# Investigating Aspect Features in Contextualized Embeddings with Semantic Scales and Distributional Similarity

**Yuxi Li**[1,2] and **Emmanuele Chersoni**[1] and **Yu-Yin Hsu**[1]
[1]Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China
[2]School of Foreign Studies, Xi'an Jiaotong University, Xi'an, China

## Abstract

Aspect, a linguistic category describing how actions and events unfold over time, is traditionally characterized by three semantic properties: *stativity*, *durativity* and *telicity*.

In this study, we investigate whether and to what extent these properties are encoded in the verb token embeddings of the contextualized spaces of two English language models – BERT and GPT-2. First, we propose an experiment using semantic projections to examine whether the values of the vector dimensions of annotated verbs for stativity, durativity and telicity reflect human linguistic distinctions. Second, we use distributional similarity to replicate the notorious Imperfective Paradox described by Dowty (1977), and assess whether the embedding models are sensitive to capture contextual nuances of the verb telicity.

Our results show that both models encode the semantic distinctions for the aspect properties of stativity and telicity in most of their layers, while durativity is the most challenging feature. As for the Imperfective Paradox, only the embedding similarities computed with the vectors from the early layers of the BERT model align with the expected pattern.

## 1 Introduction

Since the introduction of Transformer architectures in NLP (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2019), their increasing success urged researchers to get more insights about the linguistic knowledge encoded in their internal representations. The literature on *probing tasks* is a clear example of this trend: a simple classification model is asked to solve a task requiring linguistic knowledge using embeddings representations extracted from a language model (LM) with little or minimal linguistic supervision, and if the classification model is successful, one can infer that the LM's representations do encode the targeted knowledge (e.g. Tenney et al. (2019); Hewitt and Liang (2019); Goldberg (2019); Jawahar et al. (2019); Wu et al. (2020); Ravichander et al. (2020); Madabushi et al. (2020); Chen et al. (2021); Koto et al. (2021); Belinkov (2022), *inter alia*).

An alternative approach, especially popular for probing the *semantic* knowledge contained in the embeddings, involves mapping them onto human-interpretable features (Chersoni et al., 2021; Proietti et al., 2022; Wang et al., 2023). Yet the probing methodology involves a trainable classifier, and therefore the relation between the probe results and the knowledge in the original representations is not always clear (Levy et al., 2023). Moderate correlations with human ratings/norms can sometimes be obtained even by using random vectors as features (Chersoni et al., 2020), and thus alternative methods for directly analysing/modifying the structure of the semantic space have been proposed (e.g. indicator tasks, Levy et al. (2023)). A recent study by Grand et al. (2022) introduced the usage of semantic projections to interpret the content of word embeddings, by constructing subspaces corresponding to human-interpretable semantic scales. Such semantic scales were shown to be very useful in modeling human judgements for a variety of concepts in the semantics of nominals (Grand et al., 2022; Diachek et al., 2023).

In our paper, we focus on *aspect*, a concept in verb semantics that characterizes the temporal relationship of actions and events. Aspect has been shown to be important in several NLP tasks, such as next event prediction (Chambers et al., 2014) and textual entailment (Kober et al., 2019). Combining the usage of semantic scales and embedding similarity measurements, in two experiments, we address the question of whether and to what the extent the contextualized word embeddings produced by LMs encode the aspectual properties of stativity, telicity and durativity. In the first experiment, we use semantic scales to quantify the values of the

*Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 80 - 92
June 20-21, 2024 ©2024 Association for Computational Linguistics

three aspectual properties in the verb token embeddings produced by different hidden layers of BERT and GPT-2. We examine whether the projected scores reflect the binary distinction in the aspectual properties of the verbs described by Vendler (1957), assuming that verbs having different values for a property (e.g. telic vs. a) should have significantly different scores. To our knowledge, we are the first to adopt the semantic scales method for modeling verb semantics. In the second experiment, we examine the similarity between simple past and past progressive forms of telic and atelic verbs that express activities and accomplishments. According to the Imperfective Paradox in Dowty (1977), the past progressive of activity verb entails its simple past, while this entailment does not hold for accomplishment verbs. Again, we extracted the verb token embeddings from different internal layers: if a BERT/GPT-2 embedding from a given layer correctly encodes telicity, we expect that the similarity between past progressive and simple past of an activity verb will be higher - since the former entails the latter- than between the two corresponding past forms of an accomplishment verb.

We found that both LMs are capable of consistently encoding aspectual features, especially for stativity and telicity. However, BERT was more sensitive to the nuanced difference in telicity, as we found in the Imperfective Paradox experiment. Our findings reveal the extents to which prototypical LMs encode core verb properties, which has important implications for selecting LMs for downstream fine-tuning. For example, based on our results, we can hypothesize that fine-tuning BERT-family models may be proven more beneficial for improving the performance of textual entailment.

## 2 Related Work

In the semantics literature, verb aspect is generally characterized in terms of three properties: *stativity, telicity* and *durativity* (Moens and Steedman, 1988; Pruś et al., 2024).

**Stativity** refers to the distinction between states and events.Verbs of high stativity generally cannot be used in progressive forms: for example, it is not possible to use 'I am knowing/loving'. In comparison, verbs of low stativity can typically be used in progressive forms (e.g. 'I am running/swimming').

**Telicity** refers to whether an event unfolds in time in an homogeneous way, and whether any part of the process is of the same nature as the whole. Telic verbs can often be collocated with 'in' adverbial phrases but not with 'for' adverbial phrases; e.g. 'eat' can be used in 'He ate the apple in a minute' but not in 'He ate the apple for a minute'. Notice that verbs of this type describe actions/events with a natural end point (e.g. the moment in which the apple is finished). The use of 'in' signifies that the action (of eating the apple) is completed within a specific timeframe. In contrast, atelic verbs usually collocate with 'for' but not with 'in', e.g. 'He was running for an hour' but not 'He war running in an hour'.

Finally, **durativity** refers to how long an event lasts. Durative actions like 'love' can be questioned by 'How long have you loved her?', but punctual actions like 'recognize' cannot be questioned in a similar way ('How long have you recognized her?' sounds odd without additional context). These examples show that a verb can vary along the three dimensions. For example, 'love' is simultaneously stative, durative, and atelic.

The work conducted by Friedrich and Palmer (2014) focuses on the automatic classification of verb stativity in context, using a combination of distributional and manually crafted linguistic features. It is one of the first to introduce a dataset of annotated sentences specifically for this feature. Friedrich and Gateva (2017) expanded on this work, by releasing datasets also for telicity and durativity with gold and silver annotations; the latter was automatically extracted from a parallel corpus between English and Czech texts, exploiting the fact that Czech aspectual features are signaled with specific morphological markers. Kober et al. (2020) proposed an approach based on compositional distributional models to distinguish between stative and dynamic verbs, and between telic and atelic ones. Interestingly, their classification results confirmed that the tense is always a strong indicator of telicity; in particular, past tense is often correlated with telic events.

Cho et al. (2021) presented a study on using BERT surprisal to model human typicality ratings of the location arguments in natural language sentences, which were shown in the studies by Ferretti et al. (2001, 2007) to be strongly related to verb aspect: humans show priming effects for typical locations in sentences, but only when the tense of the main verb is progressive (or, in other words, the description of an action as ongoing makes the location argument more salient for human conceptual representations). BERT surprisal scores showed some sensitivity to the aspect of the verb, although they produced human-like patterns only when the entire sentence context other than the verb and the location were masked.

More recently, Metheniti et al. (2022) reported a classification experiment on telicity and durativity on English and French, suggesting that Transformer models encode a non-trivial amount of knowledge of aspect even before fine-tuning, although they have biases regards verb tense and word order. Finally, Liu and Chersoni (2023) presented a modeling study of the shortening effect that the usage of light verb constructions has on the perceived duration of event descriptions, and they also used the semantic scales method by Grand et al. (2022) to project BERT vectors onto interpretable dimensions. They showed that certain type of events (e.g. punctive) have smaller values in their DURATION-related dimensions when expressed in the light verb form (e.g. *to give a kiss* takes less time than *to kiss*).

## 3 Experiment 1: Measuring Aspect Properties with Semantic Scales

In the first experiment, we select a set of verbs from the study by Vendler (1957). For each of the three aspect properties, the verbs are divided into two groups: stative versus dynamic for stativity, telic versus atelic for telicity, and punctive versus durative for durativity.

Our primary goal is to construct a semantic scale for each property, and then to project the word embeddings of the verbs on the semantic scales, in order to assign them scores of stativity, telicity, and durativity. If the distributional space effectively captures the different value that a verb can express with respect to a given property (e.g. telicity), we expect the scores for the verbs of the two groups to be different (e.g. telic verbs should have considerably higher scores on the telicity scale compared to atelic verbs).

### 3.1 Verb Selection

To begin, we selected verbs based on the categorization in Vendler (1957) that divides verb into four classes: state, activity, accomplishment, and achievement. These classes often show differences in one crucial verb property while sharing similarities in other properties. For example, state verbs and activity verbs differ in stativity but are similar in terms of telicity and durativity. Therefore, state verbs and activity verbs can represent two extremes of stativity, with state verbs representing more stative nature and activity verbs more dynamic. Similarly, we used the ’accomplishment-activity’ contrast to capture telicity, and the ’accomplishment-achievement’ contrast to capture durativity. Selecting representative verbs for each extreme in this controlled manner can ensure that the constructed scales reflect the difference in the target property as much as possible. For each category, we prompted the ChatGPT online interface to generate 50 exemplars, and manually verified the results (See the Appendix for the full list of the experiment items).

### 3.2 Scale Construction

We followed Grand et al. (2022)’s method of identifying semantic scales from vector spaces. To obtain an ‘out-of-context’ representation for each target word, we averaged their contextualized embeddings from a sample of 20 randomly selected sentences from the British National Corpus (BNC) (Leech, 1992)[1]. If the target token was not included in the base vocabulary of a model and was split into sub-tokens, we used the average of the sub-tokens’ embeddings as the representation for the target token. The same method was consistently applied in this study when extracting the representation for a target word in context.

Next, for each target property, we randomly sampled three words from the word lists to represent each extreme of the scale and we clustered their out-of-context embeddings, following the setup of the original study by Grand et al. (2022). For example, we sampled three words from the state verbs (e.g. *exist, lack, matter*) and three words from the activity verbs (*dance, walk, drive*) to represent the extremes of stativity. The authors recommend using this clustering step in order to avoid biases specific to the lexical meaning of a single word.

Finally, we constructed the scales by subtracting the embedding of one extreme by another extreme. This yielded a vector that represents the scale of values for a specific target property from one extreme to another. Since we had three target words for each extreme, we could construct nine scales based on different extreme pairings and average them to generate the final scale, which is meant

[1] Vulić et al. (2020) actually showed that sampling more than 10 contextualized instances leads to little differences in the representation. However, to ensure more robust results, we still chose to use 20 instances to build each out-of-context representations

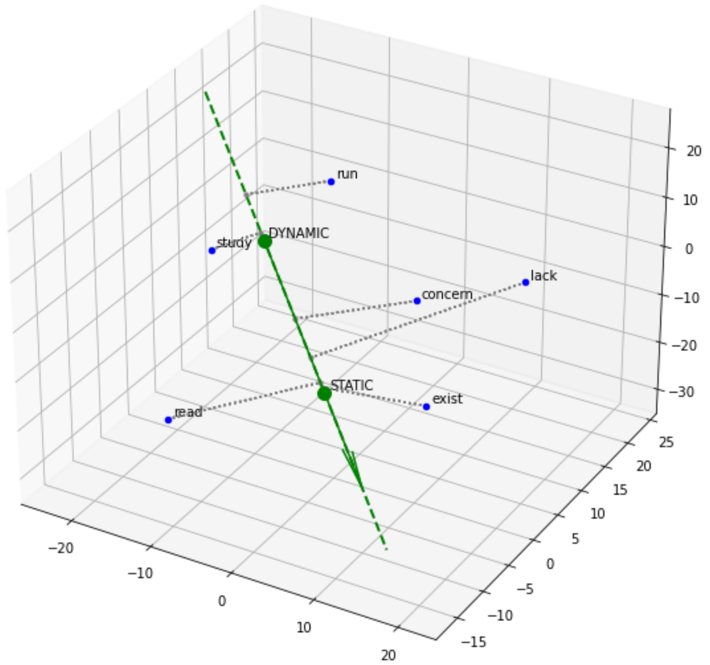

Figure 1: Semantic projection of verbs on stativity scale constructed by the 12th layer embeddings from BERT

to prevent the scale from being heavily influenced by the specific choice of antonym pairs (Grand et al., 2022). For example, if we used 'admire', 'appreciate' and 'dislike' to represent stative extreme, and 'swim', 'dance' and 'jog' to represent the dynamic extreme, we could have nine pairs, like ['admire' - 'swim'], ['admire' - 'dance'] and ['admire' - 'jog'], and subsequently average them to get the final scale.

### 3.3 Semantic Projection

After we constructed scales for the verb properties (henceforth as $\overrightarrow{stativity}$, $\overrightarrow{durativity}$, and $\overrightarrow{telicity}$), we assessed the validity of the scales by projecting other verbs onto the scales. Our hypothesis was that if the scale accurately reflected the semantic distinctions of the verbs in terms of the target property, the projection scores of one group of verbs would be significantly different from their semantic opposites. For example, we expected that the projection scores of the stative verbs on $\overrightarrow{stativity}$ to be significantly different from the projection scores of the dynamic verbs.

The projected verbs for projection are all the verbs in the original lists that are not used to build the scale extremes. For example, if we initially had fifty candidates for representing the one semantic extreme of a target property, we sampled three of them to represent the extreme, and then we used the remaining 47 words for projection. Therefore, for each property, we had in total 94 words for projection and difference testing.

To project the verbs on the scale, we used the standard scalar projection formula as follows:

$$Proj(\overrightarrow{target}) = \frac{\overrightarrow{target} \cdot \overrightarrow{property}}{\|\overrightarrow{property}\|}$$

The aggregated vector of each target event is denoted as $\overrightarrow{target}$. The result of projection is a scalar value, and a larger value indicates a higher degree of the property represented by the scale. Figure 1 provides a visualization of examples of semantic projection for the stative vs. dynamic opposition in a three-dimensional space.

After the projection, we analyzed the difference in the projection scores for the two verb groups for each scale, and we saw a significant difference as evidence that a model is able to set apart the verbs according to a specific semantic dimension (e.g. we expect stative and dynamic verb to differ significantly in their $\overrightarrow{stativity}$ scores). Specifically, we compared the projection scores of the verb groups for each scale by using the Mann-Whitney U statistical test (we chose a non-parametric test because the projected scores from some of our extraction

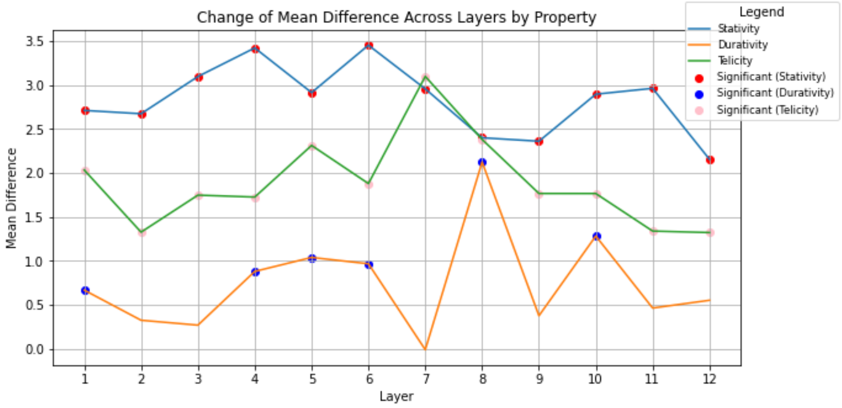

Figure 2: Layer-wise difference in the semantic projection score for stativity, durativity and telicity for each BERT layer. Dots mark the layers in which the scores for the two Vendler groups differ significantly.

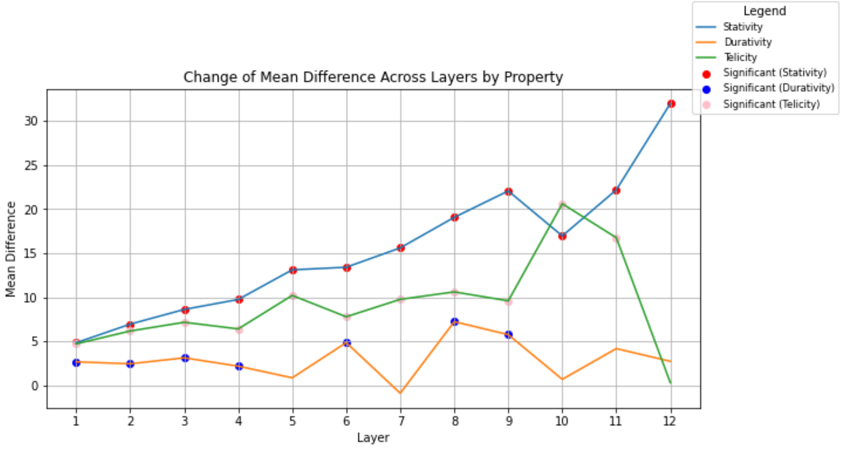

Figure 3: Layer-wise difference in the semantic projection score for stativity, durativity, and telicity for each GPT-2 layer. Dots mark the layers in which the scores for the two Vendler groups differ significantly.

experiments were not normally distributed).

### 3.4 Embedding Models

To obtain the contextualized embedding representations, we used the pre-trained BERT ('bert-base-uncased') (Devlin et al., 2019) and GPT-2 Base ('openai-community/gpt2') (Radford et al., 2019); both of them are available on HuggingFace[2]. The first model is a bidirectional, encoder-only Transformer, typically used for classification tasks, while GPT-2 is a unidirectional, decoder-only Transformer and it is often used for generation. The extraction of verb token embeddings was implemented in Pytorch. For verbs that were not included in the Transformers' vocabulary and were splitted in multiple subtokens, we obtained a single embedding via mean pooling of the embeddings of the subtokens. To have a finer-grained understanding of how Transformers encode verb properties, we ran the experiment by extracting the embeddings from all the 12 internal layers. As pointed out by Tenney et al. (2019), early Transformer layers tend to encode more permanent, 'out-of-context' features of a word (e.g. POS, syntax), while later layers tend to encode context-dependent semantics. Even if contextualized embeddings are able to model aspect properties, indeed, one may still be interested in understanding in what layers are best at separating the two verb groups for each property.

### 3.5 Results of Experiment 1

Figure 2 and Figure 3 show the layer-wise difference of projection scores of verbs of different groups on three semantic scales for BERT and GPT-2, respectively, and the dots indicate that a significant difference between the two groups at $p < 0.05$

[2] 'bert-base-uncased' can be found at: https://huggingface.co/google-bert/bert-base-uncased, 'gpt2-base' can be found at: https://huggingface.co/openai-community/gpt2

for the Mann Whitney U Test. More detailed information for scale construction and projection can be found in Appendix A.

In both Transformer models, stativity is by far the property that is better encoded (blue line): it can be observed, indeed, that the differences between stative and dynamic verbs are almost always significant across layers. This is not a surprising finding, as the difference between states and events is probably one of the main distinction in verb semantics. In the BERT model, the absolute difference between the scores of the two groups is the largest across properties and it is statistically significant in all layers; in GPT-2 the difference widens in deeper layers and remains significant for all of them.

As for telicity (green line), although the projection scores of telic and atelic verbs are closer than stative and dynamic ones, the differences are still significant for all the BERT layers. For GPT-2, the difference in telicity becomes more salient as in deeper layers and finally drops in the last one, the only layer in which it is not significant. Durativity (orange line) is the most challenging property to model, with BERT managing to set the two groups apart in the first layer, around the middle layers (4-6) and in some of the later layers (8 and 10). The GPT-2 model can distinguish the two groups in the early (layer 1-4) and in the middle layers (layer 6; 8-9), but it fails to do so in the later layers.

It can be seen that later layers of both models are less consistent in discriminating the verb groups across different properties. Probably, in the later layers the embeddings become too context-specific to reflect the distinctions: the issue could be possibly related to the *anisotropy* of contextualized vector spaces (Ethayarajh, 2019), that is, the tendency for the representations to occupy just a small cone of the vector space, with the result that the similarities even between randomly sampled words tend to be very high. Interestingly, it has been reported than GPT-2 tends to have a much higher degree of anisotropy than bidirectional models in the later layers (Ethayarajh, 2019), which could explain why the performance of BERT is more stable and consistent across properties and layers.

## 4 Experiment 2: Modeling the Imperfective Paradox with Distributional Similarity

Our first experiment showed that the models generally have a good grasp of the semantic distinctions related to the three main aspectual properties. In our second experiment, we test if the distributional similarities between verb token embeddings reflect the entailment properties of telic and atelic verbs when we manipulate their tense. With this goal in mind, we aim at replicating the Imperfective Paradox described by Dowty (1977). In his work, Dowty focuses on the opposition of activities and accomplishments in the past progressive and in the simple past tense, as in the following example:

(1) a. *Maria **was singing** the national anthem* $\models$ *Maria **sang** the national anthem* (activity - atelic)

b. *The children **were building** a sandcastle* $\not\models$ *The children **built** a sandcastle* (accomplishment - telic)

Given that our models encode telicity in the embedding representations, we extract the token verb embeddings for the verbs in the provided sentence pairs in a. and b., and for each verb we measure the distributional similarity to itself when used in the other tense. Our hypothesis is that the similarity will reflect the entailment relation between the two statements. Specifically, we expect the similarity to be significantly higher for activities than accomplishments, since the simple past is entailed by the progressive in the former, but not in the latter case.

Similar to the previous experiment, we used the 'accomplishment-activity' contrast to define telicity, e.g. accomplishment verbs are telic while activity verbs are atelic. For these two groups, we used the same verbs from Experiment 1. For each group, we constructed 100 pairs of simple/progressive past sentence pairs, resulting in a total of 200 pairs.

Initially, we extracted sentences from the BNC that contained the target verbs in the simple past tense, and for each sentence we created an equivalent sentence in the past progressive by changing the verb's aspect. For telic verbs, we used word types from the 'accomplishment' verb class, while for atelic verbs, we used word types from the 'activity' verb class. In total, we collected 100 samples for each verb group. For each verb type in the lists, we randomly sampled 10 sentences in which the verbs are in the form of past particle, and filtered those sentences that are marked as passives rather than simple past sentences. The remaining sentences were evaluated by the authors and deemed less suitable for aspect conversion. We also made sure that each verb type occurred at most 5 times

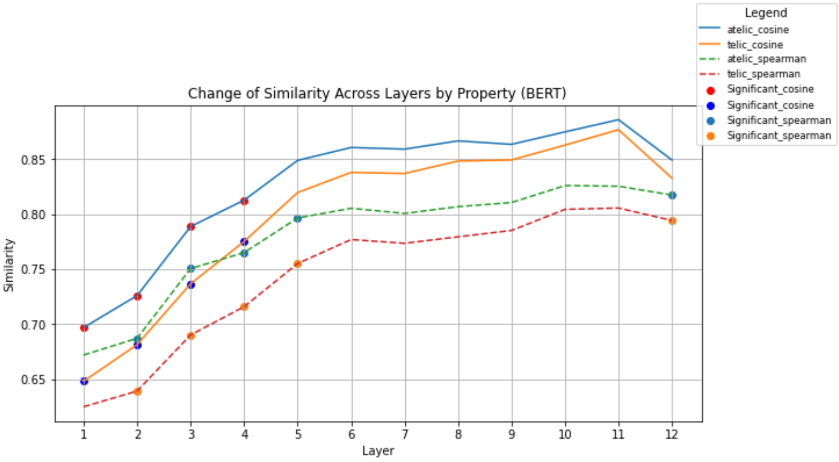

Figure 4: BERT: Layer-wise semantic similarity of the target words in simple past/past progressive pairs for the telic and atelic groups. Dots mark the layers in which the similarity scores differ significantly between two groups.

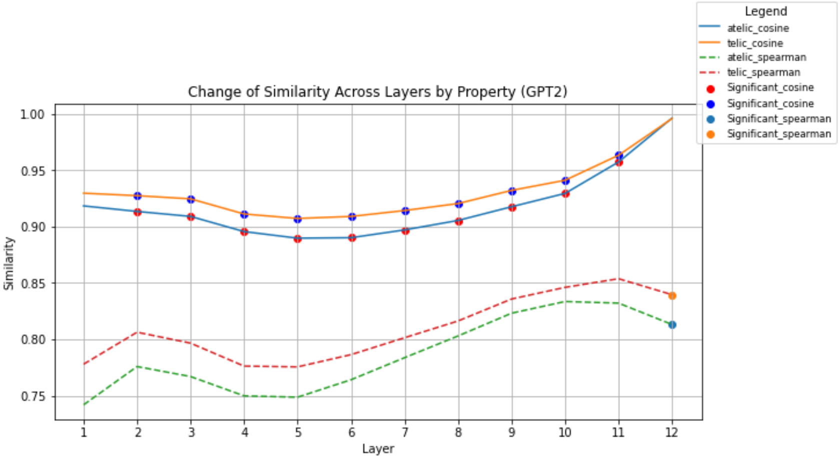

Figure 5: GPT-2: Layer-wise semantic similarity of the target words in simple past/past progressive pairs for the telic and atelic groups. Dots mark the layers in which the similarity scores differ significantly between two groups.

in the sample after filtering, to prevent the results from being too influenced by specific verb types.

As a result, we obtained 100 instances for the telic (32 verb types) and atelic group respectively (21 verb types). [3] The sentences in each pair are exactly the same, except for the main verb tense [4], and we also manually checked that they did not become incoherent due to the aspect conversion.Once obtained the sentence pairs, we extracted the verb embeddings from each of them by using Pytorch. For the sentences with the verb in the past progressive, we used the embedding of the progressive form, not including the auxiliary (e.g. from *The children were building a sandcastle*, we extract the embedding of *building*).Once again, embeddings for multi-token verbs were obtained via mean pooling of the embeddings of the subtokens. The similarity between embeddings was computed with the standard cosine metric and with Spearman correlation: we chose the latter as an additional measure because of the notorious issue of the anisotropy of contextualized vector spaces, as rank-based metrics were shown to be more robust to anisotropy and more consistently correlated with human similarity judgements (Timkey and van Schijndel, 2021).

[3] Notice that, after the filtering procedure, for several verb types we did not have any sentences left in the sample. Still, we considered the existing sample size as sufficient for statistical testing, and the diversity of verb types as high enough to make generalizations about the population.

[4] For a small number of cases (7 sentences in total) we had to adjust the additional context, as they have verbs linked by coordinate conjuctions, e.g., to convert ‘shopped’ in ‘She walked and shopped’ to ‘was shopping’ we had to change the aspect of ‘walk’ to make the sentence coherent.

### 4.1 Results of Experiment 2

Figure 4 and 5 show the two models' layer-wise semantic similarities of the target words in simple past/past progressive pairs for the telic and atelic groups, respectively; the dots on the figures indicate significant differences at threshold of $p < 0.05$. Unlike the previous experiment, more striking differences between BERT and GPT-2 are observed. Specifically, for BERT, the cosine similarity between the target words with different aspect features gradually increase across the layers. More importantly, the similarity in the telic group was constantly lower than the similarity in the atelic group, although the difference was only significant in the first four layers. This aligns with our hypothesis that telic verbs show difference in entailment compared to atelic verbs, and this difference is reflected by the distributional similarity between word vectors.

In contrast, GPT-2 embeddings behave in an unexpected way. The similarity in the telic group was almost always significantly higher than the atelic group across all the layers, except for the first and the final one. Additionally, the general similarity between the verbs in the two tenses is higher for GPT-2 than for BERT, and it gets very close to 1 in the later layers - which complies with Ethayarajh (2019)'s finding that the embeddings of autoregressive models are much more affected by anisotropy.

With Spearman, we observe that the scores are generally lower, which confirms the higher robustness to anisotropy of this metric. We can see that the similarities for BERT follow a similar pattern, with some additional significant differences in layer 5 and in the last, more contextualized layer; on the other hand, with GPT-2 the significance pattern is totally reversed, as it becomes significant only for the last layer. Once again, and surprisingly, telic verbs are more similar than atelic ones.

In general, the BERT model is the only one that approximates the expected behavior, with the atelic verbs having higher self-similarity in both tenses. Our results also confirm the recent finding that embeddings from autoregressive models are much weaker for similarity tasks, possibly because of anisotropy and of the lack of encoding of the information from later tokens (Springer et al., 2024). Specifically, GPT-2 similarities, indeed, appear to be more unstable across metrics and heavily affected by anisotropy (all the scores are increasingly close to 1 in the later layers).

Interestingly, in BERT, the difference tends to be significant only in the earlier, less contextualized layers. One possible explanation is that the model may be too "distracted" by the context in later layers. It has been reported that the capacity of BERT to reproduce human behavior in tasks related to verb semantics (e.g. selectional preference modeling, Metheniti et al. (2020); thematic fit estimation, Cho et al. (2021)) may improve by simply applying attention masks to the context words other than the verb and its arguments, which prevents the model from focusing on other elements of the sentence. Another possibility is that the semantics of these verbs in context is more ambiguous than traditionally assumed by linguists. In such cases, the decision about the existence of an entailment relation between progressive and simple past may not be straightforward even for humans (the results of Pruś et al. (2024) seem to go in this direction. Please also refer to the Limitations section).

We also conducted a qualitative analysis to identify cases whose similarity scores deviated from the majority examples. Specifically, we focused on BERT embeddings from layer 4, which was the last layer for which the difference in similarity was significant for both metrics. We defined outliers as data points with a z-score lower than -2 or higher than +2. Interestingly, we found no outliers for the telic group, while several outliers in the atelic group were found.

We further examined these outliers by projecting their past progressive form onto the three property scales, and found that besides being low in telicity, they generally have high durativity values (see also Figure 6 in the Appendix). Therefore, the conversion into the simple past form not only made them more 'bounded' by a natural end (i.e. increase in telicity), but also shortened their duration (i.e. decrease in durativity), which in turn led to lower similarity between the two aspectual forms. This finding is supported by an examination of the contexts of these outliers. For example, 'shop' in 'We were shopping in village stores as we went along, and my diary lists items of food bought rather than consumed' has low telicity and high durativity, but it has high telicity and low durativity in its simple past counterpart, as the former suggests that the shopping may last for the whole walk, while the latter suggest that they might be several times of quick shopping. Thus, in such cases telicity is not the only determinant of verb behaviour: the context might coerce the verb into wider meaning changes.

## 5 Conclusion

In our study, we presented an analysis of the contextualized verb embeddings of BERT and GPT-2 to assess to what extent they encode semantic distinctions related to the three aspectual properties of stativity, telicity, and durativity. Our first experiment, making use of the technique of the projection on a semantic scale by Grand et al. (2022), showed that both models could consistently distinguish verbs with different values for stativity and telicity, but faced more challenges with durativity, and gave less consistent results. To our knowledge, this study is the first that applies the method of semantic scales to analyse features of verb semantics.

As an additional contribution, we used the distributional similarities between the simple past and the past progressive of telic and atelic verbs to 'recreate' the Imperfective Paradox (Dowty, 1977) in a contextualized vector space. We showed that only the BERT model in the early layers reflects the distinction proposed by the theory – Progressive forms of atelic verbs, which entail their simple past, are more similar to the simple past than the corresponding forms of telic verbs.

## Limitations

Our work suffers from some obvious limitations: first of all, we run our experiments on English, so we cannot be sure that Transformer models for other languages would show similar patterns in encoding aspect properties; secondly, we focused on two types of architectures, BERT and GPT-2, but due to the limitations of our computational resources we could not test the more recent Large Language Models (Wei et al., 2022).

Finally, both of our experiments assume binary distinctions in natural language semantics, with regards to the aspect properties in Experiment 1 (stative vs. dynamic verbs, telic vs. atelic, punctive vs. durative) and with regards to the entailment in Experiment 2 (either the past progressive of a verb entails its simple past, or it does not). However, this is likely to be just a simplifying assumption: for example, the ratings collected by Pruś et al. (2024) suggest that humans tends to disagree about the entailments of verbs with the same telicity features. Future studies on the topic might need to adopt a perspectivist approach to account for differences in human semantic intuitions (Cabitza et al., 2023).

## Acknowledgements

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

## References

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. In *Proceedings of AAAI*.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense Event Ordering with a Multi-pass Architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, and Liping Jing. 2021. Probing BERT in Hyperbolic Spaces. *arXiv preprint arXiv:2104.03869*.

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

Emmanuele Chersoni, Rong Xiang, Qin Lu, and Chu-Ren Huang. 2020. Automatic Learning of Modality Exclusivity Norms with Crosslingual Word Embeddings. In *Proceedings of *SEM*.

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*.

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*.

Evgeniia Diachek, Sarah Brown-Schmidt, and Sean M Polyn. 2023. Items Outperform Adjectives in a Computational Model of Binary Semantic Classification. *Cognitive Science*, 47(9):e13336.

David R Dowty. 1977. Toward a Semantic Analysis of Verb Aspect and the English 'Imperfective' Progressive. *Linguistics and Philosophy*, pages 45–77.

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

Todd R Ferretti, Marta Kutas, and Ken McRae. 2007. Verb Aspect and the Activation of Event Knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(1):182.

Todd R Ferretti, Ken McRae, and Andrea Hatherell. 2001. Integrating Verbs, Situation Schemas, and Thematic Role Concepts. *Journal of Memory and Language*, 44(4):516–547.

Annemarie Friedrich and Damyana Gateva. 2017. Classification of Telicity Using Cross-linguistic Annotation Projection. In *Proceedings of EMNLP*.

Annemarie Friedrich and Alexis Palmer. 2014. Automatic Prediction of Aspectual Class of Verbs in Context. In *Proceedings of ACL*.

Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. *arXiv preprint arXiv:1901.05287*.

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.

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

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of ACL*.

Thomas Kober, Malihe Alikhani, Matthew Stone, and Mark Steedman. 2020. Aspectuality Across Genre: A Distributional Semantics Approach. In *Proceedings of COLING*.

Thomas Kober, Sander Bijl de Vroe, and Mark Steedman. 2019. Temporal and Aspectual Entailment. In *Proceedings of IWCS*.

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*.

Tal Levy, Omer Goldman, and Reut Tsarfaty. 2023. Is Probing All You Need? Indicator Tasks as an Alternative to Probing Embedding Spaces. In *Findings of EMNLP*.

Chenxin Liu and Emmanuele Chersoni. 2023. On Quick Kisses and How to Make Them Count: A Study on Event Construal in Light Verb Constructions with BERT. In *Proceedings of the EMNLP Workshop on Analysing and Interpreting Neural Networks (BlackBoxNLP)*.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT Meets Construction Grammar. In *Proceedings of COLING*.

Eleni Metheniti, Tim Van de Cruys, and Nabil Hathout. 2020. How Relevant Are Selectional Preferences for Transformer-based Language Models? In *Proceedings of COLING*.

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*.

Marc Moens and Mark Steedman. 1988. Temporal Ontology and Temporal Reference. *Computational Linguistics*.

Mattia Proietti, Gianluca E Lebani, and Alessandro Lenci. 2022. Does BERT Recognize an Agent? Modeling Dowty's Proto-Roles with Contextual Embeddings. In *Proceedings of COLING*.

Katarzyna Pruś, Mark Steedman, and Adam Lopez. 2024. Human Temporal Inferences Go Beyond Aspectual Class. In *Proceedings of EACL*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT. In *Proceedings of *SEM*.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2024. Repetition Improves Language Model Embeddings. *arXiv preprint arXiv:2402.15449*.

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

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, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*.

Zeno Vendler. 1957. Verbs and Times. *The Philosophical Review*, 66(2):143–160.

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

Shaonan Wang, Yunhao Zhang, Weiting Shi, Guangyao Zhang, Jiajun Zhang, Nan Lin, and Chengqing Zong. 2023. A Large Dataset of Semantic Ratings and its Computational Extension. *Scientific Data*, 10(1):106.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent Abilities of Large Language Models. *arXiv preprint arXiv:2206.07682*.

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

## A Verb Lists for Experiment 1

The verbs selected for Experiment 1 -divided into States, Activities, Accomplishments and Achievements- can be found in Table 1.

## B Qualititative Analysis of Experiment 2

As a complement to the final qualitative analysis in Section 4.1, Figure 6 shows an illustration of the projection of the embeddings of Experiment 2 onto the three semantic scales that we used for Experiment 1. Outliers are displayed in red.

| **State** | **Activity** | **Accomplishment** | **Achievement** |
|---|---|---|---|
| admire | dance | construct | discover |
| cherish | play | compose | recognize |
| dislike | jog | win | reach |
| fear | swim | deliver | spot |
| perceive | draw | encode | quit |
| pertain | sing | bond | forfeit |
| savor | cook | rebuild | explode |
| wish | travel | harvest | solve |
| disagree | study | decorate | die |
| deny | read | complete | notice |
| exist | run | bake | arrive |
| lack | chat | translate | find |
| concern | explore | repair | retire |
| depend | listen | fall | cure |
| equal | cycle | illustrate | hire |
| involve | push | produce | espouse |
| possess | hunt | train | score |
| rely | knit | freeze | break |
| signify | garden | thrive | invent |
| vary | exercise | drown | crack |
| value | sketch | organize | finalize |
| hope | juggle | renovate | overcome |
| weigh | weave | navigate | disappear |
| regret | drift | install | detect |
| know | browse | educate | unlock |
| appear | shop | cultivate | depart |
| imply | wait | assemble | ignite |
| matter | daydream | migrate | collide |
| include | hike | generate | elect |
| respect | fish | formulate | vanish |
| appreciate | wander | activate | baptize |
| resemble | babble | unveil | capture |
| contain | shiver | fabricate | resign |
| desire | walk | distill | convince |
| envy | glow | master | enlist |
| remember | lounge | establish | marry |
| forget | march | restore | quantify |
| mean | quarrel | digitize | provoke |
| believe | drive | synthesize | succumb |
| have | whisper | innovate | withdraw |
| suspect | celebrate | craft | originate |
| adore | drum | demolish | conquer |
| understand | giggle | export | divorce |
| belong | hum | forge | emerge |
| doubt | nap | launch | hop |
| owe | guard | implement | erupt |
| seem | rehearse | refurbish | plunge |
| prefer | watch | paint | shatter |
| consist | sail | upgrade | topple |
| need | relax | recover | unravel |

Table 1: Verb list for Experiment 1

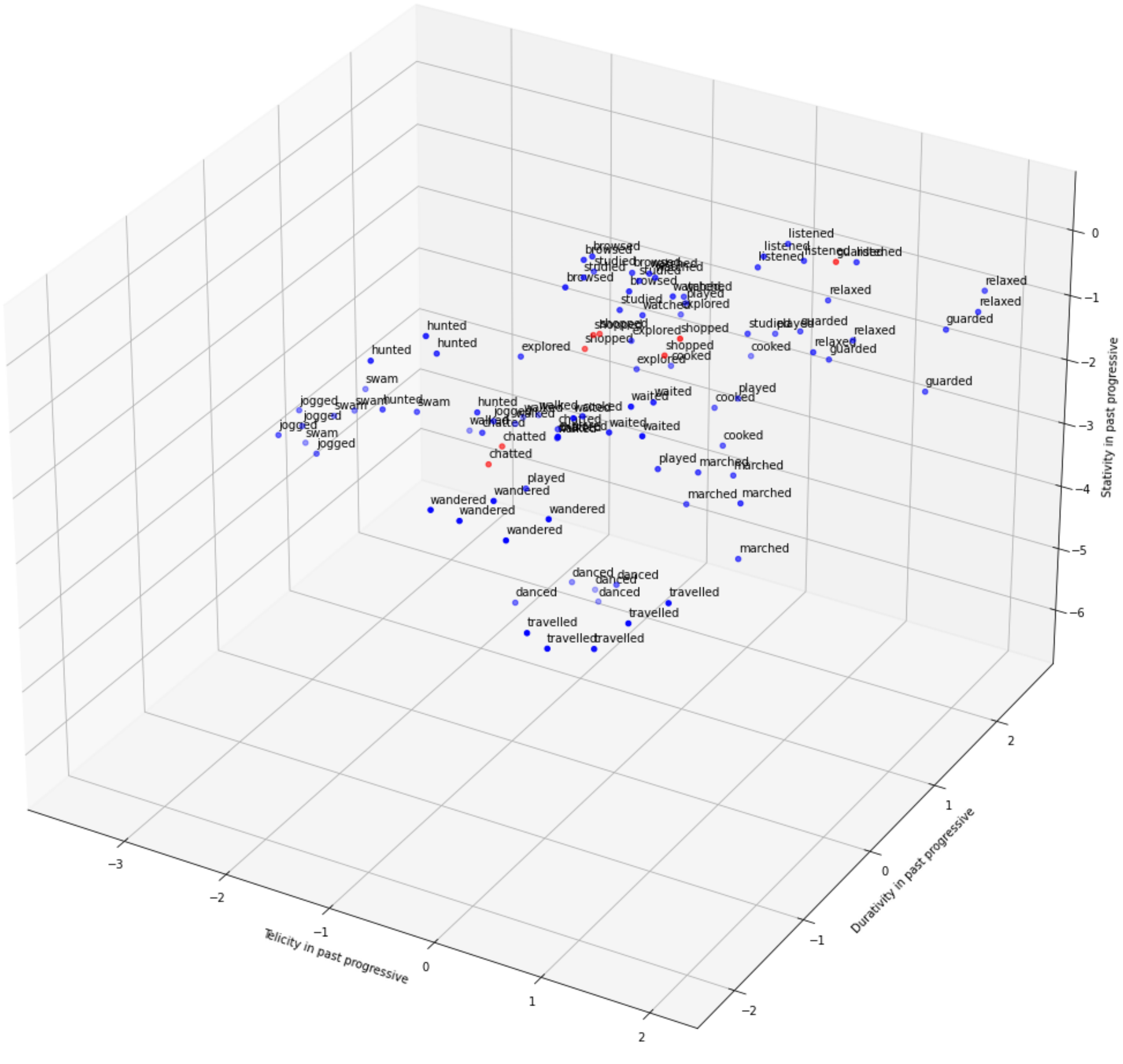

Figure 6: Visualization of aspect features of verbs in past progressive form, the red dots stand for the outliers