Human Temporal Inferences Go Beyond Aspectual Class

Katarzyna Pruś and **Mark Steedman** and **Adam Lopez**
School of Informatics
University of Edinburgh
Edinburgh, Scotland
k.prus@ed.ac.uk, {steedman, alopez}@inf.ed.ac.uk

Abstract

Past work in NLP has proposed the task of classifying English verb phrases into situation aspect categories, assuming that these categories play an important role in tasks requiring temporal reasoning. We investigate this assumption by gathering crowd-sourced judgements about aspectual entailments from non-expert, native English participants. The results suggest that aspectual class alone is not sufficient to explain the response patterns of the participants. We propose that looking at scenarios which can feasibly accompany an action description contributes towards a better explanation of the participants' answers. A further experiment using GPT-3.5 shows that its outputs follow different patterns than human answers, suggesting that such conceivable scenarios cannot be fully accounted for in the language alone. We release our dataset to support further research.

1 Introduction

Aspect is a linguistic category concerned with how actions, as described by verb phrases, unravel over time. **Situation aspect**[1] refers to the underlying semantic property of a verb phrase. For example, *to sit* is different from *to pack* in that there is a natural point at which packing is complete, but there is no such point for sitting. Situation aspect is often defined in terms of three properties: stativity, telicity and durativity. The combination of values which a verb phrase takes for each of these properties is what decides its belonging to a particular aspectual class. For example, *I love you* is stative, atelic (has no pre-determined enpoint) and durative (spans across a period of time). By contrast, *I caught the ball* is dynamic, telic (has a clearly defined endpoint) and punctual (occurs instantly).

Work on automatic aspectual classification in English has been motivated as a pre-requisite for Natural Language Understanding (NLU) in cases where temporal reasoning is required (Siegel and McKeown, 2000; Friedrich and Gateva, 2017; Kober et al., 2020; Friedrich et al., 2023). Consider the two examples:

(1) *I was listening to music* $\rightarrow$ *I listened to music*
(2) *I was winning the race* $\not\rightarrow$ *I won the race*

The entailment in (1) and lack of entailment in (2) are explained by the action descriptions belonging to different aspectual classes. Having said that, there is no empirical evidence that aspectual class is helpful for NLU tasks in practice, where more pragmatic *inferences* are favoured over strict logical *entailments* (Pavlick and Kwiatkowski, 2019).

This paper asks whether the role that aspectual classification is described to play on logical entailments is reflected in crowd-sourced data, where participants were allowed to take a less formal approach. We designed a survey with examples of verb phrases turned into sentence pairs: one in past progressive and one in past simple, like the pairs in examples (1) and (2) above. We gather participants' judgements on whether the past simple can be inferred from the past progressive. What our survey clearly shows is that aspectual classification does not best explain how non-expert participants reason about eventuality. Instead, the results are better explained by considering possible scenarios which can accompany any given verb phrase, for example whether an an action is likely to be interrupted or not. We release the the anonymised survey responses to enable further research.[2]

Finally, with an experiment using GPT- 3.5, we show that Large Language Models (LLMs) do not capture the answer patterns seen in participant answers. We speculate that this is because people's strategies for reasoning about events are an example of 'understanding' that people gain through

[1] This semantic property is often referred to in the literature as **lexical aspect**. However, especially in English, it is a property of an entire clause rather than an individual verb (Friedrich et al., 2023). We follow Bender and Lascarides (2019, p. 99) in using the term *situation aspect* to reflect the nature of this category being both lexical and compositional.

[2] https://github.com/patarzynak/beyond-aspectual-class

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1913–1923

physical experiences of the world and cannot be modelled by linguistic material alone (Bender and Koller, 2020).

2 Background

Situation aspect is a semantic property of a situation description. In literature, we often see it characterised in terms of these three properties (Moens and Steedman, 1988; Peck et al., 2013):

stativity	taking values **stative** or **dynamic**
telicity	taking values **telic** or **atelic**
durativity	taking values **durative** or **punctual**

Stativity refers to the distinction between **states** and **events**[3], where remaining in a stative situation does not require any effort, whilst remaining in a dynamic situation requires effort (Comrie, 1976, p. 49). For example, *I am* is a state and *I run* is an event. The difference between states and events is less significant when talking about the past, than it is when talking about the present (Leech, 1971), which is why in this paper the focus is on dynamic situations only.

Telicity is a concept that describes whether a situation has a culmination point. A telic situation leads up to a necessary endpoint, beyond which it cannot continue, whether that point has been reached or not (Comrie, 1976, p. 45). Conversely, an atelic situation lacks such a pre-defined finish, whether it's still in progress or not. For example, *to drown* is telic and *to dance* is atelic.

Durativity refers to the fact that certain situations span over a period of time regardless how long or short (durative), whilst others are instantaneous (punctual; Comrie, 1976, p. 41). For example, *to run* is durative, whilst *to die* is punctual.

Valid combinations of aspectual features is what defines the distinction between different situation aspect categories, often referred to as Aktionsarts (Vendler, 1967), which offers alternative terminology for talking about aspectual classification (as summarised in Table 1).

2.1 The Imperfective Paradox

The task proposed in this paper is inspired by the Imperfective Paradox as analysed by Dowty (1979). He observes that for Activities the past progressive entails the simple past, but for Accomplishments it doesn't. For example, *I was walking* entails *I walked*, but *I was building a house* does not entail *I built a house*. Dowty (1979) talks only of Activities and Accomplishments, that is durative predicates. When it comes to punctual predicates, and specifically Achievements, their progressive forms are told to be coercing them into a different aspectual reading by enforcing a durative reading (Moens and Steedman, 1988; Pustejovsky, 1991). Regardless, a sentence including a past progressive of an Achievement does not entail its past simple (e.g. *I was winning* $\not\rightarrow$ *I won*). Therefore, it is telicity that is widely pointed at as the feature, which draws the line between the predicates that evoke this entailment and the predicates that don't (Lascarides, 1991; Rastelli, 2019; Zucchi, 2020).

stative	atelic	durative	**State** *to know*
dynamic	telic	punctual	**Achievement** *to die*
		durative	**Accomplishment** *to build a house*
	atelic	punctual	**Act** *to sneeze*
		durative	**Activity** *to dance*

Table 1: Aktionsart terminology with examples.

3 Related Work

3.1 Aspect Classification

Friedrich et al. (2023) provide a comprehensive overview of works investigating aspect in the context of its computational applications. Our work is inspired by a particular line of enquiry, which focused on labelling verbs, clauses or sentences with their aspectual properties and then automating the recognition of these aspectual properties as a classification task (Siegel and McKeown, 2000; Friedrich and Palmer, 2014; Friedrich and Gateva, 2017; Kober et al., 2020; Alikhani et al., 2022). Various iterations of this task include classifying the verb types in isolation (Siegel and McKeown, 2000), the verbs in context (Friedrich and Palmer, 2014) or focusing on telicity alone (Friedrich and Gateva, 2017). All of these papers present an approach to gathering gold labels through expert annotation. Each sourced their example from one text genre only (Kober et al., 2020). As Alikhani

[3] Comrie (1976, p. 51) draws a distinction between *events* and *processes*, both being defined as dynamic situations but viewed from different perspectives. Throughout this paper, we use the terms process, event, and dynamic situation interchangeably to refer to any non-stative situation.

and Stone (2019) show for image captions, if you narrow down your dataset to one genre, it will tend to be dominated by verb phrases representing only a narrow set of possible aspectual features.

All of these papers motivated the task as necessary for reasoning about temporal relations. However, those studies are implicitly only addressing formal logical reasoning. In our study, we want to investigate this motivation in a more 'commonsense reasoning' framework — by looking into the relation between aspectual class and the inferences that non-experts make. Moreover, some of these papers rejected the examples that didn't yield sufficient inter-annotator agreement from their respective datasets. Here, we want to highlight that disagreement can be informative and therefore it is worth analysing the examples that caused it.

3.2 Natural Language Inference

Natural Language Inference (NLI) is a task, where given a pair of sentences — **premise** and **hypothesis** — one is asked to say whether the premise *entails* the hypothesis or not. There are two approaches to labelling such sentence pairs: with two labels (*entailment* and *non-entailment*) or with three (*entailment*, *neutral*, *contradiction*).

Kober et al. (2019) introduce a highly curated dataset for the entailment detection task that specifically focuses on temporality and aspect. The labelling was done by two expert annotators, who noted that 'everything appeared to be uncertain'. This resonates with our idea that there is room for disagreement in people's judgements of the Imperfective Paradox entailment.

As we intend to crowd-source entailment judgements from non-expert participants, we expect to observe that some examples will elicit a mixed response. When it comes to such disagreement in NLI tasks, Pavlick and Kwiatkowski (2019) propose that it can reflect varying approaches to resolving uncertainties and therefore cannot be dismissed as noise. We embrace that conclusion in our experiment design. With our study, we want to find out which examples of predicates caused uncertainties and attempt to interpret makes them so uncertain.

4 Human Experiment

In our experiment, we presented participants with pairs of sentences constructed from one base verb phrase: one in past progressive and one in past simple. We use the 'imperfective paradox as telicity test' setting to build an NLI type task and present it to non-expert participants. Mixed and majority 'incorrect' responses could signal one or both of two things. First, that without sufficient context telicity remains under-specified. Second, that the degree to which telicity is a factor in temporal reasoning is less significant than commonly proposed.

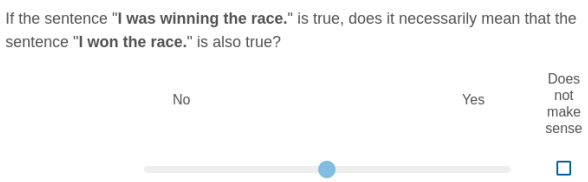

Figure 1: The slider interface used to gather participant's answers. The closer to the edge of the slider, the more certainty in the answer is expressed. Participants were instructed to use the 'Does not make sense' checkbox if they deemed any or both of the sentences ungrammatical.

4.1 Experiment Design

The participants are presented with examples in the form of a question, which varies only in a predicate *X*: *If the sentence **I was Xing** is true, does it necessarily mean that the sentence **I Xed** is also true?* This can be seen as a variation of the NLI task, where answer 'Yes' signifies entailment (*I was Xing* $\rightarrow$ *I Xed*) and answer 'No' signifies non-entailment (*I was Xing* $\not\rightarrow$ *I Xed*). The further subdivision of the lack of entailment into 'neutral' and 'contradiction' is not relevant in this case, as the two sentences will always share the use of subject *I* and predicate *X*.

As explained in Section 2, the nominally telic examples are expected to elicit a 'No', whilst atelic examples are expected to elicit a 'Yes'.[4] We wanted to capture participants' level of certainty as well as their yes/no answer, so we presented them with a slider labelled 'No' on the left and 'Yes' on the right. The slider mapped the participant's answer to a value from -50 (for certain 'No') to 50 (for certain 'Yes'), with values near 0 meaning that the participant is not confident in either answer. The participants would not be able to see the exact numeric value of their answer. A box marked 'Does not make sense' was included and participants were instructed to use it if they thought that any of the

[4]Note that atelic+punctual events are rare, and it is disputed whether they are truly punctual (Comrie, 1976, p. 42) or how to interpret their progressive form (Moens and Steedman, 1988).

Base Form: *study history*	**Past Progressive:** *I was studying history.*	**Past Simple:** *I studied history.*
Question: If the sentence *I was studying history* is true, does it necessarily mean that the sentence *I studied history* is also true?		

Table 2: An example of how the stimuli were generated by inputting the past tense forms into a question template.

two sentences was not grammatically valid. The design of an individual question page is illustrated in Figure 1. The questions were presented on a page individually; a participant needed to provide an answer before being allowed to move on to the next question.

The list of sentence pairs used in this study was collated by sourcing some of the examples proposed in the linguistics literature (Dowty, 1979; Comrie, 1976; Lascarides, 1991; Glasbey, 2004; Rastelli, 2019), events randomly drawn from the ATOMIC 2020 knowledge graph (Hwang et al., 2020), and manual alterations of the pre-selected examples. The examples were kept purposefully short and structurally simple — they present either a two-place predicate or a one-place predicate with one modifier. Moreover, all of the examples use the first person singular subject pronoun ***I*** to control for the variability that might stem from the use of different subject pronouns (or verb subjects in general; Brunyé et al., 2009). Example of how the sentence pairs are constructed from their base forms can be seen in Table 2. For brevity, we will henceforth refer to any particular example by using its base form.

Altogether the stimuli collection contains 50 examples. Each example has been annotated for telicity and durativity by one of the authors and each annotation has been verified by one more expert annotator. An agreement has been reached for most of the examples — only six caused initial disagreement amongst expert annotators. Further discussion resolved some of the initial disagreement. Nevertheless, we decided to highlight those examples in our analysis to see if initial expert disagreement can predict participant disagreement. Henceforth, the aspectual class of these examples will be referred to as 'contested'.

Two of the examples — one atelic+durative (*play at the park*) and one telic+punctual (*win the race*) — were presented to all of the participants at the start of the survey. At random, one would be presented on the instruction page as a trial example and the other would come up after as the first 'real' example. This is to minimise the risk of priming effects in our results. The answers provided to these examples are excluded from our analysis.

The remaining examples were divided into two groups. All of these examples are included in Figure 2, where the left column represents one group and the right column represents the other group. Each participant would only answer one group of questions. This was done to limit the time it takes to fill out the survey. The questions were presented to the participants in random order, intertwined with 3 attention checking questions. Randomised order, again, minimised the risk of priming effects.

4.2 Open-ended survey of approaches

At the end of the survey, the participants were asked about how they approached answering the questions. Providing this input was optional and they were given a short free-text box in which they could type their answer.

We hypothesise that despite the existence of the theoretical 'correct answer', some of the examples will show the participants' responses to be divided or contrary to the 'correct answer'. It is understood that strategies employed by crowd annotators to NLI tasks can easily result in answers different from those dictated by strict logical reasoning (Dagan et al., 2006; Bowman et al., 2015; Williams et al., 2018). Here, examples classed as 'contested' are particularly good candidates to elicit disagreement amongst participants, as they have already caused disagreement amongst experts. The purpose of this experiment is to identify examples of verb phrases which elicited 'wrong answer' or caused inter-annotator disagreement. Focusing on those examples, we can look for possible explanations at the intersection of current theories about situation aspect with insights from research on NLI.

4.3 Participants

The participants were recruited via Prolific.co. They were pre-screened to include native English speakers. A total of 120 participants were recruited, with 108 included in the analysis, and the remaining 12 rejected for failing screening criteria or at-

tention checks. The vast majority of the included participants (92%) reported English to have been the only language spoken at their home before the age of 6 and the only language in which they consider themselves fluent. The remaining 8%, whilst satisfying the same criteria of being brought up and still predominantly using English, have reported being exposed to another language during childhood or becoming fluent in an additional language in their adulthood. The participants were compensated 2.50 GBP for completing the survey and the median completion time was 7 minutes. The participants have all agreed that their anonymised responses will be used in academic publications and presentation and can be made publicly available.

4.4 Results

Distributions of answers can be seen as the upper bars in Figure 2. To plot these distributions, we mapped the slider values into 5 intervals. These intervals are (in order from signifying a 'Definite No' answer to 'Definite Yes' answer): $[-50, -31][-30, -11][-10, 10][11, 30][31, 50]$. To simplify our discussion, we note that there are three observed types of participants' answers distributions: skewed towards 'No', bimodal, and skewed towards 'Yes'. For the atelic phrases, answer distributions for almost all of the examples show participants' preference for 'Yes', which is in line with the theoretical prediction. The notable exception here is *enjoy your company*, which we will discuss in more details in the following section. For the telic phrases we observe a mixture of responses — all three distributions were observed.

Had telicity driven temporal reasoning amongst non-experts in the way that has been assumed, we would have observed a majority of the telic examples to have answer distribution skewed towards 'No'. This is clearly not the case. Moreover, durativity cannot be used as an explanation for why certain telic predicates have different distributions of answers than others. Finally, the disagreement between expert annotators was not a good predictor of participants' answers distribution being bimodal.

5 Discussion

At first glance, it should not be surprising that the answers of the participants do not match with the category-based predictions, given our survey's setup. The NLI literature observes that without providing people with explicit annotation instructions on whether they are allowed to use their real world knowledge or consider any additional context from outside the text material, they tend to take on different strategies to resolving uncertainties (Zaenen et al., 2005; Manning, 2006; Pavlick and Kwiatkowski, 2019). In our set-up, the participants were free to conjure up their own contexts and take on any approach they like, which is inline with how labels are gathered in most modern NLI datasets (Bowman et al., 2015; Williams et al., 2018). In this section we zoom in on particular examples to explore how such possible scenarios could explain the participants' answers distribution.

5.1 Atelic

We find it telling that all bar one example of atelic verb phrases have distributions very strongly skewed towards yes. Indeed, in the light of above it is noteworthy just how high the agreement was amongst participants for these examples. It shows that people have a very strong intuitive understanding that any time span of an Activity can be divided into shorter intervals, where each such interval is an instance of the same Activity. For example, if you engaged in listening to music for 30 minutes, each 1-minute interval within that 30-minute span was also an instance of listening to music. A consequence is that it is impossible to come up with scenarios where interrupting an Activity makes the action incomplete. In our example, if you have not predetermined to listen to music for any particular amount if time, but you simply got interrupted by a phone call after 30 minutes — you would still say that you listened to music.

It is therefore particularly interesting to observe that amongst atelic predicates, *enjoy your company* was the only divisive one — its answer distribution can be seen in Figure 2. To see why it is different, consider a sentence 'I was enjoying your company, until you offended me'. In this example, the statement 'I was enjoying your company' is explicitly true but interrupted by the 'you offended me' event. The speaker's reflection of the entire process causes them to say that the statement 'I enjoyed your company' is untrue. This ability to be negated by an interruption seems unique to 'enjoy' amongst other Activities. We note that it is not telicity — by any definition *enjoy your company* is atelic — but the possibility of conjuring this specific scenario that divided the answers of participants.

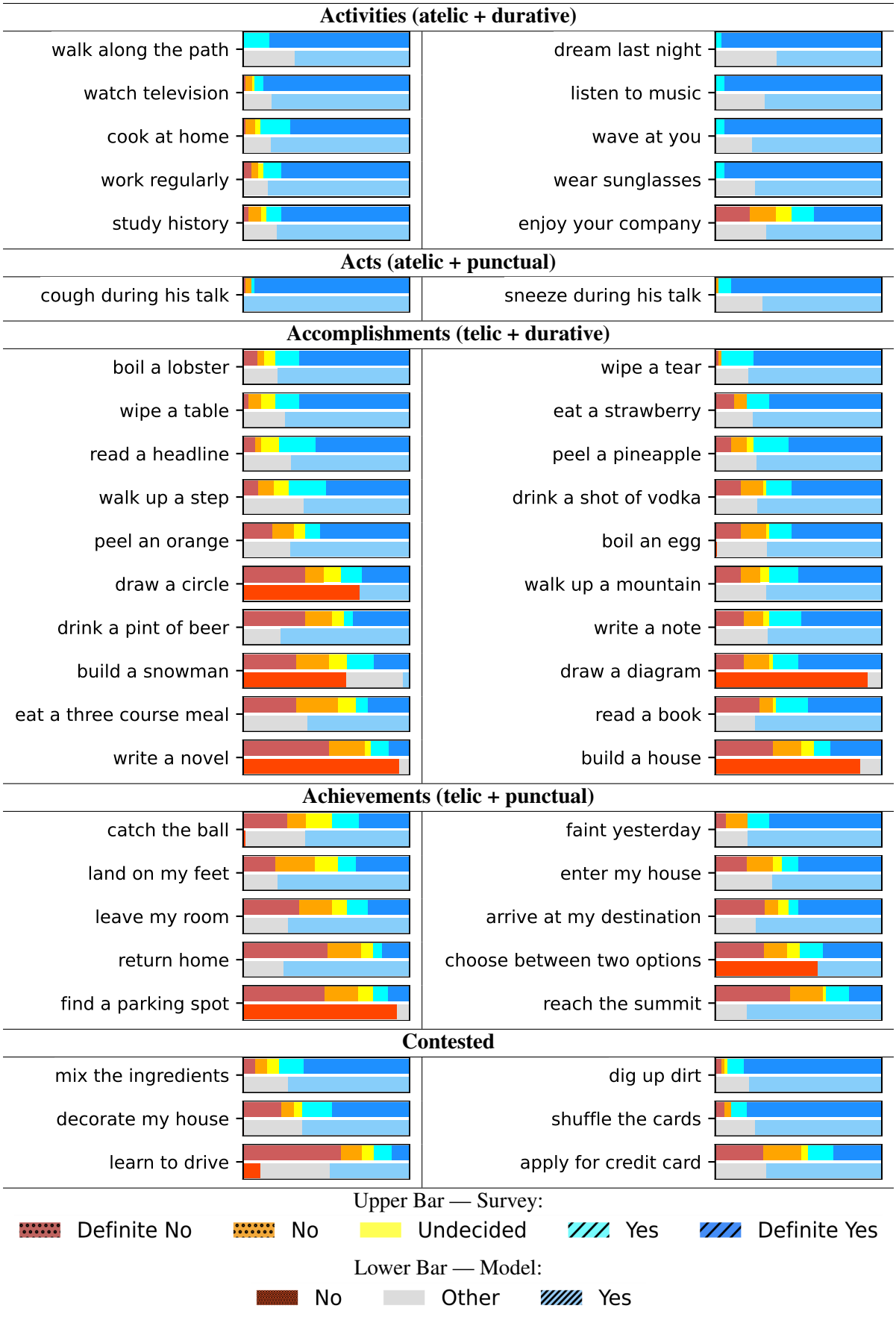

Figure 2: Results of our experiments. For each example, the upper bars present the distributions of participants' answers to our survey. The lower bar present how much probability GPT-3.5 assigned to each answer. The left column shows examples shown to one group, whilst the right column shows examples shown to the other group of participants.

5.2 Contested

Beyond the atelic predicates, the picture is less clear. Let's start by taking a look at the cases where the consensus between the two expert annotators was not immediate. It is important to remember that whilst expert annotators are as close to 'perfect logicians' as available, they are still prone to conjuring contexts that were not present in the text material. Having said that, we see two explanations for why even a 'perfect logician' might struggle to give a definitive 'Yes' or 'No' answer for some of these examples.

First, a verb phrase can be ambiguous. Let's consider the predicate *mix the ingredients*. The verb *to mix* actually functions in English both as an Activity (atelic+durative) and as an Accomplishment (telic+durative), and adding the object *the ingredients* is not sufficient context to disambiguate it. In other words, it is possible to focus on the process (synonymous with *to stir*) as well as the culmination point (synonymous with *to combine*).

Second, the linguistics literature has long noted the difficulties with assigning telicity to so-called **degree achievements** (DAs; Hay et al., 2001). Common examples of such DAs include *widen*, *straighten*, *dry*. The main characteristic of such DAs is that they can be considered both 'complete' and 'incomplete' at their intermediate stages. For example, a batch of laundry you hang out in the morning can be drier in the evening, meaning it *has dried (a bit)* and yet *hasn't dried (completely)*. We believe that our examples of *decorate my house* and *shuffle the cards* are indeed examples of DAs.

It is worth noting that these theoretical difficulties did not necessarily translate into predictors of bimodal participant answer distributions. Out of the three examples mentioned here, only *decorate my house* elicited a mixed response amongst the participants. Whilst both *mix the ingredients* and *shuffle the cards* collect some 'No' answers, their distribution is strongly skewed towards 'Yes'.

5.3 Telic

As mentioned before, all three types of distributions were observed amongst the telic predicates. Despite 'No' being the theoretical 'correct answer' for any telic example, only 4 of them (out of the 30) had distributions strongly skewed towards 'No': *write a novel*, *find a parking spot*, *reach the summit* and *return home*. It is no surprise that three of those are punctual. As mentioned in 2.1, any punctual predicate when put into a progressive form is being forced into a different meaning. However, some punctual predicates sound less natural when forced into progressive than others. In fact, a sentence *John was reaching the summit* was highlighted as grammatically incorrect by Comrie (1976). Such particularly unnaturally sounding progressive forms might have swayed the participant's choice. Having said that, this does not account for the presence of *write a novel* on this list. A possible explanation in this case, is that in practice it is quite prevalent to encounter situations where the action of writing a novel does not lead to a completion of a novel having been written.

We observe that 2 (out of 10) telic+punctual examples and 12 (out of 20) telic+durative examples have a 'Yes'-skewed distribution, whilst the remaining examples have a bimodal distribution. At least some of the 'Yes' answers to telic examples can be explained by participants adopting a pragmatic approach to inference, best illustrated by one of their free text responses:

> **Participant A:** Just weighing up the probability that the person doing the action is likely to complete the action.

For example, consider *eat a strawberry*, which is a telic predicate with a 'Yes'-skewed distribution. A non-negligible number of participants would have answered 'Yes' to this example, because in practice the action of eating a strawberry is not very likely to be abandoned. In other words, they are less likely to conjure a scenario in which the action *I was eating a strawberry* gets interrupted and therefore does not result in the strawberry having been eaten, even though such an interruption is perfectly possible in theory.

Having said that, there are participants who adopted an approach closer to formal logic:

> **Participant B:** Some actions are considered done only when they have been completed. Other actions are considered to have been done even while the action continues to be in progress.

Even participants with this approach sometimes provided answers opposed to the theoretical prediction. Consider *walk up a mountain* — a telic example to which this participant answered 'Yes'. Its well-defined endpoint, however, does not have the same 'necessity' as the endpoint of e.g. *build*

a house. For example, one can start a sentence with *I walked up a mountain* and finish it with *but not all the way up, only halfway* and some people would would find it acceptable, but some would find it objectionable. Whether one finds such a continuation to *I walked up a mountain* in principle acceptable or not is rooted in one's beliefs about the world. It is this internal model of the world, rather than categorical telicity, that likely guided them in answering the questions from our task.

What this survey clearly shows is that that neither telicity nor durativity are the main factors driving how people, particularly non-linguists, reason about eventuality. The combination of the differences between participants' approaches to the task and the differences in their beliefs of the world is a fitting explanation for the mixed responses to the nominally telic examples.

6 LLM Experiment

The training data used in the creation of any LLM contains a vast number of examples of verb use in context. However, seeing a verb in textual context is not the same as experiencing an event described by that verb. Therefore, we propose that by looking at outputs of LLMs, one can ask if the use of language alone is sufficient to account for the disagreement that certain predicates elicited amongst human participants. In other words: is the disagreement reflected in the way people talk about those events at large, or is there more to it?

To investigate this, we ran an experiment with GPT-3.5 (Brown et al., 2020). The reason for choosing GPT-3.5 is twofold. First, it is one of the more recent LLMs featured prominently in current literature. Second, unlike the newer GPT-4, it allows us to retrieve the probability values assigned to the top 5 candidates for the next predicted token. At this stage, we did not run experiments with other LLMs — our aim is not to provide a comprehensive model comparison, but rather use any LLM as a large scale language resource that implicitly encodes a multitude of possible textual contexts.

We used the text-davinci completion model. We set temperature to 0 and top_p to 1.[5] We used the following as a prompt template:

Answer the question with Y for yes and N for no. Question: If the sentence S1 is true, does it necessarily mean that the sentence S2 is also true? Answer:

where pairs of sentences from our example collection were substitutes for *S1* and *S2*. We observe the probabilities assigned to the final token produced. Amongst the top 5 candidates, we would usually observe variations of the expected answer (e.g., the model output 'Y' as per instructions, but amongst top 5 predicted tokens we observed also 'Yes' or 'yes' etc.). We therefore summed up the probabilities of all such variations. For each example we note three probability values: probability assigned to 'Yes' variations, probability assigned to 'No' variations and probability assigned to other tokens.

We would consider the model's predictions to be consistent with our survey observations if it assigned most probability (more than 0.5) to the answer towards which the distribution was skewed. We would consider the model's output as a 'mixed response' if either the probability assigned to one answer was less than 0.5 (the rest being assigned to other tokens), or if the probability assigned to both 'Yes' and 'No' was non-trivial (more than 0.1). The probability allocated by the model to 'Yes' and 'No' variations for each example are plotted as the lower bar in Figure 2.

The model seems to mirror participants' answers for some, but not all examples. It is in-line with the participants' answers, only in as much as there is no visible trend towards answering 'No' for telic examples as strict logic would dictate. Having said that, there are noticeable discrepancies between the survey results and the model experiment results.

We observe a match between the model's mixed prediction and participant's mixed answers for only a handful of examples. We observe 'mixed' model replies for only 3 examples. This is far fewer examples than the ones that resulted in a bimodal distribution amongst the participants. Of those three, *learn to drive* was 'No'-skewed amongst the participants. Moreover, seven of the model's predictions assigned most probability to 'No'. Of those, only two were 'No'-skewed amongst the participants. Finally, we observe that for some of the strongest 'No-skewed' examples amongst the participants, e.g. *return home* or *reach the summit*, the model still overwhelmingly predicts 'Yes' as an answer.

In conclusion, the results from the experiment with GPT-3.5 do not reflect either the 'perfect-logician' nor the participant's behaviour. A possible explanation is that there are limits to what

[5] We tried a few combinations of these parameters all leading to similar conclusions, so for simplicity we focus on describing the most straightforward setup.

knowledge about event structures can be captured by such a model without any experience of the physical world (Bender and Koller, 2020). Having said this, the results presented here are based on one LLM only, so our conclusions are not a definite answer, but an invitation for future research. We believe that by using LLMs as proxies for large-scale corpus analysis future research can ask interesting questions about the respective roles of textual information and physical experiences in building our 'understanding' of event structures.

7 Conclusion

It is clear from our results that whilst aspectual class can be used as a rough guide as to which inferences some subjects may draw some of the time, it is far from being the main deciding factor. We also show that predictions of GPT-3.5 are not entirely aligned with the participants' responses. This opens the door to further research into what influences human understanding of event descriptions. The examples we collected for our experiments can also be used as a dataset to explore the role of pragmatics in NLI as well as other NLU tasks.

Limitations

The main limitation of our study is its scale — at 50 examples studied it is smaller than many modern NLP works. Whilst limiting the number of examples is what allowed us to undertake a more detailed analysis of answer patterns for each individual example, it would indeed be beneficial for the research community to undertake similar experiments on a larger scale in the future. Our study shows that there is room for disagreement on the 'imperfective paradox' style questions. A larger study could investigate the magnitude of that disagreement as well the implications of such disagreement for practical applications. Similarly, our experiment with GPT-3.5 only involved one model, and so our observations should not be read as a commentary on LLMs' capabilities overall. Instead, we are hoping that this work is seen as an invitation for the community to continue research into situation aspect, with a shift from treating it as a category with an underlying ground truth label to treating it as a category that can remain under-specified on more than just a few outlier occasions.

Acknowledgements

This work was supported by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1). We thank Hannah Rohde and Vilde Reksnes for helpful discussion on the survey design; Sander Bijl de Vroe and Thomas Kober for lending their expertise on aspect and comments on early drafts; Kate McCurdy and Julie-Anne Meaney for practical pointers on survey deployment; and Sharon Goldwater, Andreas Grivas, Coleman Haley and Oli Liu for helpful comments on previous drafts of this paper.

References

Malihe Alikhani, Thomas Kober, Bashar Alhafni, Yue Chen, Mert Inan, Elizabeth Nielsen, Shahab Raji, Mark Steedman, and Matthew Stone. 2022. Zero-shot cross-linguistic learning of event semantics. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 212–224, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Malihe Alikhani and Matthew Stone. 2019. "caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Emily M. Bender and Alex Lascarides. 2019. *Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics*. Number 3 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric

Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Tad T Brunyé, Tali Ditman, Caroline R Mahoney, Jason S Augustyn, and Holly A Taylor. 2009. When you and i share perspectives: Pronouns modulate perspective taking during narrative comprehension. *Psychological Science*, 20(1):27–32.

Bernard Comrie. 1976. *Aspect : an introduction to the study of verbal aspect and related problems*. Cambridge textbooks in linguistics. Cambridge University Press, Cambridge.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, page 177–190, Berlin, Heidelberg. Springer.

David R. Dowty. 1979. *Word meaning and Montague grammar: the semantics of verbs and times in generative semantics and in Montague's PTQ*. Number v. 7 in Synthese language library. D. Reidel Pub. Co, Dordrecht ; Boston.

Annemarie Friedrich and Damyana Gateva. 2017. Classification of telicity using cross-linguistic annotation projection. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2559–2565, Copenhagen, Denmark. Association for Computational Linguistics.

Annemarie Friedrich and Alexis Palmer. 2014. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 517–523, Baltimore, Maryland. Association for Computational Linguistics.

Annemarie Friedrich, Nianwen Xue, and Alexis Palmer. 2023. A kind introduction to lexical and grammatical aspect, with a survey of computational approaches. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 599–622, Dubrovnik, Croatia. Association for Computational Linguistics.

Sheila Glasbey. 2004. Event structure, punctuality, and 'when'. *Natural Language Semantics*, 12:191–211.

Jen Hay, Christopher Kennedy, and Beth Levin. 2001. Scalar structure underlies telicity in "degree achievements". *Semantics and Linguistic Theory*, 9.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI Conference on Artificial Intelligence*.

Thomas Kober, Malihe Alikhani, Matthew Stone, and Mark Steedman. 2020. Aspectuality across genre: A distributional semantics approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4546–4562, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thomas Kober, Sander Bijl de Vroe, and Mark Steedman. 2019. Temporal and aspectual entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 103–119, Gothenburg, Sweden. Association for Computational Linguistics.

Alex Lascarides. 1991. The progressive and the imperfective paradox. *Synthese*, 87(6):401–447.

Geoffrey N. Leech. 1971. *Meaning and the English verb / Geoffrey N. Leech.* Longman, Harlow.

Christopher D. Manning. 2006. Local textual inference : It's hard to circumscribe , but you know it when you see it - and nlp needs it.

Marc Moens and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Jeeyoung Peck, Jingxia Lin, and Chaofen Sun. 2013. Aspectual classification of mandarin chinese verbs: A perspective of scale structure. *Language and Linguistics*, 4:663–700.

James Pustejovsky. 1991. *Cognition*, 41(1-3):47–81.

Stefano Rastelli. 2019. The imperfective paradox in a second language: A dynamic completion-entailment test. *Lingua*, 231:102709.

Eric V. Siegel and Kathleen R. McKeown. 2000. Learning Methods to Combine Linguistic Indicators: Improving Aspectual Classification and Revealing Linguistic Insights. *Computational Linguistics*, 26(4):595–628.

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University Press.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 31–36, Ann Arbor, Michigan. Association for Computational Linguistics.

Sandro Zucchi. 2020. Progressive: The imperfective paradox. *The Wiley Blackwell Companion to Semantics*, pages 1–32.

A Instructions

Below we include the full, verbatim text of the instructions provided to the participants on the first page of the survey — following the consent form page. Please note that this text uses *play at the park* as a practice example. Half of participants in each group were shown a version of this using *win the race* as a practice example instead.

INSTRUCTIONS:

Please read the instructions now - they will not be repeated on further pages and there will not be an option to come back to this page.

In this survey you will be presented with pairs of sentences. For each pair you will be asked to **assume that the first sentence is true**. Using your best judgement, we ask you to **indicate whether the second sentence is therefore also true**. Use the slider to indicate the confidence in your judgement - **the further away from the middle you place the slider, the more confident you are in your judgement**.

PRACTICE EXAMPLE:

Please answer this question:

If the sentence **"I was playing at the park."** is true, does it necessarily mean that the sentence **"I played at the park."** is also true?

(Here is a slider as illustrated in Figure 1)

BEAR IN MIND:

If either sentence is not interpretable or either sentence is grammatically incorrect - tick the "Does Not Make Sense" box.

If both sentences are sensible and correct, please provide an answer with the slider. Please note, that even if you are confident that you want to leave the slider in the middle - you will have to move it slightly and ultimately put it back in the middle, before you'll be able to press "Next".

There will be **three attention checking questions** in this survey - they will vary in structure from the description above.