017) and all verb-direct object pairs were extracted. The eight statistics were then calculated. This dataset will be referred to as the VN dataset. Details of the construction of the datasets are in Appendix A.

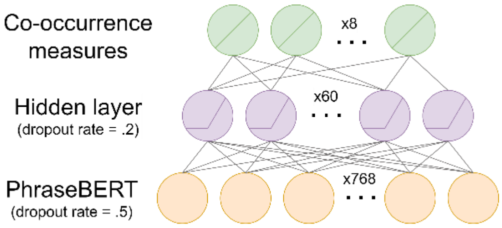

Figure 1: Architecture of the model used in Experiment 1.

4 Experiment 1: Predicting co-occurrence statistics from Phrase-BERT embeddings

The first experiment investigates whether information contained in corpus statistics is represented in Phrase-BERT in some form. This was done by attempting to predict corpus statistics from Phrase-BERT embeddings. If Phrase-BERT embeddings do contain information on association, entropy, etc., then these measures should be predictable from Phrase-BERT representations.

4.1 Methodology

A neural network (Figure 1) was used to predict co-occurrence statistics from Phrase-BERT embeddings. The model architecture consisted of an input layer containing all Phrase-BERT embeddings with dropout rate .5, a hidden layer of 60 units with ReLU activation and dropout rate .2, and finally eight output units with linear activation. The co-occurrence measures were centred and scaled before modelling, and a training-dev-test split of 8-1-1 was used. The model was implemented in Keras (Chollet et al. 2015).

4.2 Results & discussion

Figure 2 plots the predicted values from the neural network against the actual corpus statistics. As can be seen from the graph, although there are considerable deviations between the predicted and actual values of the co-occurrence statistics, the embeddings do have substantial predictive power overall. The mean squared error (calculated on the normalised corpus statistics) in the test set was .521. Were a curvilinear activation function employed, some of the predictions may be even more accurate. Moreover, it should be noted that some of the noise may come from noise in the co-occurrence statistics themselves, rather than in the ability of the embeddings to predict co-occurrence patterns. In sum, embeddings seem to encode some, though not necessarily all, of the information available in co-occurrence statistics.

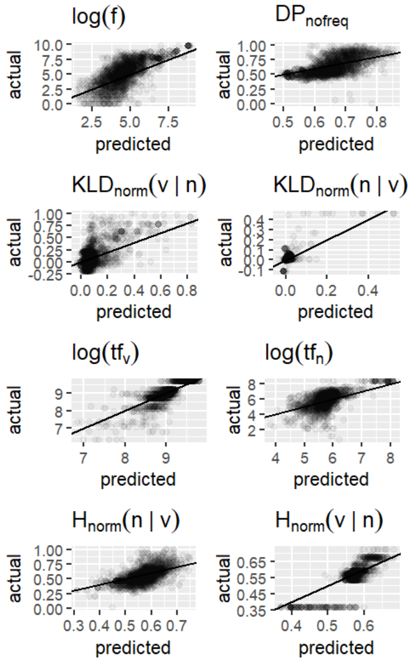

Figure 2: Predicted values of the corpus statistics using Phrase-BERT embeddings and actual values of the eight corpus statistics as calculated using the BNC. Only the test set is shown. Dots on the diagonal line have exactly equal actual and predicted values. The actual and predicted values are presented in their original scales, rather than the normalised version used in modelling.

5 Experiment 2: Relative contribution of BERT and co-occurrence statistics to light verb prediction

Since Experiment 1 found that word embeddings do encode information relevant to co-occurrence, one question is whether problems traditionally faced by corpus linguists that call for co-occurrence statistics can be solved by using word embeddings alone, or if co-occurrence measures still contain independent information that matter. In

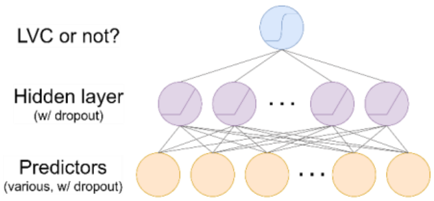

Figure 3: Architecture of the model used in Experiment 2.

Hyperparameter	Values
# of hidden layer units	15, 30, 45
Dropout rate for input layer	.2, .35, .5
Dropout rate for hidden layer	.2, .35, .5

Table 1: Hyperparameter values tested.

this section, we will consider the particular problem of extracting light verb constructions from a corpus. Imagine, for example, that we would like to teach light verb constructions in an L2 language instruction setting, and would like locate all light verb constructions in a set of level-appropriate texts to determine which readings would best serve the purpose. Would Phrase-BERT alone suffice to complete the job, or do we need traditional sources of information like co-occurrence statistics?

To answer this question, in this section, I aim to predict whether a phrase is a light verb construction from Phrase-BERT embeddings, corpus statistics, and both. If Phrase-BERT embeddings perform similar to or better than corpus statistics, and using both does not constitute an improvement over Phrase-BERT alone, then Phrase-BERT already contains all the useful information contained in the corpus statistics. If, on the other hand, using both sources of information is better than using Phrase-BERT alone, then this implies that corpus statistics contain useful information for LVC prediction that is not encoded in Phrase-BERT. I also run versions of these three models that add WordNet lexical files, dependency syntax information, or both, to see if any advantage of adding corpus statistics can be eliminated when semantic and/or syntactic information is added.

5.1 Methodology

The model trained in this section aims to predict whether a phrase is a light verb construction, based on the LVC dataset. Different combinations of predictors were used: I trained models using Phrase-BERT only, co-occurrence statistics only, or both, with syntactic information, semantic information, or both. Note that although both the corpus statistics and the Tu & Roth light verb judgements used the BNC, the Tu & Roth judgements were not involved in the calculation of corpus statistics, so there is no information leak.

The model architecture (Figure 3) consisted of an input layer containing the various variables, a hidden layer, and a sigmoid output layer for the choice between LVC vs non-LVC. Class weights were proportional to the reciprocal of the sample size of each class. Decision thresholds were tuned to maximise F1 using a grid search between 0 and 1 (exclusive) and a step size of .01. Grid search was used to determine the number of hidden layer units and dropout rates; all combinations of the values in Table 1 were tried, and for each combination of variables, I took the hyperparameter combination that resulted in the highest F1 in the validation set. As with Experiment 1, scaled and centred corpus statistics were used, and the training-dev-test split was 8-1-1.

Model	*P*	*R*	*F1*	*AUC*
BERT	0.937	0.970	0.953	0.955
STAT	0.910	0.935	0.922	0.835
BERT + STAT	0.950	0.964	**0.957**	**0.958**
BERT + SYN	0.951	0.958	0.954	0.952
STAT + SYN	0.898	0.969	0.932	0.846
BERT + STAT + SYN	0.953	0.960	0.956	**0.958**
BERT + SEM	0.946	0.961	0.953	0.949
STAT + SEM	0.901	0.961	0.930	0.870
BERT + STAT + SEM	0.940	**0.972**	0.956	0.955
BERT + SYN + SEM	**0.955**	0.948	0.952	0.954
STAT + SYN + SEM	0.915	0.955	0.935	0.890
BERT + STAT + SYN + SEM	0.951	0.958	0.955	0.956

Table 2: Results of Experiment 2 based on the test set. P = precision, R = recall, F1 = F1-value, AUC = area under the curve, BERT = Phrase-BERT embeddings, STAT = co-occurrence statistics, SEM = WordNet lexical files, SYN = noun modifiers' presence.

5.2 Results

Precision, recall, F1 and AUC values of all the models trained were shown in Table 2. Phrase-BERT alone performs substantially better than corpus statistics on all metrics. Yet when we combine both, the resulting model does better on all metrics but recall compared to the model with BERT alone. This pattern (adding statistics improves most metrics) largely persists even after adding syntactic dependencies and/or semantic categories to the model, though the model with just BERT and statistics remains the best model in terms of F1. Thus, co-occurrence statistics contain useful information beyond what is encoded in Phrase-BERT, syntactic dependencies on the noun, and WordNet lexical files.

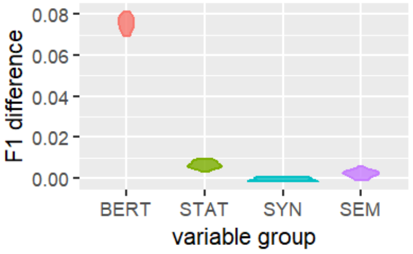

Figure 5: Permutation variable importance of the four variable groups, as calculated by drop in F1 after shuffling the relevant variable group.

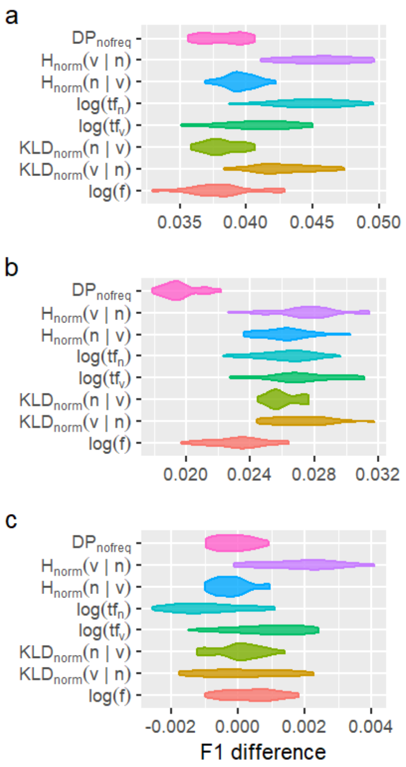

Figure 4: Permutation variable importance of the tupelised co-occurrence statistics in (a) the STAT model, (b) the STAT + SYN + SEM model, (c) the BERT + STAT + SYN + SEM model. Note that the x-axis is different in each graph, with the scale of the x-axis in (c) much smaller than (a) and (b).

5.3 Discussion

To examine how important corpus statistics were, I used a permutation variable importance approach on the maximal model. I randomly shuffled the values of each of all four groups of variables, and examined the impacts on the F1 in the test set. I did this reordering 20 times per variable group. As seen in Figure 5, the biggest drop in F1 by far came from reordering BERT, but reordering corpus statistics still resulted in a rather more substantial drop in performance than the semantic or syntactic variables. This suggests that corpus statistics have a small, but still substantial contribution towards the predictive power of the model.

But which statistics exactly are still important in this full model, i.e. are not captured by Phrase-BERT or by the syntactic and semantic properties? I repeated the permutation variable importance process, but this time shuffling each statistic independently, for three models: (a) statistics only, (b) statistics with syntax and semantics, (c) statistics with BERT, syntax and semantics (Figure 4). Going from model (a) to (b), there is a drop in all of the variables' importance, but all of them still matter, so semantics and syntax only capture a small part of the useful information from co-occurrence statistics. Unsurprisingly, once BERT is added, all the statistics' importance drop drastically, though $H_{norm}(v|n)$ remains important.

To further examine how exactly co-occurrence statistics contribute to better predictions in qualitative linguistics terms, I qualitatively compared the predictions of the full model (c) with the model with everything but co-occurrence statistics (hereafter the no-stats model). I looked at cases in which one model got something wrong that the other got right.

Firstly, I looked at cases of phrases labelled as non-LVCs in the original dataset but one of the two models judge as an LVC. These cases are especially important as the two models differ substantially in precision. Phrases that were classified as false positives in the full model and true negatives in the no-stats model often seem to be mislabelled in the original data or edge cases, e.g. *take effect* or *do some work* (many similar phrases were counted as LVCs in the data). On the other hand, if we look at the opposite situation – phrases that were false positives in the no-stats model but true negatives in the full model – there were fewer apparently mislabelled items. Instead, many were clear non-LVCs where the verb is seemingly light (and is light in many other contexts), but in the specific phrase retains the non-light lexical meaning, e.g. *made a profound impression* (where the verb indicates the subject is actually creating something) or *get credit* (where the subject metaphorically receives something). In these cases, the useful contribution from corpus statistics likely comes from the ability to relate the noun to the verb rather than considering them separately. For example, *get* is a frequent verb often appearing in LVCs and *credit* is an abstract noun, which are often associated with LVCs. So looking at *get* and *credit* separately, one may be tempted to classify this as an LVC. But the noun is not strongly attracted to the verb (z-score of $KLD_{norm}(n|v) = -.21$). Out of the 815 input variables, the most negative Shapley value is $KLD_{norm}(n|v)$ (Shapley value = -0.03), suggesting that it was a major factor that pushed the maximal model to treat this phrase as non-LVC. This suggests such information was not encoded as well in Phrase-BERT alone.

I then examined cases where phrases labelled as LVCs in the original dataset were classified as non-LVCs by one of the two models. Very few phrases were false negatives in the full model but true positives in the no-stats model. There were no clear patterns in phrases that were false negatives in the no-stats model but true positives in the full model, except that they sometimes have less frequent nouns, like *booking* (seen once training data) or *injection* (seen twice). This is unsurprising given that the models are close in terms of recall.

Of course, these results do not imply that corpus statistics are always needed on top of Phrase-BERT for LVC classification. I did not consider the context surrounding the LVCs, so I do not know whether Phrase-BERT better captures surrounding contextual information than corpus statistics like previous and next word entropy (Zhào et al. 2016). Moreover, the workflow for my system requires the user to first locate candidate verb-object combinations, rather than getting a list of LVCs from a raw text corpus; statistics may be hard to use in this situation. Still, the results suggest that corpus statistics remain relevant in at least *some* situations relevant to the corpus linguist.

5.4 Follow-up experiment

Since Experiment 2 found that much of the useful information in corpus statistics is found in Phrase-BERT, one may ask *how* Phrase-BERT uses this implicit co-occurrence information to make predictions about LVC membership. To do this, I used the syntactic, semantic predictors and co-occurrence statistics to predict the behaviour of the BERT-only model. Again, a neural network with a single ReLU hidden layer of 15 units was used, with the same dropout rates as Experiment 1. The output layer has linear activation, and predicts the estimated probability from the BERT-only model, with a logit transformation applied to the probability so that it can be any real number.

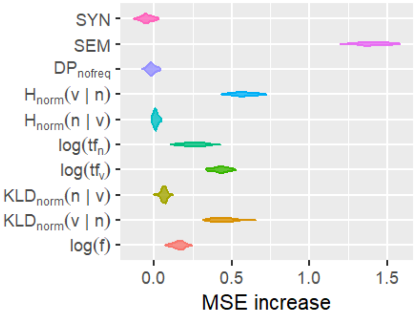

Figure 7: Permutation variable importance of the statistics in the follow-up experiment, as calculated by drop in F1 after shuffling the relevant variable group.

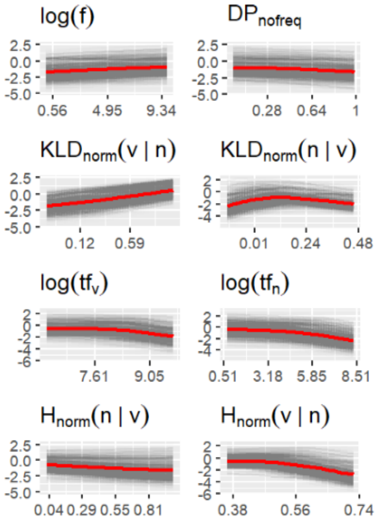

Figure 6: Partial dependency plots of the six statistics in the test set. Note that these are based on z-scores, not original values.

Permutation variable importance (Figure 7) shows that WordNet semantic information is the most important, and as before, $KLD_{norm}(v|n)$, $H_{norm}(v|n)$ and the type frequencies stand out as the most important predictors based on co-occurrence statistics. To see the exact way in which statistical information encoded in Phrase-BERT is used to predict light verb construction status, partial dependency plots of the relationship between the statistics and the prediction of the BERT-only model are shown in Figure 6. The strongest relationships are: $KLD_{norm}(v|n)$ (i.e. the verb's attraction to the noun) is positively associated with LVC status, while $KLD_{norm}(n|v)$ is positively associated for very low values but negatively associated elsewhere. These results can be interpreted as Phrase-BERT having learnt that in LVCs, the verbs are generally strongly attracted to the noun, and the nouns are somewhat, but not very, attracted to the verb. The productivity of the noun with respect to the range verbs it combines with, as measured by type frequency and entropy, is also negatively associated with LVC status.

These results may be compared to those obtained for Tibetan LVCs in Lai (in press). However, there are several important differences between the two studies. Firstly, in this paper, noun-verb combinations are investigated regardless of frequency, whereas in Lai (in press), only combinations with the highest frequency were taken. Secondly, in this study, only verbs that appear in at least one LVC are considered, whereas Lai (in press) makes no such restriction.

The relationship between $KLD_{norm}(v|n)$ and $KLD_{norm}(n|v)$ and LVC status is mostly in accord with the Tibetan findings. The initial positive relationship between $KLD_{norm}(n|v)$ and LVC status found here is absent from the Tibetan study, likely because low-frequency noun-verb combinations were not considered there. Lower entropy of the verb slot given the noun $H_{\text{norm}}(v|n)$ and type frequency of the noun $\log(tf_n)$ being associated with LVC status is also consistent with the Tibetan findings. In the Tibetan study, higher values of $H_{\text{norm}}(n|v)$ and $\log(tf_v)$ were visually found to be associated with LVC status (though the statistical test was insignificant), contrary to the weak negative association found here. This small difference, however, does not necessarily indicate a typological difference, as it can likely be attributed to the fact that the present study excludes verbs that never appear in LVCs: such verbs were likely absent from LVCs precisely because they appear with fewer nouns, and their inclusion would have tipped the scales the other way.

6 Conclusion

In this study, we showed that a considerable amount of information in co-occurrence statistics is encoded in Phrase-BERT, though not all. We saw that tupleised corpus statistics only do slightly worse than Phrase-BERT at predicting whether a verb-object combination is an LVC, and moreover, the statistics have an independent contribution to LVC detection beyond information also encoded in Phrase-BERT, mostly coming from $H_{norm}(v|n)$, the normalised entropy of the verb slot for each noun. Finally, corpus statistics can be used to partially interpret how Phrase-BERT identifies LVCs. Indeed, the patterns found through this analysis largely accord with findings in Lai (in press) for Tibetan, showing that the power and robustness of tupleised corpus statistics for LVC detection crosslinguistically. Importantly, this would not be possible in a traditional single-statistic approach, which would not capture e.g. the fact that noun-to-verb attraction is mostly *negatively* associated with LVC status but verb-to-noun attraction is *positively* associated.

Thus, tupleised corpus statistics can aid in interpreting black-box systems and improving the performance of such systems when added as additional predictors. Tupleisation contributes to the lasting relevance of co-occurrence statistics for corpus linguists in the age of LLMs.

References

Belinkov, Yonatan, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi & James Glass. 2017. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In Greg Kondrak & Taro Watanabe (eds.), *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1–10. Taipei, Taiwan: Asian Federation of Natural Language Processing. https://aclanthology.org/I17-1001.

Bonial, Claire, Julia Bonn, Kathryn Conger, Jena D. Hwang & Martha Palmer. 2014. PropBank: Semantics of new predicate types. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis in the sample (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3013–3019. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/1012_Paper.pdf .

Chollet, François et al. 2015. Keras. https://keras.io.

Desagulier, Guillaume. 2019. Can word vectors help corpus linguists? *Studia Neophilologica* 91(2). 219–240. https://doi.org/10.1080/00393274.2019.1616220.

Dras, Mark & Michael Johnson. 1996. Death and lightness: using a demographic model to find support verbs. In *International Conference on the Cognitive Science of Natural Language Processing (5th: 1996)*. Dublin City University Natural Language Group.

Evert, Stephanie. 2005. *The statistics of word cooccurrences: word pairs and collocations*. Germany: Universität Stuttgart PhD Thesis.

García Salido, Marcos. 2016. Error analysis of support verb constructions in written Spanish learner corpora. *The Modern Language Journal* 100(1). 362–376.

Gries, Stefan Th. 2022a. Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach. *Lexis* (19). https://doi.org/10.4000/lexis.6231.

Gries, Stefan Th. 2022b. What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies* 5(1). 1–33. https://doi.org/10.1075/jsls.21028.gri.

Gries, Stefan Th. 2022c. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies* 5(2). 171–205.

Gries, Stefan Th. 2024. *Frequency, Dispersion, Association, and Keyness: Revising and tupleizing corpus-linguistic measures* (Studies in Corpus Linguistics). Vol. 115. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.115.

Honnibal, Matthew & Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen & Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1195–1205. New Orleans, Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1108.

Kanclerz, Kamil & Maciej Piasecki. 2022. Deep Neural Representations for Multiword Expressions Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 444–453. Dublin, Ireland: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-srw.36.

Jawahar, Ganesh, Benoît Sagot & Djamé Seddah. 2019. What does BERT learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics.

Leech, Geoffrey Neil. 1992. 100 million words of English: the British National Corpus (BNC). 어학연구. 서울대학교 언어교육원.

Lai, Ryan Ka Yau (in press). Beyond bidirectional association: Distinguishing light verb constructions from other conventionalised noun-verb combinations in modern Tibetan. In Jens Fleischhauer & Anna Riccio. (eds.), *Light verb constructions from a cross-linguistic perspective*. Berlin: Mouton de Gruyter.

Lin, Yongjie, Yi Chern Tan & Robert Frank. 2019. Open Sesame: Getting inside BERT's Linguistic Knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 241–253. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4825.

Piasecki, Maciej & Kamil Kanclerz. 2022. Non-Contextual vs Contextual Word Embeddings in Multiword Expressions Detection. In Ngoc Thanh Nguyen, Yannis Manolopoulos, Richard Chbeir, Adrianna Kozierkiewicz & Bogdan Trawiński (eds.), *Computational Collective Intelligence* (Lecture Notes in Computer Science), vol. 13501, 193–206. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-16014-1_16.

Stefanowitsch, Anatol & Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1). 1–43. https://doi.org/10.1515/cllt.2005.1.1.1.

Tan, Yee Fan, Min-Yen Kan & Hang Cui. 2006. Extending corpus-based identification of light verb constructions using a supervised learning framework. In Paul Rayson, Serge Sharoff & Svenja Adolphs (eds.), *Proceedings of the Workshop on Multi-word-expressions in a multilingual context*, 49–56. Trento, Italy: Association for Computational Linguistics.

Tayyar Madabushi, Harish, Laurence Romain, Dagmar Divjak & Petar Milin. 2020. CxGBERT: BERT meets Construction Grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4020–4032. International Committee on Computational Linguistics (ICCL). https://doi.org/10.18653/v1/2020.coling-main.355.

Taslimipoor, Shiva & Omid Rohanian. 2018. SHOMA at PARSEME shared task on automatic identification of VMWEs: Neural multiword expression tagging with high generalisation. arXiv. https://doi.org/10.48550/ARXIV.1809.03056.

Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. *In International Conference on Learning Representations*. New Orleans, Louisiana: Association for Computational Linguistics.

Thrush, Tristan, Ethan Wilcox & Roger Levy. 2020. Investigating novel verb learning in BERT: Selectional preference classes and alternation-based syntactic generalization. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 265–275. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.blackboxnlp-1.25.

Tiun, Sabrina, Saidah Saad, Nor Fariza Mohd Noor, Azhar Jalaludin & Anis Nadiah Che Abdul Rahman. 2020. Quantifying semantic shift visually on a Malay domain specific corpus using temporal word embedding approach. *Asia-Pacific Journal of Information Technology and Multimedia* 09(02). 1–10. https://doi.org/10.17576/apjitm-2020-0902-01.

Tu, Yuancheng & Dan Roth. 2011. Learning English light verb constructions: contextual or statistical. In Kordoni Valia, Carlos Ramicsh & Aline Villavicencio (eds.), *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, 31–39. Portland, Oregon, USA: Association for Computational Linguistics.

Uchida, Satoru. 2024. Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. *Applied Corpus Linguistics* 4(1). 100089. https://doi.org/10.1016/j.acorp.2024.100089.

Van Hoey, Thomas. 2023. ABB, a salient prototype of collocate–ideophone constructions in Mandarin Chinese. *Cognitive Linguistics* 34(1). https://doi.org/10.1515/cog-2022-0031.

Vincze, Veronika, István Nagy T. & János Zsibrita. 2013. Learning to detect English and Hungarian light verb constructions. *ACM Transactions on Speech and Language Processing* 10(2). 1–25. https://doi.org/10.1145/2483691.2483695.

Wahle, Jan Philip, Terry Ruas, Norman Meuschke & Bela Gipp. 2021. Incorporating word sense disambiguation in neural language models. arXiv preprint arXiv:2106.07967.

Wang, Shufan, Laure Thompson & Mohit Iyyer. 2021. Phrase-BERT: Improved phrase embeddings from BERT with an application to corpus exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10837–10851. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.846.

Waszczuk, Jakub, Rafael Ehren, Regina Stodden & Laura Kallmeyer. 2019. A neural graph-based approach to verbal MWE identification. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 114–124. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5113.

Weischedel, Ralph, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, et al. 2011. OntoNotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium* 17.

Weissweiler, Leonie, Abdullatif Köksal & Hinrich Schütze. 2024. Hybrid human-LLM corpus construction and LLM evaluation for rare linguistic phenomena. *arXiv preprint arXiv:2403.06965*.

Wittenberg, Eva & Maria Mercedes Piñango. 2011. Processing light verb constructions. *The Mental Lexicon* 6(3). 393–413. https://doi.org/10.1075/ml.6.3.03wit.

Yu, Danni, Luyang Li, Hang Su & Matteo Fuoli. 2024. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies. *International Journal of Corpus Linguistics*. John Benjamins Publishing. https://doi.org/10.1075/ijcl.23087.yu.

Zhao, Weina, Lin Li, Huidan Liu & Jian Wu. 2016. Tibetan trisyllabic light verb construction recognition. *Himalayan Linguistics* 15(1). https://doi.org/10.5070/H915130102.

A Details of data extraction

To create the LVC dataset, Tu & Roth's data was used as-is, with no modifications except replacing the underscores with spaces. To extract non-LVC verb-object combinations from PropBank, I looked for verbs (`pos` = `V`), and then looked for an `ARG1` whose constituency tree representation starts with `(NP` in the corresponding proposition. To extract LVC verb-object combinations, I looked for verbs again, but this time looked for a word labelled `ARGM-PRR` which indicates it is the head of a light verb nominal. If this is not immediately adjacent to the verb, then the closest word to the `ARGM-PRR` whose constituency tree representation starts with `(NP` is considered the start of the light verb nominal. Otherwise, the word itself is considered the entirety of the light verb nominal.

Phrase-BERT representations of the examples of the LVC dataset were computed for the string of words starting with the verb and ending in the light verb nominal, including anything in between, such as indirect object pronouns (e.g. *throw **them** a curveball*).

The object nominals were dependency-parsed and dependents on the object were extracted, including *a*, *the*, *no*, *some*, *any*, *good*, *this*, *little*, *more*, *great* and *first*. The syntax features used in this paper are simply Boolean features indicating the presence of these words.

The WordNet lexical files were based on the head of the object alone. I used `nltk` to get the synsets corresponding to the head, and then used Wahle et al.'s model to find the most appropriate meaning given the LVC instance. A sample input is as follows:

```
question:    which    description
describes the word " explanation "
best in the following context? \
descriptions:[  " a statement that
makes something comprehensible by
describing the relevant structure
or operation or circumstances etc.
", " thought that makes something
comprehensible ", or " the act of
explaining; making something plain
or intelligible " ]
context: gave us an " explanation
" .
```

I then took the lexical file of the synset whose definition was deemed most appropriate.

To create the VN dataset, sentences were first extracted from the HTML version of the BNC. Then I used spaCy to dependency-parse and lemmatise everything in the corpus. Direct objects (`dobj`) and passive subjects (`nsubj:pass`) were extracted from the corpus along with their verbal heads. Statistics were then calculated based on extracted verb-object combinations.