

Linguistically-informed Computational Approaches to Verbal Aspect

Bachelor Thesis

Date, Year

Samuel Innes

innes@cl.uni-heidelberg.de

Institut für Computerlinguistik
Ruprecht-Karls-Universität Heidelberg

Supervisor Prof. Dr. Michael Herweg

Reviewer Prof. Dr. Anette Frank

Abstract

Abstract in English

Zusammenfassung

Zusammenfassung auf Deutsch

Contents

List of Figures	VI
List of Tables	VII
1 Introduction	1
2 Aspectology: an introduction	3
2.1 A (short) phenomenology of aspect	3
2.1.1 Lexical vs grammatical aspect	4
2.1.2 Lexical aspect	5
2.1.2.1 Some attempts at event classification	8
2.1.3 Grammatical aspect	9
2.2 Aspect in Slavic languages	10
2.2.1 Slavic Perfective-Imperfective distinction	11
2.3 Formal models of aspect (or put this in related work?)	11
3 Related Work	12
3.1 Previous computational approaches to modeling aspect	12
3.1.1 Situation entities!!	12
4 Methods	13
4.1 Which aspect types do I use?	13
4.1.1 UMR	13
4.1.2 How does the Slavic Perfective-Imperfective relate to UMR classes . . .	15
4.2 Use of approximative LMs as sources of linguistic knowledge??	16
4.3 Where do LLMs fail?	16
5 Experiments	17
5.1 Dataset creation	17
5.1.1 Dataset annotation with LLMs	17

5.1.2	Language model fine-tuning	21
5.1.2.1	Ablation (aspectualisation and contextualisation)	21
5.1.2.2	Combined Rus/Eng model	21
5.1.3	Probing	21
5.1.4	Aspect latent space	21
5.1.4.1	Verb of motion clustering	21
5.1.4.2	Prefix clustering	21
6	Discussion	22
7	Conclusion	23
A	Appendix	24
A.1	Experiments with ChatGPT	24
A.1.1	Temporal reasoning	24
	Bibliography	25

List of Figures

1	A set-theoretical representation of the relationship between time, situation and event (<i>left</i>) along with the categorisation of the three tempora phenomena discussed (<i>right</i>).	4
2	Comrie's classification of aspectual oppositions. Reproduced from Comrie [1976].	10
3	UMR aspect classification tree [Jens Van Gysel and Xue, 2022]	14

List of Tables

1	Representation of (1),(2),(3) and (4) with regards to the reference time referred to by the utterance and inherent run-time of the event. Concept adapted from van Hout [2016].	7
2	Classification of Vendlerian event types by binary aspectual paramters [Smith, 1991]	9
3	Some of the example verb phrases given in Vendler [1957] for the classification of events.	9
4	Comparison of aspectual classes. Adapted and extended from Egg, Prepens, and Roberts [2019].	15
5	Aspect classes of annotated verbal events in the UMR dataset.	18

1 Introduction

Motivation

Despite being one of the most studied areas in linguistics, I COULDNT FIND ANY statistical approaches to aspect.

Why aspect?

One may justifiably beg the question why a “computational approach” to aspect (OR INDEED TO ANY PROBLEM IN THEORETICAL LINGUISTICS) is necessary or indeed useful: in an age of LLMs

The use of such a study comes down to the purpose of computational linguistics as an area of study. Computational linguistics has changed a lot since its conception in the mid 20th century, at some points being closer to linguistics and at others (arguably including right now) being closer to computer science. However the position of the field lying at the intersection between more well-established and well-defined areas of study has led to a fruitful exchange of ideas between the disciplines.¹

One reason is of course the contribution to the linguistic community: computational approaches to language have SPURRED ON LOTS OF PROGRESS (cf Chomsky!!!). A

In the true nature of the interdisciplinarity of the field I wish to WORK AT BOTH AIMS IN PARALLEL and show how they complement each other.

Importance on engineering side:

- Zero-shot performance of ChatGPT comparable with BERT [Zhong et al., 2023]
- I also experienced poor (?) performance with own experiments
- but fine-tuning lead to large improvements

1 Just to name a few examples: formal languages, artificial neural networks and SOMETHING ELSE!!!!!!

- UMR annotates aspect, and this can be used to extract habitual events or states, which are typical knowledge forms

The fact that fine-tuning can lead the model to

Therefore the purpose of this study is two-fold: firstly to further explore the phenomenology of aspect using methods from computational linguistics such as neural embeddings, and secondly to look at how current state-of-the-art approaches deal with this phenomenon and investigate how this could be improved.

I wish to challenge Chomsky's assertion that large language models are "not a contribution to science"²

Main contributions

- First in-depth study using computational approaches to study the phenomenology of aspect
- First probing of neural models applied to the task
- SOMETHING ELSE

Structure of this thesis

² <https://www.youtube.com/watch?v=axuGfh4UR9Q> WHAT TIME DOES HE SAY THIS???

2 Aspectology: an introduction

Aspect is one of the most studied areas of linguistics [Sasse, 2002], particularly Slavic linguistics (for reasons I will discuss later), and hence a thorough theoretical discussion of the linguistic phenomenon which does full justice to the work on the field is not possible within the constraints of this thesis. Nevertheless, in order to make a productive contribution to the area I will touch on some of the main findings in the area of aspectology.

2.1 A (short) phenomenology of aspect

The Concise Oxford Dictionary of Linguistics [Matthews, 2014] defines aspect thus:

General term, originally of specialists in Slavic languages, for verbal categories that distinguish the status of events, etc. in relation to specific periods of time, as opposed to their simple location in the present, past, or future.

As noted, one distinction which must be made right away is that between *aspect*, and another temporal phenomenon *tense*. Many works recall Bernard Comrie's differentiation in his book *Aspect* [Comrie, 1976] between the deictic nature of *tense* and the focus on the "internal temporal constituency of a situation" of *aspect*. In other words: "tense¹ relates the time of the situation referred to to some other time, usually to the moment of speaking", and thus by relating the time of the situation to the time of the utterance it is deictic [Comrie, 1976]. Aspect on the other hand gives a *situation-internal* description of the events in that situation, such how they relate to each other temporally or how an individual event is temporally characterised. This brings us to another important distinction: that between lexical and grammatical aspect.

1 In many of the world's languages this is grammaticalised as past, present and future; in others such as English by some accounts [Jespersen, 1933] it is a binary distinction such as past and non-past, whereas Some languages such as Greenlandic (Kalaallisut) which some linguists have even argued to be tenseless Bittner [2005]. See also the contentious debate about Hopi time Whorf [2012], Malotki [1983].

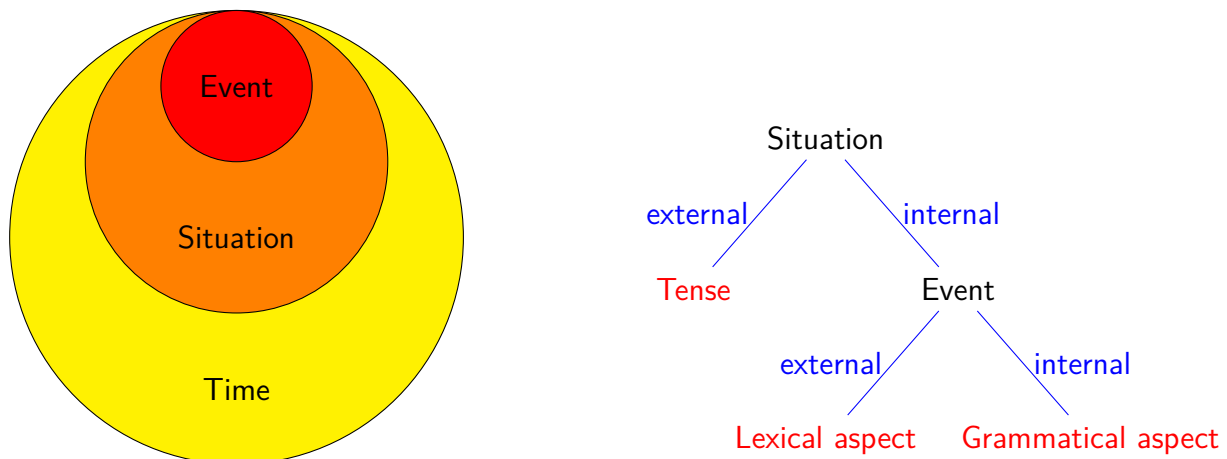


Figure 1: A set-theoretical representation of the relationship between time, situation and event (left) along with the categorisation of the three tempora phenomena discussed (right).

2.1.1 Lexical vs grammatical aspect

The polysemy of the term “aspect” in the linguistic community is unfortunate. In an area of language which seems to have surprisingly far-reaching interactions with other parts of linguistic systems these ambiguities are particularly unhelpful.² As already mentioned, aspect³ is a term often used to refer to both lexical and grammatical aspect. Lexical aspect (also referred to as *Aktionsart*, *situation aspect* or *inner aspect*) is the inherent property of a verb or verb phrase which “characterizes the temporal profile of event descriptions” [van Hout, 2016]. For example, to “crack open an egg” is inherently a short event describing the change of one state (the egg being whole) to another (the egg being cracked). Grammatical aspect (or *viewpoint aspect*, *outer aspect*), on the other hand, describes the “internal temporal constituency” [Comrie, 1976] of an event, such as its habituality or ongoing nature and can be seen more as an external lens imposed on a verbal phrase. In English, one example of this lens is the progressive, which is formed by the verb *be* + gerund, as in the phrase “Sue was running”. Figure 1 shows a taxonomy of tense, and grammatical and lexical aspect.

The concrete linguistic realisation of these categories is very often unclear or non-existent and exhibits a relatively wide variety of encodings throughout the world’s languages [Dahl, 1985]. It therefore proves tricky to uphold this distinction in empirical studies “in the real world”. Nevertheless

² Among other things, aspect is also intertwined with case, mood and voice [Franks, 2005, Kiparsky, 2004].

³ Boogaart uses the term *aspectuality* to remove the aforementioned ambiguity [Boogaart, 2004]. However, I will stick with the term *aspect* due to its prevalence in the literature, further specifying where necessary.

the theoretical differentiation of these two interrelated subdomains is an important one to make, and each describes a variety of concepts which will be useful later.

2.1.2 Lexical aspect

Telicity

A fundamental distinction of lexical aspect is that of telicity (from Ancient Greek *télos* meaning “end”). Telicity describes whether an event has an inherent goal or end-point after which the event can be regarded as completed: for example "go climbing" would be atelic whereas "climb the mountain" is telic (since it involves the agent reaching the summit of a mountain). A classic test for telicity⁴ is whether the verb phrase admits a completing adverb such as "in an hour" and does not admit a durative adverb such as "for an hour" [Krifka, 1998].

Dahl misleadingly defines telicity as "involv[ing] the presence of a boundary or the attainment of a specific result-state" [Östen Dahl, 2015], which, however can lead to confusion with the term *(un)boundedness*.

Boundedness

(Un)boundedness refers to the existence (or lack of) a *temporal* boundary marking the end of an action. Boundedness must be distinguished from telicity⁵ in order to avoid the so-called 'Imperfective Paradox' highlighted by Linguistics et al. [2005], here verbalised by Zucchi [2020]:

How is it possible that a statement of the form *x was F-ing* is true and yet there is no time at which *x was F-ed* is true?

More concretely, it asks the question why (1) entails (2) but (3) doesn't entail (4) in the examples below:

⁴ Though Xiao and Mcenery [2006] note that this test is flawed and propose an alternative test scheme.

⁵ Friedrich et al. [2023] seem to mistakenly conflate the two, stating that "[t]elicity is also sometimes referred to as boundedness (e.g., by Loáiciga and Grisot, 2016)" referring to the 2016 paper Loáiciga and Grisot [2016], which, however, clearly distinguishes between the two notions.

- (1) The man was running.
- (2) The man ran.
- (3) The man was building a house.
- (4) The man built a house.

Depraetere [1995] shows that this apparent "paradox" can be resolved by distinguishing between these two concepts of telicity and boundedness,⁶ and this further serves to show the dangers of misuse of terminology. Table 2.1.2 visualises the relationship between the reference time (temporal boundaries) and the run-time (the inherent "time schema" of the verb phrase). This allows for a clearer definition of boundedness, namely whether the right-hand side of the run-time boundary lies within the reference time or not (IS THIS TRUE?).

Stativity

Another parameter of lexical aspect I would like to discuss is stativity, which describes a state of being such as "know", "love" or "be", rather than an action [Binnick, 1991]. A classic test for stativity in English is non-admittance of the progressive or the imperative [McIntosh, 1975].⁷ Consider the following examples:

- (5) She resembles her grandmother.
- (6) * She is resembling her grandmother.

Durativity

Durativity denotes whether an event takes time (i.e. has duration) or happens in an instant, and this can be checked in English by the compatibility of durative adverbs such as *for an hour* [Wilhelm, 2007]. For example:

-
- 6 By definition of telicity as whether an even has an inherent end-point, and boundedness as whether it has a temporal boundary (separate from its intended end-point) we can distinguish between whether an event has reached its termination or whether it was ended before reaching this end-point. Therefore while it the case that the both (1) and (2) are bounded, only (3) and (4) contain telic events, and it seems to be the case that the progressive's focus on temporal boundary nullifies the end-point inherent in the verb phrase "build a house". This serves as an interesting example of the interaction between grammatical and lexical aspect.
 - 7 Interestingly, however, Granath and Wherrity [2013] find that, assuming a functional-semantic definition of stativity, the usage of stative verbs with the progressive is much higher in spoken language, than in written language.

Sentence	Representation	Telicity	Boundedness
The man was running.	<p>The diagram shows a horizontal timeline with an arrow at the end. A bracket labeled 'Reference time' spans the entire duration shown. Another bracket labeled 'Run-time' is positioned below the timeline, starting at the beginning and extending for a significant portion of the reference time, but not reaching the end, indicating it is unbounded.</p>	atelic	unbounded
The man ran.	<p>The diagram shows a horizontal timeline with an arrow at the end. A bracket labeled 'Reference time' spans the entire duration shown. Another bracket labeled 'Run-time' is positioned below the timeline, starting at the beginning and extending to the end of the reference time, indicating it is bounded.</p>	atelic	bounded
The man was building a house.	<p>The diagram shows a horizontal timeline with an arrow at the end. A bracket labeled 'Reference time' spans the entire duration shown. Another bracket labeled 'Run-time' is positioned below the timeline, starting at the beginning and extending for a significant portion of the reference time, but not reaching the end, indicating it is unbounded. A vertical line labeled 'Goal' is positioned at the end of the run-time bracket, indicating the goal of the event.</p>	telic	unbounded
The man built a house.	<p>The diagram shows a horizontal timeline with an arrow at the end. A bracket labeled 'Reference time' spans the entire duration shown. Another bracket labeled 'Run-time' is positioned below the timeline, starting at the beginning and extending to the end of the reference time, indicating it is bounded. A vertical line labeled 'Goal' is positioned at the end of the run-time bracket, indicating the goal of the event.</p>	telic	bounded

Table 1: Representation of (1),(2),(3) and (4) with regards to the reference time referred to by the utterance and inherent run-time of the event. Concept adapted from van Hout [2016].

- (7) Andrea was painting a picture for an hour.
(8) * Andrea was reaching the summit for an hour.

2.1.2.1 Some attempts at event classification

Vendler [1957]

It was the seminal work *Verbs and Times* [Vendler, 1957] of philosopher Zeno Vendler which initiated the discussion on inner aspect in the linguistic tradition.⁸ Vendler begins his discussion with the following premise:

Indeed, as I intend to show, if we focus our attention primarily upon the time schemata presupposed by various verbs, we are able to throw light on some of the obscurities which still remain in these matters. [...] There are a few such schemata of very wide application. Once they have been discovered in some typical examples, they may be used as models of comparison in exploring and clarifying the behaviour of any verb whatever.⁹

That is to say, the "time schema" of any verb can be described through comparison with a set of prototypical classes (see 2.1.2.1), which can be easily identified. In order to arrive at these prototypical "schemata" he uses an analytical method consisting of classifying verbs according to their behaviour regarding certain elements of English grammar, as was also used to outline the aspectual parameters above.¹⁰ For example, the first distinction he makes is between English verbs that permit the progressive and those that don't. This signals the first class of events known as 'states'. The article then goes on to outline the other three Vendlerian classes *activity*, *accomplishment* and *achievement*, and their character is summarised thus:

- **State** - non-dynamic, static and durative situation

⁸ A similar classification was also developed independently by Anthony Kenny in *Action, Emotion and Will* [Kenny, 1963], however combining Achievement and Accomplishment into one single class [Mourelatos, 1978]. Hence it is sometimes referred to as the Vendler-Kenny scheme of verb-types.

⁹ Vendler [1957]

¹⁰ The issue of Anglocentrism is one which has plagued many a linguistic theory throughout the years, with work from Chomsky's generative grammar [Levisen, 2019] to Abstract Meaning Representation [Damonte and Cohen, 2018] being criticised for their too heavy focus on English.

	Static	Durative	Telic
State	+	+	-
Activity	-	+	-
Accomplishment	-	+	+
Achievement	-	-	+

Table 2: Classification of Vendlerian event types by binary aspectual parameters [Smith, 1991]

State	Activity	Accomplishment	Achievement
know	running	paint a picture	reach the summit
understand	pushing	build a house	spot the plane
love	smoking	deliver a sermon	recognise

Table 3: Some of the example verb phrases given in Vendler [1957] for the classification of events.

- **Activity** - open-ended, dynamic and durative processes without an end-point
- **Accomplishment** - dynamic and durative processes with a natural end-point
- **Achievement** - instantaneous or near-instantaneous events (such as semelfactives)

Or to use the parameters of lexical aspect introduced above:

Vendler states in his introduction that verbs "presuppose" certain time schemata and hence assigns a category to each verb (as in 2.1.2.1). However it must also be stated that the true profile of a verb phrase such as those listed in 2.1.2.1 depends heavily on the context. Hence a typical semelfactive such as "sneeze" could also be used reinterpreted as a process when combined with a progressive auxiliary as in "Harry was sneezing" [Moens and Steedman, 1988]. Thus the verbs (or verb phrases) mentioned by Vendler as belonging to a certain category are ones that typically lend themselves to one class or another.

Moens and Steedman [1988]

2.1.3 Grammatical aspect

Grammatical (or *viewpoint*) aspect provides a lens through which a particular event is viewed and it "is typically expressed by overt grammatical morphemes (hence the label grammatical aspect)" [Patard et al., 2019]. Comrie [1976] provides a hierarchical classification of aspectual oppositions

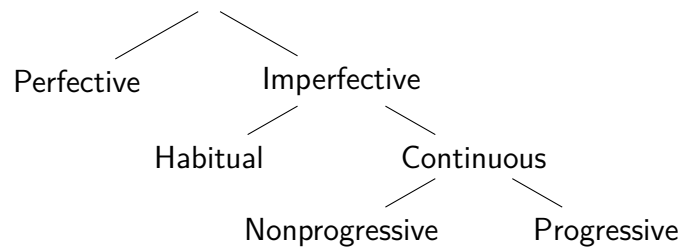


Figure 2: Comrie's classification of aspectual oppositions. Reproduced from Comrie [1976].

shown in 2, and in the following section I will look at some of the main oppositions described in this classification.

Perfective vs. Imperfective

The main aspectual distinction made by Slavic languages and, as can be seen in figure 2, one of the main distinctions made in theoretical aspectology generally is that between perfective and imperfective. In order to avoid confusion with the concrete realisation of these theoretical categories in the aforementioned Slavic group I will follow the nomenclature used by DAHLL and refer to the theoretical

Progressive

Iterative

2.2 Aspect in Slavic languages

The Slavic language group has a special place in the study of aspectology due to its overt encoding of aspectual phenomena, otherwise rather uncommon [Tomelleri, 2010].¹¹

WE CAN SEE THE DIACHRONICAL DEVELOPMENT FROM LEXICAL ASPECT TO GRAMMATICAL ASPECT

¹¹ Tomelleri [2010] also notes that Georgian and Ossetian exhibit some similar properties with preverbs (mostly of spatial origin) used to denote aspectual meaning.

2.2.1 Slavic Perfective-Imperfective distinction

Show that imf and perf different to telicity/boundedness

2.3 Formal models of aspect (or put this in related work?)

I would like to outline some of formal logical approaches that have been made to this area

3 Related Work

3.1 Previous computational approaches to modeling aspect

Aspect has received comparatively very little attention from the computational linguistics community[Friedrich et al., 2023], especially from the Natural Language Processing (NLP) part of the field.¹ However there have been some works in recent years looking at this area:

3.1.1 Situation entities!!

title

¹ As opposed to those with more emphasis on the *linguistics* part of computational linguistics.

4 Methods

This is the description of the methods.

4.1 Which aspect types do I use?

One inherent drawback of computational methods is exactly their one unifying characteristic: their computability. Sadly, the requirement for computability means, in many cases, sacrificing the nuance that comes with more qualitative approaches. In this concrete case this means settling for a single aspect classification system. A consequence of the glut of literature in the field is a glut of classification systems to go with it, each with their own idiosyncrasies and each having their own advantages and drawbacks. I will attempt to discuss what I see as the main candidates before deciding on one system and explaining this decision - NO - this was done in aspectology already!!!!SEE ABOVE

reason is of course the contribution to the linguistic community: computational approaches to language have SPURRED ON LOTS OF PROGRESS (cf Chomsky!!!).

The framework I decided on was Uniform Meaning Representation (UMR) [Van Gysel et al., 2021].

4.1.1 UMR

UMR [Van Gysel et al., 2021] was introduced in order to expand and further generalise the attempt to design an abstract semantic representation, as was most successfully pioneered by Banarescu et al. [2013] with Abstract Meaning Representation (AMR). In contrast to AMR, UMR aims to be a typologically-informed abstraction away from English structures, making it more suitable for other similar languages, or, in their own words, making it "a practical and cross-linguistically valid meaning representation designed to meet the needs of a wide range of NLP applications" [Van Gysel et al., 2021].

WHY IS UMR IMPORTANT???? Get Apaho English comparison from Bonn et al. [2023]

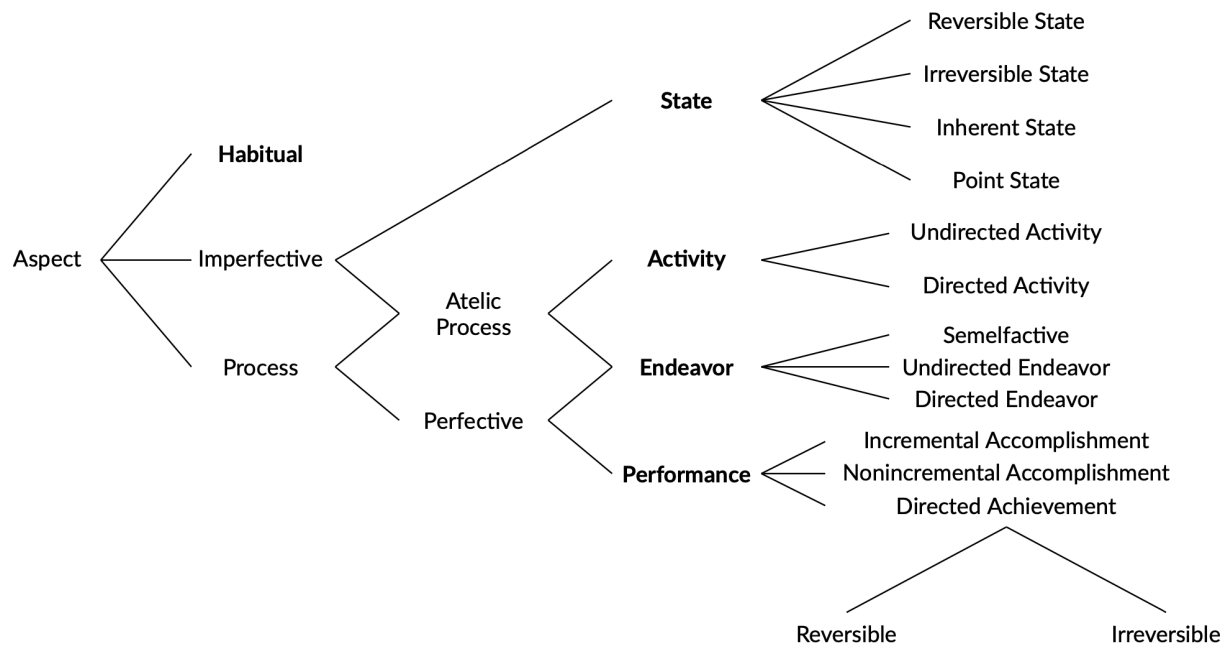


Figure 3: UMR aspect classification tree [Jens Van Gysel and Xue, 2022]

Aspect in UMR

UMR describes the following 5 coarse-grained aspect classes: (descriptions taken from Van Gysel et al. [2021]), also depicted in figure 3:

- **state** - an unspecified type of state
- **habitual** - an event that occurs regularly in the past or present, including generic statements
- **activity** - an event that has not necessarily ended and may be ongoing at Document Creation Time (DCT)
- **endeavour** - a process that ends without reaching completion (i.e., termination)
- **performance** - a process that reaches a completed result state

How these relate to other aspectual classes can be shown in table 4.

Interesting to note is that these classes conflate the distinction made earlier between lexical and grammatical aspect, most clearly in the class "habitual", which is a paradigmatic example of an outer aspect, rather than one inherent in the verb event itself. However, as can be seen in figure

Vendler [1957]	Moens and Steedman [1988]	Egg [2005]	Egg, Prepens, and Roberts [2019]		Van Gysel et al. [2021]
state	state (habitual state)	stative predicate (CHECK THIS)	stative -		state habitual
activity	process	process predicate	dynamic	unbounded	activity
accomplishment	culminated process	intergressive predicate		bounded	extended/no change endeavour
		change predicate			extended/change performance
achievement	point	intergressive predicate			punctual/no change endeavour
	culmination	change predicate			punctual/change performance

Table 4: Comparison of aspectual classes. Adapted and extended from Egg, Prepens, and Roberts [2019].

3, while classes which would usually be seen as grammatical aspect are to be found nearer the top of the tree (on the left), the leaves further down the tree are more examples of *Aktionsarten*. That Van Gysel et al. [2021] make no mention of these different types of aspect is not necessarily surprising, given one of the main design goals of UMR being scalability, itself entailing learnability for annotators. As we have seen (HAVE WE?), the theoretical distinction of inner and outer aspect is "very difficult to apply in practice" [Dahl, 1985], hence a clear separation would often be difficult - and indeed not very fruitful - for annotators. This conflation of two phenomena, however, can also be an advantage, since it shows the relationship between classes of each type (i.e. that *irreversible states* are imperfective and *directed achievements* perfective etc.), subsuming them all into *one* semantic parameter space concerning aspect. A further advantage of UMR is the flexibility of annotation levels that can be seen in figure 3. This allows for annotation both at a fine-grained level (using the classes at the bottom of the tree), but also on a more coarse-grained level if instances are unclear or annotators unsure.

4.1.2 How does the Slavic Perfective-Imperfective relate to UMR classes

Habitual = imperfective State = imperfective (general imp?) Activity = imperfective (see motion verbs) endeavour =

performance = perfective

4.2 Use of approximative LMs as sources of linguistic knowledge??

4.3 Where do LLMs fail?

Dataset creation

In order to create the dataset for training the model, later to be used for the neural analysis, it is

Since UMR aspect is an AMR/UMR node parameter, rather than a clausal one (see for example Smith's situation entity types [ADD LINK HERE](#)), it is necessary to first parse any sentence into an AMR graph. However, since no parsers are available for Russian, I had to resort to another option. One possibility would be to develop a system, possibly using tools already available for Russian such as syntax parsers [ADD EXAMPLE HERE](#), however since the focus is on one small part of the AMR/UMR graph - nodes describing an event - this could be an unnecessary distraction. In theory at least, AMR graphs should have a strong correlation with the syntax tree representation of the sentence, and since most events are verbal phrases, this fact can be utilised to simplify the problem. I therefore opted to solve the problem by using a dependency parser (OR SYNTAX PARSER???)

5 Experiments

5.1 Dataset creation

In order to carry out further computational analysis of the topic it is helpful to have an annotated dataset. Since no datasets exist beyond for certain aspectual parameters such as telicity [Friedrich and Gateva, 2017] or stativity [Friedrich and Palmer, 2014], I had to create my own [Friedrich et al., 2023]. The traditional route of hand-annotating including has been a standard paradigm for many years however is highly time-intensive. Crowdsourcing is a more time- and cost-effective way of creating data, however has been shown to be of variable quality [Li, 2024], especially when the task requires more expert knowledge. Furthermore there are a range of biases which can affect data quality [Beck, 2023]. The huge success of Large Language Models in recent years has prompted some to look at utilising them for dataset annotation either on their own [He et al., 2023, Yu et al., 2023, Gilardi et al., 2023] or combined with human annotators [Goel et al., 2023].

5.1.1 Dataset annotation with LLMs

I decided to annotate the

Choice of LM model

There has been an explosion of LMs in recent years

While many previous studies looking at LLM capabilities choose to look at ChatGPT models due to their notoriety in recent years and general good performance. However, as Törnberg [2024] notes, these models come with several issues: it is unknown what the training data for the model is, leading to problems with transparency, and ChatGPT models have been shown to evolve over time, meaning reproducibility is hindered. For these reasons, and also in order to host the model locally for better control (rather than using an API), I chose to use a different model. Meta's Llama 2 model seemed to offer a good balance between the apparent current trade-off between

Class	Support
Performance	124
State	55
Activity	36
Endeavour	17
Habitual	2
Total	234

Table 5: Aspect classes of annotated verbal events in the UMR dataset.

performance and scientific good practice, performing comparatively well in previous studies [Yuan et al., 2023] + OTHER EXAMPLES

Data creation

As the only currently existing data containing UMR aspect classes, I used the example UMR dataset¹ provided by the creators of UMR. It contains HOW MANY sentences in 6 languages (Arapaho, Chinese, English, Kukama, Navajo and Sanapana) annotated according to the UMR schema. The texts come from WHERE

In order to extract the verbs in the sentences I used Stanford NLP’s Stanza POS tagger CITEEE, utilising the alignment given in the training data to find the corresponding node in the UMR graph and its aspect annotation. Table 5.1.1 shows the class distribution of verbal events extracted from the dataset.

Here it is clear that the *habitual* class is underrepresented, corroborating the results of Dahl [1985]’s study, and hence I manually added some datapoints to improve the balance. The added sentences were either found in existing online datasets and simply labelled with a UMR aspect class by hand or they were both manually composed and then labelled.

Törnberg [2024] points out that, since it is unknown exactly which training data was used to train many LMs, one should exercise caution when evaluating their performance on a test set, since the model may have seen the test data before during pre-training. In this case, while it is impossible to rule out that the UMR dataset was used for Llama 2 pretraining, it is highly unlikely that it has seen the data in this form (i.e. as a sentence, a verb from this sentence and a UMR aspect label for this verb), and since the model was pre-trained with a next-token prediction task, it is

1 ENTER LINK FOR WHERE TO FIND IT!!!!!!

highly unlikely that it would have memorised the labels for the data it is being tested on in my experiments. Nevertheless this cannot be ruled out and must be kept in mind when analysing the results.

One interesting feature of the UMR :aspect parameter, is that it is used not only with verbs in the source sentence but also with nouns and adjectives:

(9) 200 dead , 1,500 feared missing in Philippines landslide .

Since Russian NLP tools are scarce, it would be difficult to perform event extraction, so, in order to keep the results comparable between both languages, I chose to just focus on verbal events. In my analysis the :aspect parameter occurred with a verb in the source sentence 74.8% of the time, meaning most of the data from the UMR dataset could be used.

Prompt engineering

The model was given the following instruction for each datapoint:

The annotation distinguishes five base level aspectual values—State, Habitual, Activity, Endeavor, and Performance. The State value corresponds to stative events: no change occurs during the event. It also includes predicate nominals (be a doctor), predicate locations (be in the forest), and thetic (presentational) possession (have a cat). The Habitual value is annotated on events that occur regularly in the past or present. The Activity value indicates an event has not necessarily ended and may be ongoing at Document Creation Time (DCT). Endeavor is used for processes that end without reaching completion (i.e., termination), whereas Performance is used for processes that reach a completed result state.

followed by the following question:

Which class does "{verb}" belong to in this sentence: state, habitual, activity, endeavor, or performance?"

Due to hardware constraints, I had to use a technique to improve the efficiency of the model training process since the smallest llama model is 7 billion parameters. In this case I chose to use PEFT (Parameter efficient fine-tuning) (ADD CITATION), which leverages the insight that LM

fine-tuning usually only updates parameters at the end of the model. This made it possible to fine-tune the model without resorting to the huge amount of computing power usually necessary for fine-tuning LLMs.

I experimented both with the standard Llama-2-7b and Llama-2-7b-chat [Touvron et al., 2023], however the latter ended up having difficulties responding in a formulaic way and rather added unnecessary or unrelated information, which is to be expected since it has been fine-tuned to fare well in a dialogue environment. This made it often difficult to extract the model's predicted label to use for evaluation.

Best practices for using LLMs as annotators

Törnberg [2024] formulated a set of best practices when using LLMs as text annotators, which I aim to follow.

Systematic coding procedure

Develop a prompt codebook

With annotation guide for human annotators

Validate the model

Prompt engineering

Return as JSON

Examine model stochasticity

5.1.2 Language model fine-tuning

5.1.2.1 Ablation (aspectualisation and contextualisation)

5.1.2.2 Combined Rus/Eng model

5.1.3 Probing

5.1.4 Aspect latent space

5.1.4.1 Verb of motion clustering

5.1.4.2 Prefix clustering

6 Discussion

Discussing the results.

7 Conclusion

And finally a conclusion.

A Appendix

A.1 Experiments with ChatGPT

Over the last few years, Large Language Models (LLMs) such as GPT-3 [Brown et al., 2020] have achieved great amounts of success over a range of tasks in the NLP domain.

However it has been shown that models such as ChatGPT struggle with mathematical and logical reasoning.

A.1.1 Temporal reasoning

Bibliography

B. Comrie. *Aspect: An Introduction to the Study of Verbal Aspect and Related Problems*. Camb. Textbk. Ling. Cambridge University Press, 1976. ISBN 9780521211093. URL <https://books.google.it/books?id=1RVcXZjV0j0C>.

Jin Zhao Jens Van Gysel, Meagan Vigus and Nianwen Xue. Developing a uniform meaning representation for natural language processing. Tutorial, 2022. URL https://lrec2022.lrec-conf.org/media/filer_public/94/45/9445a9b9-4cfb-4898-a5cc-a89143ab2a2a/umr_lrec_tutorial_slides_6.pdf.

Angeliek van Hout. Lexical and Grammatical Aspect. In *The Oxford Handbook of Developmental Linguistics*. Oxford University Press, 07 2016. ISBN 9780199601264. doi: 10.1093/oxfordhb/9780199601264.013.25. URL <https://doi.org/10.1093/oxfordhb/9780199601264.013.25>.

Carlota Smith. The parameter of aspect. 1991. URL <https://api.semanticscholar.org/CorpusID:61127772>.

Zeno Vendler. Verbs and times. *The Philosophical Review*, 66(2):143–160, 1957. ISSN 00318108, 15581470. URL <http://www.jstor.org/stable/2182371>.

Markus Egg, Helena Prepens, and Will Roberts. Annotation and automatic classification of aspectual categories. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3335–3341. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1323. URL <https://doi.org/10.18653/v1/p19-1323>.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert, 2023.

Hans-Jürgen Sasse. Recent activity in the theory of aspect: Accomplishments, achievements, or just non-progressive state? *Linguistic Typology*, 6, 2002. URL <https://api.semanticscholar.org/CorpusID:121510831>.

- P.H. Matthews. *The Concise Oxford Dictionary of Linguistics*. Oxford Paperback Reference. OUP Oxford, 2014. ISBN 9780199675128. URL <https://books.google.it/books?id=1mg3BQAAQBAJ>.
- Otto Jespersen. *Essentials of English grammar*. Routledge, 1933.
- Maria Bittner. Future Discourse in a Tenseless Language. *Journal of Semantics*, 22(4):339–387, 08 2005. ISSN 0167-5133. doi: 10.1093/jos/ffh029. URL <https://doi.org/10.1093/jos/ffh029>.
- Benjamin Lee Whorf. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. The MIT Press, 2012. ISBN 9780262517751. URL <http://www.jstor.org/stable/j.ctt5hhbx2>.
- Ekkehart Malotki. *Hopi Time*. De Gruyter Mouton, Berlin, New York, 1983. ISBN 9783110822816. doi: doi:10.1515/9783110822816. URL <https://doi.org/10.1515/9783110822816>.
- Steven Franks. The slavic languages. *Handbook of Comparative Syntax*, pages 373–419, 2005.
- Paul Kiparsky. Partitive case and aspect. 2004. URL <https://api.semanticscholar.org/CorpusID:118440489>.
- Ronny Boogaart. *Aspect and Aktionsart*, pages 1165–1180. De Gruyter Mouton, Berlin ■ New York, 2004. ISBN 9783110194272. doi: doi:10.1515/9783110172782.2.14.1165. URL <https://doi.org/10.1515/9783110172782.2.14.1165>.
- Östen Dahl. Tense and aspect systems. 1985. URL <https://api.semanticscholar.org/CorpusID:18132338>.
- Richard Xiao and Tony Mcenery. Can completive and durative adverbials function as tests for telicity? evidence from english and chinese. *Corpus Linguistics and Linguistic Theory*, 2(1):1–21, 2006. doi: doi:10.1515/CLLT.2006.001. URL <https://doi.org/10.1515/CLLT.2006.001>.
- Manfred Krifka. The origins of telicity. 1998. URL <https://api.semanticscholar.org/CorpusID:55535400>.
- Östen Dahl. Tense, aspect, mood and evidentiality, linguistics of. In James D. Wright, editor, *International Encyclopedia of the Social and Behavioral Sciences (Second Edition)*, pages 210–213. Elsevier, Oxford, second edition edition, 2015. ISBN 978-0-08-097087-5. doi:

<https://doi.org/10.1016/B978-0-08-097086-8.52025-X>. URL <https://www.sciencedirect.com/science/article/pii/B978008097086852025X>.

Annemarie Friedrich, Nianwen Xue, and Alexis Palmer. A kind introduction to lexical and grammatical aspect, with a survey of computational approaches. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 599–622, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.44. URL <https://aclanthology.org/2023.eacl-main.44>.

Sharid Loáiciga and Cristina Grisot. Predicting and using a pragmatic component of lexical aspect of simple past verbal tenses for improving English-to-French machine translation. *Linguistic Issues in Language Technology*, 13, 2016. URL <https://aclanthology.org/2016.lilt-13.3>.

Pacific Linguistics, Segura Carlota, and Evert Jan. Dowty, david. 1979. word meaning and montague grammar: The semantics of verbs and times in generative semantics and in montague's ptq. dordrecht: Reidel. 2005. URL <https://api.semanticscholar.org/CorpusID:57525242>.

Sandro Zucchi. *Progressive: The Imperfective Paradox*, pages 1–32. 11 2020. ISBN 9781118788318. doi: 10.1002/9781118788516.sem138.

Ilse Depraetere. On the necessity of distinguishing between (un)boundedness and (a)telicity. *Linguistics and Philosophy*, 18(1):1–19, 1995. ISSN 01650157, 15730549. URL <http://www.jstor.org/stable/25001576>.

R.I. Binnick. *Time and the Verb: A Guide to Tense and Aspect*. Filologia y linguistica. Oxford University Press, 1991. ISBN 9780195062069. URL <https://books.google.it/books?id=LDXnCwAAQBAJ>.

C. McIntosh. The semantics of stativity. *Journal of Literary Semantics*, 4(Jahresband):35–42, 1975. doi: doi:10.1515/jlse.1975.4.1.35. URL <https://doi.org/10.1515/jlse.1975.4.1.35>.

Solveig Granath and Michael Wherrity. I'm loving you – and knowing it too: Aspect and so-called stative verbs. *Rhesis. International Journal of Linguistics, Philology and Literature*, 4(1):6–22, Dec. 2013. doi: 10.13125/rhesis/5575. URL <https://ojs.unica.it/index.php/rhesis/article/view/5575>.

Andrea Wilhelm. *Telicity and durativity: a study of aspect in D ne Sul/ine (Chipewyan) and German*. Studies in linguistics. Routledge, New York, 2007. ISBN 9780415976459. Includes bibliographical references (p. 315-325) and indexes.

Anthony Kenny. *Action, Emotion and Will*. Humanities Press, Ny, 1963.

Alexander P. D. Mourelatos. Events, processes, and states. *Linguistics and Philosophy*, 2(3): 415–434, 1978. ISSN 01650157, 15730549. URL <http://www.jstor.org/stable/25000995>.

Carsten Levisen. Biases we live by: Anglocentrism in linguistics and cognitive sciences. *Language Sciences*, 76:101173, 2019. ISSN 0388-0001. doi: <https://doi.org/10.1016/j.langsci.2018.05.010>. URL <https://www.sciencedirect.com/science/article/pii/S0388000117303558>. Biases in Linguistics.

Marco Damonte and Shay B. Cohen. Cross-lingual abstract meaning representation parsing, 2018.

Marc Moens and Mark Steedman. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28, 1988. URL <https://aclanthology.org/J88-2003>.

Adeline Patard, Rea Peltola, and Emmanuelle Roussel. *Chapter 1 Introduction: Cross-Linguistic Perspectives on the Semantics of Grammatical Aspect*, pages 1 – 9. Brill, Leiden, The Netherlands, 2019. ISBN 9789004401006. doi: 10.1163/9789004401006_002. URL <https://brill.com/view/book/9789004401006/BP000001.xml>.

Vittorio S. Tomelleri. Slavic-style aspect in the caucasus. *Suvremena lingvistika*, 36(69), 2010. URL <http://suvlin.ffzg.hr/index.php/en/browse-archive/39-vol-36-no-69-07-2010/432-slavic-style-aspect-in-the-caucasus,pages=>.

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O’Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Haji , James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. Designing a uniform meaning representation for natural language processing. *KI - K nstliche Intelligenz*, 35(3–4):343–360, April 2021. ISSN 1610-1987. doi: 10.1007/s13218-021-00722-w. URL <http://dx.doi.org/10.1007/s13218-021-00722-w>.

L. Banarescu, Claire Bonial, Shu Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, Martha Palmer, and N. Schneider. Abstract meaning representation for sembanking. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, 01 2013.

Julia Bonn, Skatje Myers, Jens E. L. Van Gysel, Lukas Denk, Meagan Vigus, Jin Zhao, Andrew Cowell, William Croft, Jan Hajič, James H. Martin, Alexis Palmer, Martha Palmer, James Pustejovsky, Zdenka Urešová, Rosa Vallejos, and Nianwen Xue. Mapping AMR to UMR: Resources for adapting existing corpora for cross-lingual compatibility. In Daniel Dakota, Kilian Evang, Sandra Kübler, and Lori Levin, editors, *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)*, pages 74–95, Washington, D.C., March 2023. Association for Computational Linguistics. URL <https://aclanthology.org/2023.tlt-1.8>.

M. Egg. *Flexible Semantics for Reinterpretation Phenomena*. CSLI Lecture Notes (CSLI- CHUP) Series. CSLI Publications, 2005. ISBN 9781575865027. URL <https://books.google.it/books?id=jGVsAAAAIAAJ>.

Annemarie Friedrich and Damyana Gateva. Classification of telicity using cross-linguistic annotation projection. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2559–2565, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1271. URL <https://aclanthology.org/D17-1271>.

Annemarie Friedrich and Alexis Palmer. Automatic prediction of aspectual class of verbs in context. In *Annual Meeting of the Association for Computational Linguistics*, 2014. URL <https://api.semanticscholar.org/CorpusID:1832412>.

Jiyi Li. A comparative study on annotation quality of crowdsourcing and llm via label aggregation, 2024.

Jacob Beck. Quality aspects of annotated data: A research synthesis. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 17(3–4):331–353, November 2023. ISSN 1863-8163. doi: 10.1007/s11943-023-00332-y. URL <http://dx.doi.org/10.1007/s11943-023-00332-y>.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. Annollm: Making large language models to be better crowdsourced annotators, 2023.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies. 05 2023.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), July 2023. ISSN 1091-6490. doi: 10.1073/pnas.2305016120. URL <http://dx.doi.org/10.1073/pnas.2305016120>.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. Llms accelerate annotation for medical information extraction, 2023.

Petter Törnberg. Best practices for text annotation with large language models, 2024.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. Back to the future: Towards explainable temporal reasoning with large language models, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

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. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20*, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.