On Quick Kisses and How to Make Them Count: A Study on Event Construal in Light Verb Constructions with BERT

Chenxin Liu and **Emmanuele Chersoni**
The Hong Kong Polytechnic University
Department of Chinese and Bilingual Studies
Yuk Choi Road 11, Hung Hom, Kowloon, Hong Kong (China)
chenxinl@uw.edu, emmanuelechersoni@gmail.com

Abstract

Psycholinguistic studies suggested that our mental perception of events depends not only on the lexical items used to describe them, but also on the syntactic structure of the event description. More specifically, it has been argued that light verb constructions affect the perception of duration in event construal, such that the same event in this type of constructions is perceived by humans as taking less time (*to give a kiss* takes a shorter time than *to kiss*).

In our paper, we present two experiments with BERT using English stimuli from psycholinguistic studies to investigate the effects of the syntactic construction on event duration and event similarity. We show that i) the dimensions of BERT vectors encode a smaller value for duration for both punctive and durative events in count syntax, in line with human results; on the other hand, we also found that ii) BERT semantic similarity fails to capture the conceptual shift that durative events should undergo in count syntax.

1 Introduction

Temporality of event representations is at the core of human cognition, as the event duration in linguistic descriptions is likely correlated to the way we represent time in our mind (Coll-Florit and Gennari, 2011). When we talk about an event, whether it is a hug between friends, an academic speech or the advice given by a doctor after a visit, we generally do not make explicit the duration information, since this information is supposed to be known by everyone from personal experience. Moreover, duration is usually not encoded by grammar.

However, previous studies in psycholinguistics suggested that grammatical cues in the message significantly affect the event representation built by language comprehenders for aspects such as causation and event structure (Fausey and Boroditsky, 2010; Johnson and Goldberg, 2013). One of the main hypotheses of our research work is that the perceived duration of an event is also one of the aspects that is influenced by grammar.

Events can be individuated or not (Barner et al., 2008; Wellwood et al., 2018): *She does runs* and *She does running* describe something similar, but the **count syntax** in the first sentence makes us think about several occurrences of an activity, while the **mass syntax** in the second one makes us think about a more generic action whose temporal boundaries are not specified [1]. Moreover, some events are more easily thought as **atomic**: when we hear that *Mary kissed John* we can easily imagine that she did it several times, and each time could count as a separate *kissing*-event. On the other hand, *speaking* is **non-atomic**: if we say that *The king spoke to the soldiers of the army*, we might be plausibly describing a situation where the king had a break in his speech and then he started again, but this would still count as the same *speaking*-event.

Atomicity is the main criterion to classify events into **punctive** (like *kissing*) or **durative**. Punctive events refer to verbs that tend to be instantaneous and are usually bounded by a natural end point (e.g. in *Mary kissed John*, the set start and end point come with the contact and the separation of two lips). Besides, they often receive iterative reading when taking place in progressive form or over a more protracted duration (e.g. by reading *Mary kicked the table for an hour*, one could imagine that there should be more than one kicking action). Durative events usually are not naturally bounded and not understood iteratively. For instance, the sentence *Mary talked for an hour* describes a single protracted event rather than multiple events.

A study by Wittenberg and Levy (2017) analyzed

[1] According to the Number Asymmetry hypothesis by Barner and Snedeker (2005), the count syntax uses number as the uniform dimension of measurement (e.g., *two cups*, *a dance*, *a jump*). By contrast, the mass syntax is unspecified and open to comparison using various measuring dimensions, such as mass, volume and time (e.g., *some water*, *some dancing*, *some jumping*).

Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 367–378
December 7, 2023. ©2023 Association for Computational Linguistics

the interaction of mass-count syntax with punctive and durative events when they are used within a light verb construction (e.g. *to give a kiss, to give a speech*). In the first experiment, the author asked human subjects to estimate the duration of events, and the events were denoted either by a verb in transitive construction, or by a corresponding light verb construction[2], and the results revealed that **light verb syntax has a shortening effect on the perceived duration of the event** (*to give a kiss* takes a shorter time than *to kiss*). Moreover, in a second experiment, human raters had to decide whether two events occurring in two different constructions were the same or not. Noticeably, while an event in the transitive or in the light verb construction was rated to be similar to itself in the cases of punctive events (e.g. *to kiss* vs. *to give a kiss*) in count syntax and durative events in mass syntax (e.g. *to advise* vs. *to give advice*), **durative events undergo a significantly greater semantic shift when they are described with count syntax**. In other words, a higher semantic distance is perceived between event descriptions such as *to talk* and *to give a talk*. The event in the light verb construction, although it is still conceptually related to the one in the transitive frame, is conceived as of a different type.

Event and temporal knowledge are useful for many Natural Language Processing (NLP) applications, including information retrieval, story generation, question answering and text summarization (Zhou et al., 2020; Ma et al., 2021). Performance in NLP applications vastly improved with the advent of Language Models (LMs) based on Transformer architectures (Vaswani et al., 2017; Devlin et al., 2019), and consequently, a lot of research focused on analyzing their linguistic abilities, and questioning whether the representations they learn are compatible with linguistic theory (Li, 2022). We would like to verify, therefore, whether the semantic representations of LMs are able to capture changes in the representations of time that are encoded in subtle variations of the linguistic input.

In our study, we used the popular BERT language model (Devlin et al., 2019) to reproduce the experiments of Wittenberg and Levy (2017). First, we analyzed the contextualized vector representations produced by different versions of BERT, and used the technique of *semantic projection* (Grand et al., 2022) to test if the value of the semantic dimension of duration is actually shorter for events described with light verb constructions than in the corresponding ditransitive constructions; secondly, we measured the semantic similarity between the same events in the two constructions, to see if the distances between vectors capture the same meaning shifts that have been detected by humans.

In our first experiment, we found that the shortening effect can be found across event types and projection conditions, with the contextualized vector of the light verb construction showing significantly shorter duration values. In the second experiment, however, the similarity between BERT vectors largely fails to reproduce the pattern observed in the original study, as no significant meaning shifts were observed across event categories. We hypothesized this might be due to an inherent shortcoming of distributional similarity in distinguishing between fine-grained meaning relations between linguistic expressions (Baroni and Lenci, 2011; Xiang et al., 2020; Schulte Im Walde, 2020). To the best of our knowledge, our study is the first to study meaning shifts in light verb constructions with contextualized vector spaces, and in general one of the first to analyze the representation of duration in distributional models. The materials to reproduce our experiments can be found at `https://github.com/xinxinlaoshi/QuickKisses`.

2 Related Work

2.1 Probing Linguistic Knowledge in Language Models

A large number of studies in the literature on language models (LMs) has been dedicated to the analysis of the linguistic knowledge that they encode. The most popular methodology is probably the one employing *probing tasks*, in which a simple model is asked to solve a task requiring linguistic knowledge using a representation derived from a LM, with little or no specific linguistic supervision. If the model achieves a good performance, then one can infer that the LM representation encodes the target linguistic knowledge (Tenney et al., 2019a,b; Hewitt and Liang, 2019; Wu et al., 2020; Vulić et al., 2020; Sorodoc et al., 2020; Ettinger, 2020; Geiger et al., 2021; Koto et al., 2021; Chersoni et al., 2021a; Conia and Navigli, 2022; Kim and Linzen, 2020; Arps et al., 2022; Misra et al., 2022).

[2] Since *give* in the light verb construction is ditransitive, i.e. taking both a direct object and an indirect object, "light verb construction" and "ditransitive frame" were used interchangeably to refer to the same syntactic construction in Wittenberg and Levy (2017). This paper follows this usage.

Some previous computational work specifically investigated how grammatical cues of sentence inputs impact the predictions of Transformer language models. Cho et al. (2021) studied the priming effect of verb aspect on BERT predictions of event locations in English, and they found that BERT correctly assigns higher probability scores to typical event locations, but that it is not particularly affected by the aspect. Humans, on the other hand, activate specific expectations for event locations only when the verbs describing those events are in the imperfective form (e.g. the location in *The boy was fishing at the lake* is more salient than in *The boy had fished at the lake*, as the event is represented as still ongoing). The work by Metheniti et al. (2022) focused again on the BERT model and on the aspectual features of telicity and duration. Their setup included a classification task in English and French, and their results proved that in both languages BERT was adequately capturing information on telicity and duration, even in the non-finetuned forms, although it also showed some bias to verb tense and word order.

2.2 Modeling Conceptual Shifts in Computational Semantics

The phenomena of coercion and metonymic interpretation have widely been investigated in NLP, either with classical distributional models (Zarcone and Padó, 2011; Zarcone et al., 2012; Chersoni et al., 2017; McGregor et al., 2017; Chersoni et al., 2021b) or with Transformer-based language models (Rambelli et al., 2020; Pedinotti and Lenci, 2020; Ye et al., 2022; Gu, 2022). Most studies focused on *complement coercion*, a type clash between an event selecting verb and an entity denoting noun, that triggers a hidden event interpretation (e.g. *The composer began the symphony* → *The composer began* ***writing*** *the symphony*).

Some works focused instead on the mass-count coercion in nominals. Katz and Zamparelli (2012) considered pluralisation as a proxy of count usage (e.g. *wine* (mass noun) → *wines* (count usage), which is more likely to refer to *glasses of wine* rather than to the liquid), and built a vector space model with separate vector representations for the singular and the plural of a list of candidate mass and count nouns. Consistently with their initial hypothesis, they found that the vector similarity between singular and plural is higher for count nouns than for mass nouns, since the latter undergo a meaning shift when they are pluralized. The follow-up work by Hürlimann et al. (2014) analyzed the factors affecting the similarity scores in the data by Katz and Zamparelli (2012), reporting that abstract and highly polysemous nouns undergo greater semantic shifts as a consequence of pluralization. Finally, Liu and Chersoni (2022) used BERT vectors to study the meaning shift of coercion, and they found that mass noun vector have more pronounced shifts (i.e. a lower similarity between token vectors) when used in count contexts.

A recent work by Chronis et al. (2023) combined the analysis of meaning shifts with interpretability, by using regression to map distributional vectors on interpretable feature spaces. One of their case studies is focusing on the Article + Adjective + Numeral + Noun construction (e.g. *a beautiful three days in Rome*), where the noun modified by the numeral, in virtue of the event construal associated with this construction, behaves as a single collective unit. The authors showed that, indeed, the BERT representations of the nouns in the context of this construction assign more prominent values to measure- and unit-related semantic features.

3 The Study by Wittenberg and Levy (2017): Effects of Light Verb Constructions on Event Duration and Similarity

The goal of Wittenberg and Levy (2017) was to investigate whether describing an event in mass or count syntax with a light verb construction affected the construals of event duration and similarity in comprehenders. The authors built sentences for three groups of verbs (see also Table 1 for a schematic illustration of the findings): a) punctive events described with transitive verbs (e.g. *to kiss*) vs. described in count syntax with light verb constructions (e.g. *to give a kiss*); b) durative events described with transitive verbs (e.g. *to advise*) vs. described in mass syntax with light verb constructions (e.g. *to give advice*); c) durative events described with transitive verbs (e.g. *to talk*) vs. described in count syntax with light verb constructions (e.g. *to give a talk*).

The sentences were built to insert the event descriptions in natural contexts, and in each verb group, they differed only by construction type (e.g. *After their first date, Douglas* ***kissed*** *Mary* vs. *After their first date, Douglas* ***gave a kiss*** *to Mary*). In the Experiments 1-2, human participants were asked

Findings/Event Type	**Punctive Count** (kiss → give a kiss)	**Durative Mass** (advise → give advise)	**Durative Count** (talk → give a talk)
Event Duration	Shorter in constructions	Non significant	Non significant
Event Similarity	Event semantically similar	Event semantically similar	Event conceptually different

Table 1: Table adapted from Wittenberg and Levy (2017), with a summary of the findings of the study (Experiments 1-2 for event duration, Experiment 4 for event similarity).

to read the sentences and estimate how long the described event probably took. [3] While the results suggested that light verb constructions are generally associated with shorter durations in both experiments, consistent significant effects were found only in punctive events with count syntax (the tendency was present also for durative mass events, but findings were less consistent across settings).

In Experiment 4, the participants were asked to indicate the semantic similarity between the same event in transitive and ditransitive construction on a 7-point Likert scale. The durative count pairs were rated significantly less similar to each other than the punctive count and durative mass pairs.

Combing the results from these experiments, the authors suggested that the light-verb encoding with count/mass syntax can also lead to a change in the general construal of an event, besides the shortening effect on the event duration. They argued that durative events are similar to mass nouns in terms of atomicity, as mass nouns like *milk* can also be partitioned arbitrarily, and one can get two portions of exactly the same substance by dividing (Cheng, 1973; Link, 1983; Rothstein, 2017). When mass nouns like *wine* or *iron* occur in a count context, the denotation of the resulting count noun phrase (several wines, an iron) is expected to change from the substance to another object that is different from but arbitrarily related to the substance, such as various types of wine or a piece of flatiron.

Given the analogy between the durative events and mass nouns, a similar conceptual shift would also occur when durative verbs are represented in the form of deverbal nouns with count syntax. For example, it is intuitive to think that *giving a speech*, while keeping the core meaning of utterance, is conceptually further from *speaking* than *giving a hug* is from *hugging*. However, the direction of the shift depends on the context and is hard to predict, thus any change in duration can be coincidental. Therefore, the count light-verb encoding in this case resulted in a significant meaning difference rather than in a consistent shortening effect.

4 Experiment 1: Modeling Event Duration with Semantic Projections

The objective of our first experiment was to analyze the feature of duration in the embedding representations of events in language models and compare the results with the findings of Experiments 1-2 in Wittenberg and Levy (2017).

But how to quantify duration in BERT embeddings? For this goal, we adopted the semantic projection technique introduced by Grand et al. (2022). The authors of the study suggested that one can infer semantic properties of objects and entities as semantic subspaces in a distributional model. Subspaces were found with the following procedure: 1) identify multiple words that can represent extreme values of those properties on a scale, e.g. for SIZE they could be *big, huge, gigantic* on one extreme, and *tiny, small, minuscule* on the other extreme; 2) average the word vectors at the two extremes, and then connect the two extremes with a line. This line will represent the human mental scale for SIZE; 3) given a list of words/concepts to be ordered by their SIZE, project their embeddings onto the SIZE line and take the relative ordering of their values. Applying this simple method to GloVe embeddings (Pennington et al., 2014), the authors were able to predict human judgements across different semantic categories and for different types of properties (e.g. TEMPERATURE, SPEED, AROUSAL, INTELLIGENCE etc.). Our idea is to apply the same method to contextualized embeddings representing concrete usages of the verbs and the constructions from Wittenberg and Levy (2017), by projecting them onto a DURATION subspace. To generate the embeddings, we used the popular BERT model (Devlin et al., 2019), in its base, uncased version for English.

The target events are adapted from those identified by Wittenberg and Levy (2017), and our dataset consisted of descriptions of those events either in bare verb forms or in ditransitive construc-

[3] Notice that the experiments differed in the response options offered to the participants: in one case they were open estimates, in the other they were predefined time bins. Nonetheless, the findings were consistent across settings.

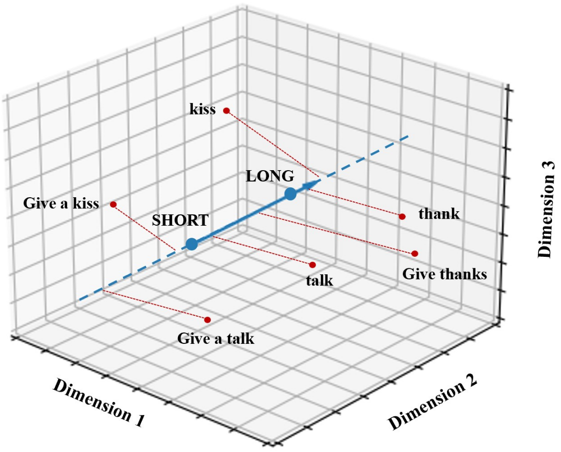

Figure 1: Illustration of semantic projection for the verbs *to kiss, to thank, to talk* and corresponding light verb constructions.

tions led by *give*. There are 3 categories of phrase pairs in total:

1a. Punctive events in count syntax: *kiss – give a kiss, hug – give a hug, kick – give a kick, shake – give a shake, cuddle – give a cuddle, wink – give a wink*;

1b. Durative events in count syntax: *talk – give a talk, address – give an address, lecture – give a lecture, present – give a presentation, speak – give a speech, check – give a check*;

1c. Durative events in mass syntax: *advise – give advice, thank – give thanks, assure – give assurance, encourage – give encouragement, recognize – give recognition, support – give support*.

We first extracted sentences in which the target phrases occurred from the British National Corpus (BNC) and obtained a total of 161,752 sentences. Generally, all three categories of events have higher frequency when occurring as a transitive verb, and punctive count events have the lowest frequency in both transitive[4] and ditransitive contexts (detailed statistics for context extraction can be found in Table 5, in the Appendix).

For each target event, we sampled 40 sentences where it occurs in a transitive context and 40 sentences where it occurs in a ditransitive context from the British National Corpus (Leech, 1992), then generated their vector representation via the BERT architecture, using the MINICONS library[5] (Misra, 2022). For transitive verb sentences, we extracted the embeddings of the verbs from the last layer; for light verb constructions, we used the embedding of the nominal (e.g. we used the contextualized $\overrightarrow{kiss}$ vector to represent *to give a kiss*).

Short	Long
Adjectives: brief, short, immediate, short-term. **Nouns**: minute, moment, second.	**Adjectives**: long, long-term, lengthy. **Nouns:** ages, years, decades, centuries

Table 2: List of words representing the extremes of the DURATION scale.

To realize the semantic projection, we followed (Grand et al., 2022) in projecting the vectors of our sampled target events onto a 1-dimensional subspace (i.e., a line). The feature subspace should extend from the concept vector of $\overrightarrow{short}$ (duration) to the concept vector of $\overrightarrow{long}$ (duration). Each concept vector will be obtained by averaging multiple word vectors related to *long* or *short*. A list of 14 words (7 for *short* and 7 for *long*) was selected to represent the concepts, each with a minimum frequency of 1000 in the British National Corpus (more detailed statistics about the context extraction can be found in the Appendix). For the words "long" and "short" themselves, we only use sentences where they are followed by "time" or "period", to discard the occurrences of the spatial meaning. In the end, the DURATION subspace will be the difference between the average of vectors representing $\overrightarrow{long}$ and the average of vectors representing $\overrightarrow{short}$. By averaging, the approximation of the feature subspace will be less likely to be biased by a specific word choice (Grand et al., 2022).

We tested the semantic projection under two different settings: (i) we used all the words to build the two concept vectors; (ii) we used the top 3 most frequent feature words (as in Grand et al. (2022)'s original setting). Whenever a word is selected, we randomly sampled 1000 sentences including it from the BNC, and we averaged its Transformer-generated vectors to obtain a sort of "out-of-context" representation. Then we averaged all word vectors related to one extreme of the "time" continuum to get the two concept vec-

[4] Indeed, *wink*, *talk*, and *speak* are not transitive verbs as they precede prepositional phrases (*wink *(at) the girl*). Wittenberg and Levy (2017) used "transitive" to refer to the two-place argument structures with either a direct object or a prepositional phrase, and we follow their usage.

[5] Minicons library provides the intuitive and efficient extraction of word/phrase representations from transformer models that are accessible on huggingface hub.

tors. Lastly, the $\overrightarrow{\text{DURATION}}$ vector was obtained by subtracting the aggregated vector $\overrightarrow{short}$ from the aggregated vector $\overrightarrow{long}$.

For projecting the vectors on DURATION, we used the standard scalar projection formula:

$$Proj = \frac{\overrightarrow{target} \cdot \overrightarrow{\text{DURATION}}}{||\overrightarrow{\text{DURATION}}||}$$

where the aggregated vector of each target event is denoted as $\overrightarrow{target}$. The result obtained by this operation is a scalar value, where larger values correspond to the estimate of a longer event duration.

Figure 1 shows the pattern of semantic projection results under conditions (i) and (ii) respectively. The largest mean cut-off resulting from the ditransitive light verb construction occurs in punctive events with count syntax under all conditions, whereas the smallest reduction in projection is observed in durative events with mass syntax. Differently from predictions, durative events are also estimated to take less time in ditransitive construction with count syntax, to a higher degree than when they occur with mass syntax.

We built linear mixed-effects models with R's lme4 package (Bates et al., 2014) to analyze the main effect of construction (ditransitive or transitive) and event category (punctive count, durative count, or durative mass). We computed p-values by performing likelihood-ratio tests on models that differ only for the presence or absence of the fixed-effect parameter(s) under consideration (construction, event category, and interaction between construction and event category). The coding of the two categorical predictors, construction and event category follows lme4's default coding.

We found significant main effects of construction in both types of projection conditions and significant interaction effects of construction and event category. This means that, while the transitive construction can generally lead to a higher projection value (i.e. indicates a longer duration estimated by BERT embeddings) the magnitude of effect also depends on event category. The effect of event category only reaches marginal significance in condition (ii) (see full output in the Appendix, in Tables 6 and 7). We then ran pairwise comparisons between transitive and ditransitive constructions within each event category. While the count syntax consistently shortens the duration of punctive events, such an effect is not systematically present for durative events with mass syntax. Surprisingly, the ditransitive construction with count syntax can predict a significantly shorter duration for durative events in both projection conditions, similarly to punctive events. This suggests that count syntax is generally associated with shorter duration for both punctive and durative events, while mass syntax does not produce similar effects for durative ones.

The results of semantic projection align with the prediction that punctive events in count syntax (*give a hug*) are construed as taking less time than in the transitive verb form. This suggests that the subtle effect of the syntactic alternations on events' temporal structure is encoded in the BERT vector space, and this aspect of knowledge can be successfully recovered using the semantic projection technique. Among the proposed projection conditions, type (i) projection with all the feature words seems to be generally better for it produces the most significant effect of construction, as well as the most significant interaction between construction and event category.

While the data analysis reveals that the shortening effect of light verb construction is more significant for punctive events with count syntax than for durative events with mass syntax, in line with Wittenberg and Levy (2017)'s results, we also observed a difference: in the original study, the shortening effect was clearly present as a tendency also for durative count events, but it did not consistently reach significance; in our experiment, light verb constructions with durative count events have a significantly shorter DURATION in both settings.

5 Experiment 2: Investigating the Conceptual Shift of Events in Count/Mass Syntax

Our second experiment aims at reproducing with BERT the finding that the durative events in count syntax (*give a talk*) are conceptually further apart from their transitive verb counterparts (*to talk*) than punctive events in count syntax, or durative events in mass syntax. The prediction is drawn from the analogy to mass nouns, such as *glass* or *iron*, when they undergo mass-to-count coercion. When the mass nouns are coerced to be used in the count contexts, the denotation of the noun is enriched in a way that is conceptually related but further from its original sense. If durative events in count syntax behave similarly, they are expected to be conceptually further apart from their transitive counterparts than the other two categories of events.

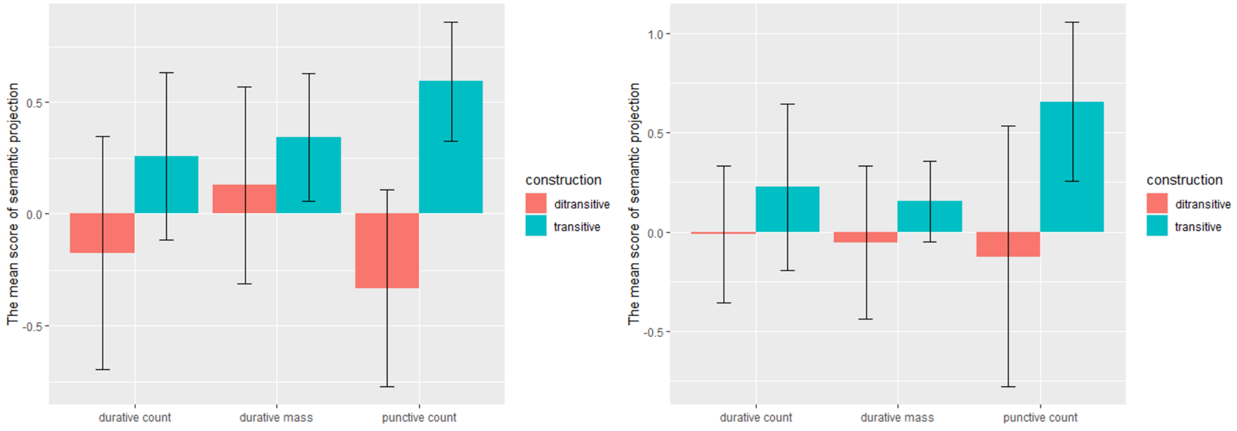

Figure 2: Semantic projection scores with setting (i) on the left (all words are used to build the two concept vectors for the extreme values of the scale) and setting (ii) on the right.

We used again BERT and the MINICONS library to generate semantic representations of the target events in context: the idea is to measure the semantic similarity scores of each event (punctive count, durative count, or durative mass) to itself for randomly sampled sentences. We carried out the sampling either i) by selecting context pairs where the target events occur in both cases in its transitive or ditransitive contexts (within the same context type); or ii) by selecting context pairs where the target event occurs once in the transitive context and once in a ditransitive context (between context types). This means that each noun type will have its occurrences sampled in three different ways:

(1) All context pairs sampled from transitive contexts;

(2) All context pairs sampled from ditransitive contexts;

(3) The context pair composed of one occurrence in the transitive context and one in the ditransitive context.

The similarity comparison between (1) and (3) is the most relevant one for our study: we expect that similarities in (3) to be much lower than in (1), to an extent proportional to the meaning shift that the event undergoes when the count/mass syntax is introduced. Conceptually, the difference between (2) and (3) should be similar to the difference between (1) and (3), since this difference approximates the degree of meaning shift that the events undergo when going from a ditransitive frame to the transitive contexts. However, given the relatively low frequency of the three categories of events in ditransitive constructions, we deemed more appropriate to compare (3) with (1) rather than (2), as the limited sentences in (2) may not fully represent the meaning of the events in count/mass syntax. For each category of event, we repeat the sampling ten times for each group, and for each time we randomly extract 10 different context pairs to generate the vectors.

For context pairs classified as occurring in transitive contexts, we simply used the vector of the bare verb as the semantic representation of the target events. For context pairs in the ditransitive frame, we used the embedding of the nominal representing the target event, rather than the whole *give* construction. This choice is motivated by the fact that, according to the linguistic theory (Butt, 2010; Wiese and Maling, 2005; Wittenberg and Levy, 2017), *give* in the light verb construction only communicates the directionality of the action whereas the bulk of the action meaning is conveyed by the event nominal. Therefore, we deemed it better to use the vectors of the deverbal noun instead of the whole phrase to represent the meaning of the target events. As a similarity score, instead of cosine, we used the Spearman correlation between vectors. The reason is that contextualized vector spaces are affected by anisotropy (Ethayarajh, 2019), with a small number of dimensions having disproportionately high variance. Metrics like cosine have been shown to be severely affected by outlier dimensions, while rank-based metrics like Spearman are better correlated with human similarity judgements (Timkey and van Schijndel, 2021).

The results are reported in Table 3. The average

Context pairs	Event	Spearman ρ
Transitive	Punctive Count	0.47
Ditransitive	Punctive Count	0.65
Both	Punctive Count	0.45
Transitive	Durative Count	0.41
Ditransitive	Durative Count	0.58
Both	Durative Count	0.37
Transitive	Durative Mass	0.43
Ditransitive	Durative Mass	0.58
Both	Durative Mass	0.38

Table 3: Average Spearman scores for each event category under the six sampling conditions.

similarity of context pairs where the target event occurs in transitive or ditransitive contexts suggests how semantically similar the event is to itself when the event is represented in the simple verb form or in the light verb construction, while the average similarity across different contexts type reflects the similarity between the events encoded in the ditransitive contexts and their verbal counterparts. Therefore, the difference between the two scores should quantify the meaning shift of the target events when imposed the effect brought by changes in linguistic framing. Here is an ideal example of durative count events to illustrate the proposed meaning shift:

s1. We did not talk about 'Robin Hood' schemes, not at all. (transitive context)

s2. Americans love to talk. (transitive context)

s3. I see she hasn't actually given a talk, but she's going to. (ditransitive context)

The event talk in s1 and s2 refer in both cases to the means of communication or conveying information by spoken words, while talk in s3 is tended to be interpreted as a more formal activity. Accordingly, the similarity of s1 and s2 is 0.60, and the similarity of s1 and s3 is 0.51. The difference between the similarity scores should reflect the degree of conceptual shift when entering in the count syntax. Notice that, when the sentence pairs are sampled across contexts, it is intuitive to hypothesize that their average similarity will be lower than that of the same contexts as the part-of-speech of the word representing the event has changed (e.g. *speak – speech*). However, if our results are to replicate the findings of Wittenberg and Levy (2017), we expect the pairs of durative count events to undergo a greater meaning shift than the other categories.

From Table 3, it can be seen all events have a lower similarity in transitive contexts than in ditransitive contexts. Since all events in our dataset, regardless of their event categories, are more frequent when occurring in the bare verbs than occurring in the light-verb encoding, there might be a higher degree of contextual variation in their patterns of usage which can explain the relatively low similarity score. Also, as predicted, the lowest similarity is found when the sentence pairs are sampled from different contexts. However, it is easy to see that the meaning shift Δ (the difference in average Spearman ρ between sampling just in transitive contexts, and sampling transitive and ditransitive ones) is very small for all the event categories (average values are summarized in Table 4). We built a linear regression model with the Δ score for each event as a target variable and the event category as a predictor, and indeed we did not find any significant effect of event category ($\chi^2 = 1.2060$, $p > 0.1$; cf. Table 8 in the Appendix).

A possible reason for the absence of significant conceptual shifts might be an inherent shortcoming of vector similarity in distributional spaces: metrics like cosine and Spearman tell us that two word meanings are related, but they fail to tell us *in which way* those meanings are related. A large body of literature in Distributional Semantics focused on this issue regarding nominals, and pointed out the struggle in teasing apart relations such as synonymy and hypernymy/hyponymy (Baroni and Lenci, 2011; Xiang et al., 2020; Schulte Im Walde, 2020), but the same problem could apply to the relation between durative count verbs and the corresponding light verb constructions. For example, the meaning of *to give a lecture* is still related to the meaning of *to lecture*, while being a more specific type of lecturing-event (Gagné et al., 2020), similarly to the hyponymy relation in nominals. It is thus possible that such relations are not clearly distinguishable from near-synonymy between verb phrases on the basis of distributional similarity.

Event type	Δ
Punctive Count	0.022
Durative Count	0.039
Durative Mass	0.053

Table 4: Average meaning shift Δ for each event type, computed as the difference between the average Spearman correlations by sampling transitive contexts and sampling transitive and ditransitive ones.

6 Conclusions

In our paper, we presented an analysis of subtle meaning changes in event construal, comparing transitive verbs and light verb constructions, using the BERT model to represent the meaning of events in a distributional semantic space.

In Experiment 1, we focused on event duration, by identifying a DURATION dimension in our BERT vector space via the semantic projection technique (Grand et al., 2022). Similarly to the original study by Wittenberg and Levy (2017), we found that the light verb construction has a general shortening effect, with the vectors for the construction having generally lower values along this dimension than the ones of the corresponding transitive verbs. We take this result as initial evidence that the BERT vector space encodes subtle meaning nuances related to the representation of time in natural language sentences.

However, in Experiment 2, we compared the distributional similarities of transitive verbs vs. constructions pairs, to see if the model was sensitive enough to spot the meaning shift that durative count events undergo in light verb constructions (cf. the Experiment 4 in Wittenberg and Levy (2017)). In this case, the answer was negative, and no significant differences in the meaning shifts across event categories was observed. We suggested that the lack of specificity of vector similarity as a semantic relation may explain this negative result.

Future work for specializing contextualized vector spaces, similarly to what has been done for static models (Mrkšić et al., 2017), may be needed to handle fine-grained semantic distinctions.

Limitations

Our work has some clear limitations: we studied only a specific type of construction in English, and using just a limited set of verbs.

Moreover, we only employed a single, bidirectional Transformer model (BERT Base) to generate the vector representations, and thus we cannot be sure whether our considerations are generalizable to other architectures.

Finally, concerning Experiment 1, the choice of the words for building the prototypes of the extremes of the DURATION scale is likely to affect the results, but given the space constraints we only explored two possible settings and left a more systematic investigation to future work.

Acknowledgements

We would like to thank the three anonymous reviewers for their constructive feedback.

References

David Arps, Younes Samih, Laura Kallmeyer, and Hassan Sajjad. 2022. Probing for Constituency Structure in Neural Language Models. *arXiv preprint arXiv:2204.06201*.

David Barner and Jesse Snedeker. 2005. Quantity Judgments and Individuation: Evidence that Mass Nouns Count. *Cognition*, 97(1):41–66.

David Barner, Laura Wagner, and Jesse Snedeker. 2008. Events and the Ontology of Individuals: Verbs as a Source of Individuating Mass and Count Nouns. *Cognition*, 106(2):805–832.

Marco Baroni and Alessandro Lenci. 2011. How We BLESSed Distributional Semantic Evaluation. In *Proceedings of the GEMS Workshop on GEometrical Models of Natural Language Semantics*.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting Linear Mixed-effects Models Using lme4. *arXiv preprint arXiv:1406.5823*.

Miriam Butt. 2010. The Light Verb Jungle: Still Hacking Away. *Complex Predicates in Cross-Linguistic Perspective*, pages 48–78.

Chung-Ying Cheng. 1973. Response to Moravcsik. *Approaches to Natural Language*.

Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Logical Metonymy in a Distributional Model of Sentence Comprehension. In *Proceedings of *SEM*.

Emmanuele Chersoni, Enrico Santus, Chu-Ren Huang, and Alessandro Lenci. 2021a. Decoding Word Embeddings with Brain-based Semantic Features. *Computational Linguistics*, 47(3):663–698.

Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2021b. Not All Arguments Are Processed Equally: A Distributional Model of Argument Complexity. *Language, Resources and Evaluation*, pages 1–28.

Won Ik Cho, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Modeling the Influence of Verb Aspect on the Activation of Typical Event Locations with BERT. In *Findings of ACL-IJCNLP 2021*.

Gabriella Chronis, Kyle Mahowald, and Katrin Erk. 2023. A Method for Studying Semantic Construal in Grammatical Constructions with Interpretable Contextual Embedding Spaces. *arXiv preprint arXiv:2305.18598*.

Marta Coll-Florit and Silvia P Gennari. 2011. Time in Language: Event Duration in Language Comprehension. *Cognitive Psychology*, 62(1):41–79.

Simone Conia and Roberto Navigli. 2022. Probing for Predicate Argument Structures in Pretrained Language Models. In *Proceedings of ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of EMNLP*.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Caitlin M Fausey and Lera Boroditsky. 2010. Subtle Linguistic Cues Influence Perceived Blame and Financial Liability. *Psychonomic Bulletin & Review*, 17(5):644–650.

Christina L Gagné, Thomas L Spalding, Patricia Spicer, Dixie Wong, Beatriz Rubio, and Karen Perez Cruz. 2020. Is Buttercup a Kind of Cup? Hyponymy and Semantic Transparency in Compound Words. *Journal of Memory and Language*, 113:104110.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal Abstractions of Neural Networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic Projection Recovers Rich Human Knowledge of Multiple Object Features from Word Embeddings. *Nature Human Behaviour*, 6(7):975–987.

Yuling Gu. 2022. Measure More, Question More: Experimental Studies on Transformer-based Language Models and Complement Coercion. *arXiv preprint arXiv:2212.10536*.

John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. In *Proceedings of EMNLP*.

Manuela Hürlimann, Raffaella Bernardi, and Denis Paperno. 2014. Nominal Coercion in Space: Mass/Count Nouns and Distributional Semantics. In *Proceedings of CLiC-it*.

Matt A Johnson and Adele E Goldberg. 2013. Evidence for Automatic Accessing of Constructional Meaning: Jabberwocky Sentences Prime Associated Verbs. *Language and Cognitive Processes*, 28(10):1439–1452.

Graham Katz and Roberto Zamparelli. 2012. Quantifying Count/Mass Elasticity. In *Proceedings of the West Coast Conference on Formal Linguistics*.

Najoung Kim and Tal Linzen. 2020. COGS: A Compositional Generalization Challenge Based on Semantic Interpretation. In *Proceedings of EMNLP*.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse Probing of Pretrained Language Models. In *Proceedings of NAACL*.

Geoffrey Neil Leech. 1992. 100 Million Words of English: The British National Corpus (BNC). *Language Research*.

Bai Li. 2022. *Integrating Linguistic Theory and Neural Language Models*. Ph.D. thesis, Department of Computer Science, University of Toronto.

Godehard Link. 1983. *The Logical Analysis of Plurals and Mass Terms: A Lattice-theoretical Approach*, volume 127. Blackwell Oxford.

Chenxin Liu and Emmanuele Chersoni. 2022. Exploring Nominal Coercion in Semantic Spaces with Static and Contextualized Word Embeddings. In *Proceedings of the AACL-IJCNLP Workshop on Cognitive Aspects of the Lexicon*.

Mingyu Derek Ma, Jiao Sun, Mu Yang, Kung-Hsiang Huang, Nuan Wen, Shikhar Singh, Rujun Han, and Nanyun Peng. 2021. Eventplus: A Temporal Event Understanding Pipeline. In *Proceedings of NAACL-HLT: Demonstration*.

Stephen McGregor, Elisabetta Ježek, Matthew Purver, and Geraint Wiggins. 2017. A Geometric Method for Detecting Semantic Coercion. In *Proceedings of IWCS*.

Eleni Metheniti, Tim Van De Cruys, and Nabil Hathout. 2022. About Time: Do Transformers Learn Temporal Verbal Aspect? In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*.

Kanishka Misra. 2022. minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint arXiv:2203.13112*.

Kanishka Misra, Julia Taylor Rayz, and Allyson Ettinger. 2022. A Property Induction Framework for Neural Language Models. In *Proceedings of CogSci*.

Nikola Mrkšić, Ivan Vulić, Diarmuid O Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic Specialization of Distributional Word Vector Spaces Using Monolingual and Cross-lingual Constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Paolo Pedinotti and Alessandro Lenci. 2020. Don't Invite BERT to Drink a Bottle: Modeling the Interpretation of Metonymies Using BERT and Distributional Representations. In *Proceedings of COLING*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of EMNLP*.

Giulia Rambelli, Emmanuele Chersoni, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2020. Comparing Probabilistic, Distributional and Transformer-based Models on Logical Metonymy Interpretation. In *Proceedings of AACL-IJCNLP*.

Susan Rothstein. 2017. *Semantics for Counting and Measuring*. Cambridge University Press.

Sabine Schulte Im Walde. 2020. Distinguishing between Paradigmatic Semantic Relations across Word Classes: Human Ratings and Distributional Similarity. *Journal of Language Modelling*, 8(1):53–101.

Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda. 2020. Probing for Referential Information in Language Models. In *Proceedings of ACL*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of ACL*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations. In *Proceedings of ICLR*.

William Timkey and Marten van Schijndel. 2021. All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality. In *Proceedings of EMNLP*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing Pretrained Language Models for Lexical Semantics. In *Proceedings of EMNLP*.

Alexis Wellwood, Susan J Hespos, and Lance Rips. 2018. The Object: Substance:: Event: Process Analogy. *Oxford Studies in Experimental Philosophy*, 2:183–212.

Heike Wiese and Joan Maling. 2005. Beers, Kaffi, and Schnaps: Different Grammatical Options for Restaurant Talk Coercions in Three Germanic Languages. *Journal of Germanic Linguistics*, 17(1):1–38.

Eva Wittenberg and Roger Levy. 2017. If You Want a Quick Kiss, Make It Count: How Choice of Syntactic Construction Affects Event construal. *Journal of Memory and Language*, 94:254–271.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT. In *Proceedings of ACL*.

Rong Xiang, Emmanuele Chersoni, Luca Iacoponi, and Enrico Santus. 2020. The CogALex Shared Task on Monolingual and Multilingual Identification of Semantic Relations. In *Proceedings of the COLING Workshop on the Cognitive Aspects of the Lexicon*.

Bingyang Ye, Jingxuan Tu, Elisabetta Jezek, and James Pustejovsky. 2022. Interpreting Logical Metonymy through Dense Paraphrasing. In *Proceedings of CogSci*.

Alessandra Zarcone and Sebastian Padó. 2011. Generalized Event Knowledge in Logical Metonymy Resolution. In *Proceedings of CogSci*.

Alessandra Zarcone, Jason Utt, and Sebastian Padó. 2012. Modeling Covert Event Retrieval in Logical Metonymy: Probabilistic and Distributional Accounts. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal Common Sense Acquisition with Minimal Supervision. In *Proceedings of ACL*.

A Appendix

Descriptive Statistics

The descriptive statistics for the context extraction from the BNC can be found in Table 5.

Likelihood Estimation Results

The output of the likelihood estimation tests under the different projection conditions can be found in Table 6 and 7 (Experiment 1).

Table 8 shows instead the output of the test for the Δ meaning shift scores by Event category (Experiment 2).

We also used the F-test the examine the main effect of event category on the degree of meaning shift, and the results are in line with the likelihood estimation results ($F(2, 15) = 0.5197, p > 0.1$). The pairwise comparisons between different categories of events suggested that the degree of meaning shift does not differ significantly by event categories (punctive count vs. durative count: $F(1, 10) = 0.3199, p > 0.1$; punctive count vs. durative mass: $F(1, 10) = 0.9044, p > 0.1$; durative count vs. durative mass: $F(1, 10) = 0.2439, p > 0.1$).

Event category	**Context**	**Avg. Freq**	**Min. Freq**	**Max. Freq**
Punctive Count	Transitive	2806.17	301	8549
Punctive Count	Diransitive	155.17	54	354
Durative Count	Transitive	14071.33	523	29366
Durative Count	Ditransitive	211.83	43	384
Durative Mass	Transitive	9400.83	2919	18580
Durative Mass	Ditransitive	712.67	202	1683

Table 5: Descriptive statistics for the context extraction from the BNC: average, min and max frequency for each event category – context type.

	Degree of freedom	χ^2	$p - value$
Construction	1	14.285	< .001 ***
Event category	2	1.1386	> .1 n.s.
Construction × Event category	2	8.4669	< .05 *
Punctive count – construction	1	14.94	< .001 ***
Durative count - construction	1	7.1604	< .01 **
Durative mass - construction	1	1.1415	> .1 n.s.

Table 6: Likelihood estimation results under the projection condition (i) for duration estimates, testing the main effects of construction, event category, and their interaction (upper part), and the results of testing the main effect of construction in pairwise comparisons within each event categories (lower part).

	Degree of freedom	χ^2	$p - value$
Construction	1	11.994	< .001 ***
Event category	2	1.1888	> .1 n.s.
Construction × Event category	2	9.0736	< .05 *
Punctive count – construction	1	9.5146	<.01 **
Durative count - construction	1	5.9238	<.05 *
Durative mass - construction	1	1.7316	> .1 n.s.

Table 7: Likelihood estimation results under the projection condition (ii) for duration estimates, testing the main effects of construction, event category, and their interaction (upper part), and the results of testing the main effect of construction in pairwise comparisons within each event categories (lower part).

	Degree of freedom	χ^2	$p - value$
Event category	2	1.2060	> .1 n.s.
Punctive count vs. durative count	1	0.3779	> .1 n.s.
Durative count vs. durative mass	1	0.2892	> .1 n.s.
Punctive count vs. durative mass	1	1.0389	> .1 n.s.

Table 8: Likelihood estimation results for meaning shift Δ scores, testing the main effect of event category, and related pairwise comparisons.