Identifying and Quantifying Aspectual Ambiguity with LMs

Reversing the NLP Pipeline

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. Katja Markert

Abstract

Abstract in English

Zusammenfassung

Zusammenfasssung auf Deutsch

Contents

List of Figures

List of Tables

1 Introduction

Motivation

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

It also serves/aims??? to be a case study in the use of Language Models in linguistic research.

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]

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

- I also experienced poor (?) performance with own experiments
- but fine-tuning lead to large improvements
- 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 (REALLY?)
- First probing of neural models applied to the task
- First attempt at aspectual ambiguity recognition

Structure of this thesis

Acknowledgements

Herweg Valentin Dad Mum JN Housemates

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

2 Aspectology: an introduction

Many works on the area begin with the assertion that aspect is one of the most studied areas in 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 give an introduction to the fundamental elements of the following study and 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:

> [Aspect is a g]eneral 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 here, one helpful 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: "tense[1] 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 as how they relate to each other temporally or how an individual event is temporally characterised. This introduces 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) some linguists have even argued to be tenseless Bittner [2005]. See also the contentious debate about Hopi time Whorf [2012], Malotki [1983].

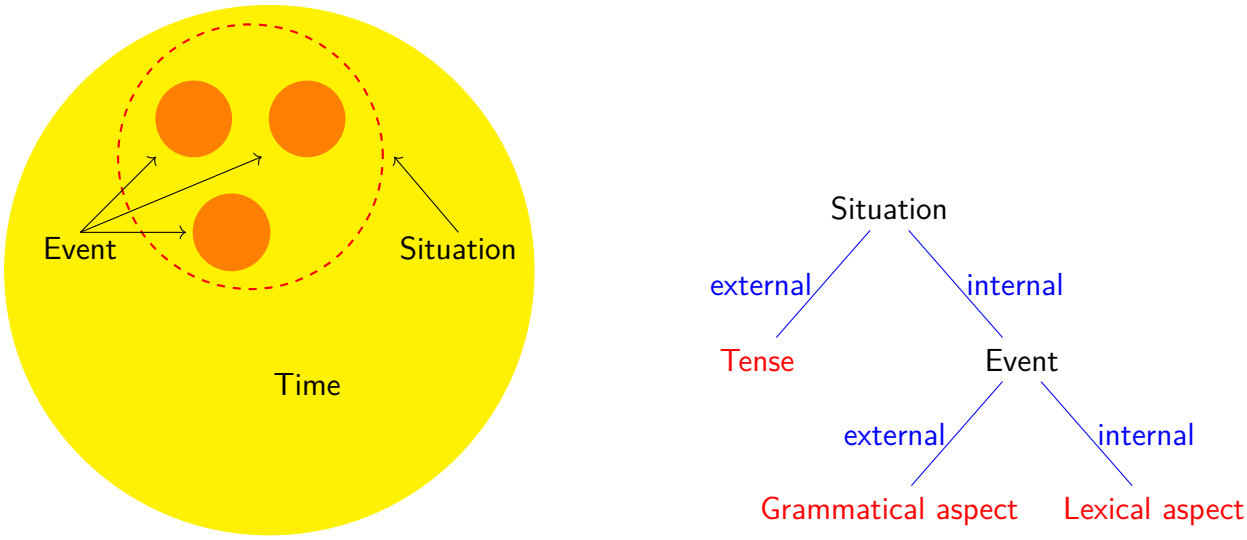


Figure 1: A visual representation of the relationship between time, situation and event *(left)*, along with the categorisation of the three temporal phenomena discussed *(right)*. Here each situation is a (non-empty) set of events ($\mathbf{situation} \subseteq \mathbf{events}$) and events (and by derivation situations) are anchored in time. In this (very simple) model, timeless statements such as "2+2=4" are also anchored in time, however this time period is infinite. This model serves purely for the visualisation of tense and aspect and will not be further employed.

2.1.1 Lexical and 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, aspect[3] is an umbrella term often used to refer to two slightly different phenomena: 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, irreversible 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, which is not an inherent feature but can be seen rather 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* +

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

gerund, as in 1.

(1) Sue was running.

Figure 1 provides a visual representation of the concepts introduced in this section for clarity. Here each situation is a (non-empty) set of events ($\mathbf{situations} = \mathcal{P}(\mathbf{events}) \setminus \emptyset$) and events (and by derivation situations) are anchored in time.[4]

The concrete linguistic realisation of these categories is very often unclear or not explicit 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". This has led some to question to what extent this distinction makes sense [Sasse, 2002]. Those who question the contrast between these two semantic dimensions, described as unidimensionalists in Sasse [2002], claim that aspectual distinctions in both dimensions can be reduced to a common set of semantic primitives, which can be applied to all levels of analysis, i.e. the boundedness of lexical aspect is the same as the boundedness which marks the distinction between the perfective and imperfective parameters.

I have chosen to introduce this distinction in order to introduce the terminology and clarify key concepts, however I will not strictly uphold the theoretical differentiation between these two in the experimental part of this study for reasons of pragmatism.

2.1.2 Lexical aspect parameters

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 (seeing as going climbing has no inherent goal), whereas "climb the mountain" is telic (since it involves the agent reaching the summit of a mountain). A classic test for telicity[5] is whether the verb phrase admits

4 In this (very simple) model, timeless statements such as "2+2=4" are also anchored in time, however this time period is infinite.

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

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* (see 2.1.4).

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], which is usually described as "dynamic". A classic test for stativity in English is non-admittance of the progressive or the imperative [McIntosh, 1975].[6] Consider the following examples:

(2) She resembles her grandmother.

(3) * 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:

(4) Andrea was painting a picture for an hour.

(5) * Andrea was reaching the summit for an hour.

(6) ? Andrea was cracking an egg for an hour.

6 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 and indeed seems to be characteristic of Modern American English (cf. the McDonald's tagline: "I'm loving it." [Freund, 2016])

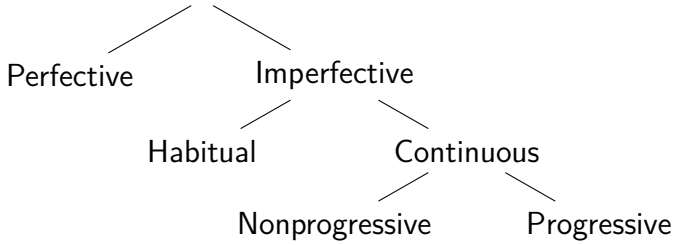


Figure 2: Comrie's classification of aspectual oppositions. Reproduced from Comrie [1976].

2.1.3 Grammatical aspect parameters

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 shown in 2, and in the following section I will briefly describe some of the main oppositions described in this classification: the perfective vs. imperfective opposition, (un)boundedness and habituality,

Perfectivity

The main aspectual distinction made by Slavic languages and, as can bee seen in figure 2, one of the main distinctions made in theoretical aspectology generally as that between perfective and imperfective. As noted by Dahl [1985], the theoretical distinction between perfectivity and imperfectivity is different to their concrete realisation in the Slavic language group. To avoid confusion, I will therefore henceforth use the terms to 'perfective' and 'imperfective' to refer to the theoretical binary opposition, and only refer to the Slavic (im-)perfectivity explicitly.

Returning to the definition of grammatical aspect as *situation-internal* but *event-external* [Kibort, 2008], perfectivity can be seen as a focus on the event as a whole indivisible entity (i.e. its result), and imperfectivity implies a focus on its internal temporal constituency [Comrie, 1976, Östen Dahl and Velupillai, 2013]. In practice this often means that perfective verbs refer to "completed" events such as 7, and imperfective verbs to incomplete or long-lasting events like 8, however note that this is not always the case.

(7) We walked the Camino de Santiago.

(8) It was raining for weeks.

2.1.4 Boundedness

(Un)boundedness refers to whether a situation described has reached a temporal boundary or not [Depraetere, 1995]. For example, while the act of crossing the road in 9 has no temporal bound on it (i.e. it is unclear when or if the event was ended), in 10 does have a definite end (bound). It is hard to fit into either of the categories of lexical or grammatical aspect, since it is neither an inherent feature of an event, nor is it usually imposed through grammatical means, but rather a derived feature of a predicate in context.

(9) We were crossing the road yesterday, when [...].

(10) We had crossed the road yesterday, when [...].

Boundedness must be distinguished from telicity[7] in order to avoid the so-called 'Imperfective Paradox' highlighted by Dowty [2012], 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 (11) entails (12) but (13) doesn't entail (14) in the examples below:

(11) The man was running.

(12) The man ran.

(13) The man was building a house.

(14) 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,[8] and this further serves to show the dangers of

7 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 Loáiciga and Grisot [2016], which, however, clearly distinguishes between the two notions.

8 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 is the case that the both (11) and (12) are bounded, only (13) and (14) 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.

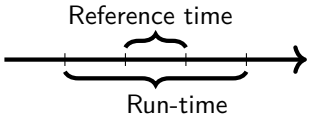
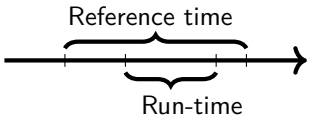
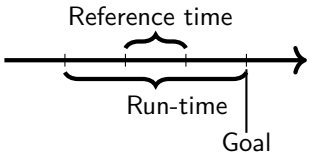
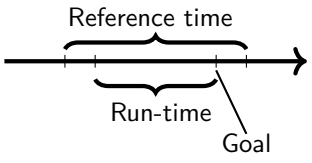
Sentence	Representation	Telicity	Boundedness
The man was running.		atelic	unbounded
The man ran.		atelic	bounded
The man was building a house.		telic	unbounded
The man built a house.		telic	bounded

Table 1: Representation of (11),(12),(13) and (14) with regards to the reference time referred to by the utterance and inherent run-time of the event. Concept adapted from van Hout [2016].

misuse of terminology. Table 2.1.4 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.

Habituality

Habitual events refer to events which happen with regularity or often. They are closely related to generic statements and can indeed be both at once in what [Dahl, 1985] calls "habitual generics", such as 2.1.4.

(15) Cherry trees bloom in April.

While usually still coming under the broad umbrella of aspectual phenomena,[9] habituality is generally considered to be located at a different level to other aspect features. Dahl [1985] notes that the prevalence of habituals in his questionnaire was low, which corroborates my findings in the dataset I used (see table 5.1.1), and that most languages that explicitly mark it use periphrastic means.

2.1.5 Some attempts at event classification

Vendler [1957]

It was the seminal paper *Verbs and Times* [Vendler, 1957] of philosopher Zeno Vender which initiated the discussion on inner aspect in the linguistic tradition.[10] Vendler begins his discussion in the paper 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 behavior of any verb whatever.

That is to say, the "time schema" of any verb can be described through comparison with a set of prototypical classes (see 2.1.5), 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.[11] 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:

9 However, Boneh and Doron [2010] argue that it is a first and foremost a *modal* category "which can only indirectly be characterized in aspectual terms".

10 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.

11 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 and lack of applicability to other languages.

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

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

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

Vendler states in his introduction that verbs "presuppose" certain time schemata and hence assigns a category to each verb (as in 2.1.5). However it must also be stated that the true profile of a verb phrase such as those listed in 2.1.5 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 when they took the photo." [Moens and Steedman, 1988]. Thus, the verbs (or verb phrases) mentioned by Vendler as belonging to a certain category are ones that are prototypical for the particular class. This is a fact that will become important later on (see 2.3.2), since it is not a trivial question whether all verb phrases inherently tend towards one particular lexical aspect which is saved in the mental lexicon, or whether some verb phrases have an under- or unspecified aspectual class. This is a question I will attempt to shed some light on with the experiments in this study.

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 example verb phrases given in Vendler [1957] for the classification of events.

Moens and Steedman [1988]

Moens and Steedman take a slightly different approach, focussing on the sentence level, where each sentence has a predicative *nucleus*, either a state or a process. While the former has no definitive start of end point, in the latter these are present and processes can be divided up into four classes by whether they are atomic or extended, and whether they imply a consequent state, i.e. the state after the action is different to the one before (see table 5 for the relationship between these classes and other classification schemata). Figure

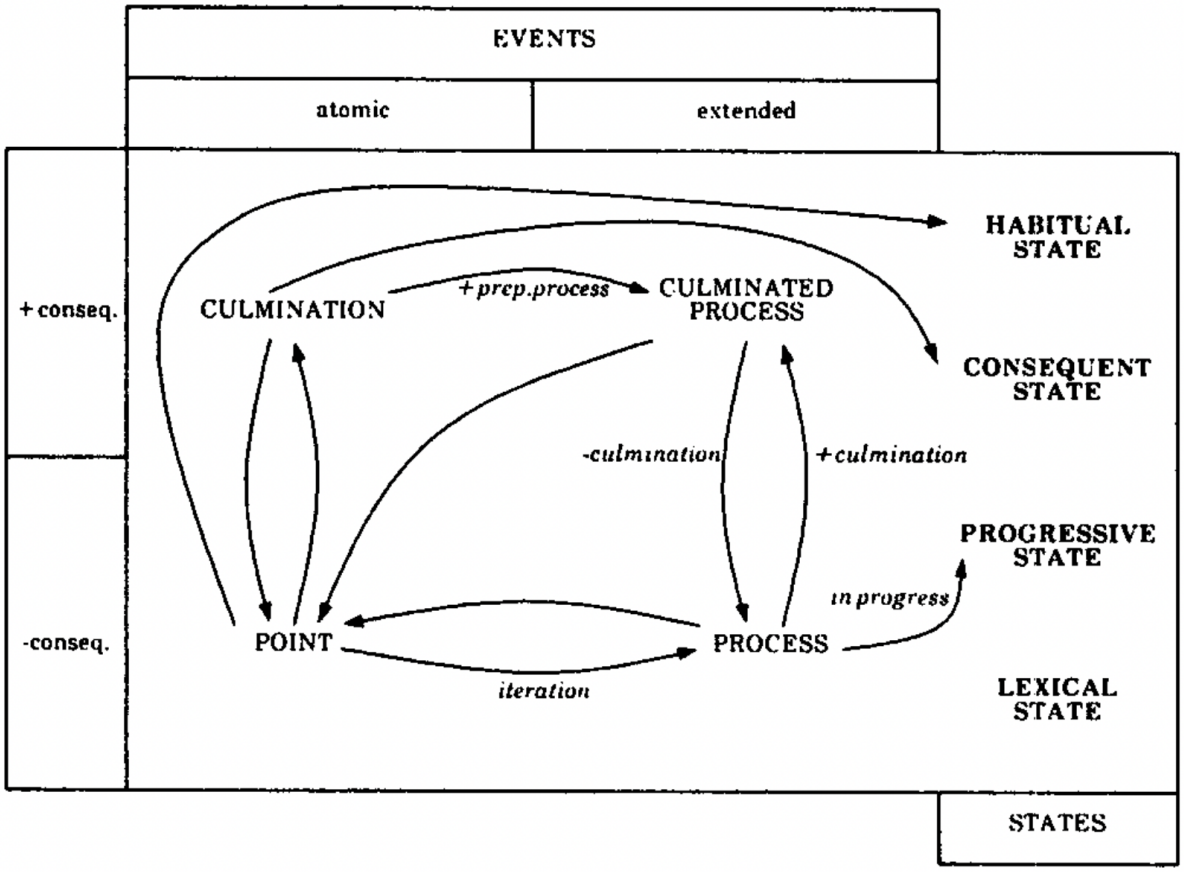


Figure 3: A visualisation of event- and process types with their corresponding transitions [Moens and Steedman, 1988]

2.2 Aspect in Slavic languages

The Slavic language group has a special place on the study of aspectology due to its overt encoding of aspectual phenomena, otherwise rather uncommon [Tomelleri, 2010].[12] Verbs in Slavic languages, with few exceptions,[13] are either perfective or imperfective, meaning the speaker must explicitly state the aspect of the verb event referenced. Consider the following sentences:

(16) Ja pročita-ju ėtu knig-u.
I read.PERF-FUT this book-ACC
'I will read this book.'

(17) Ja budu čita-t' ėtu knig-u.
I will read.IMPF-INF this book-ACC
'I will read (be reading) this book.'

While both are in the future tense, the former uses the perfective form of the verb to read, implying that the speaker will read the book *and* finish it (i.e. read it from the beginning to the end), whereas the second uses the imperfective form and thus expresses that the speaker will be reading the book at some point in the future, but not necessarily finish it.

As mentioned, the opposition between these two lexical categories differs slightly from the "purer" theoretical distinction discussed in 2.1.3. A classic example where the Slavic categories do not correspond to the semantic ones is the pair 18 and 19, where the former is a prototypical usage of the perfective to denote a whole completed event, the latter imperfective form denotes that the event has since been undone [Franks, 2005].

(18) Kto otkry-l okno?
Who open.PERF-PST.SG.MASC window?
'Who opened the window?'

(19) Kto otkryva-l okno?
Who open.IMPF-PST.SG.MASC window?
'Who opened the window?'

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

13 Note that there is a small group of Polish verbs which can be used as either imperfective or perfective (for more details see Kipka [1990])

It is also interesting to note that modern Slavic aspect, now usually seen as a *grammatical* feature, developed from spatial prefixes [Dickey, 2017] The most common perfectivising prefix *po-* used to refer to "up to", "across", or "at" in Proto-Slavic [Derksen, 2008], however has slowly developed into a temporal marker and finally into a grammaticalised aspect marker. The gradual lexicalisation of aspect systems in Slavic languages (i.e. the development from a system based more on contextual information, such as in other Indo-European languages, to one where grammatical aspect information is encoded in prefixes and sometimes even in the main lexeme) is further evidence in favour of a less clear distinction between grammatical and lexical aspect.

It is due to this relatively unique covert aspect marking that I decided to include Slavic languages in this study.

2.3 Aspectual ambiguity

While some verb phrases, such as *to tend to*, overwhelmingly only express one certain aspect which implies a habitual event, in many situations they can be ambiguous:

(20) Your soul was made to be **filled** with God Himself. *(Brown corpus, cited by Friedrich and Palmer [2014])*

In this example the verb 'filled' can be read as both a stative and a dynamic event.[14] Nevertheless, in many cases the reader tends to prefer a particular interpretation, such as in 21, where the sentence is most likely to be a habitual, a common use of the English simple present, and yet could also be interpreted as a historical present as in 22.

(21) He walks to school.

(22) He walks to school, and guess who he sees on the way!

This poses a practical problem for annotators when asked to provide a single class for a sentence (or sometimes even just for a verb phrase as in Siegel and McKeown [2000]), but also an

14 Indeed, while not specified in English, this is explicitly encoded in other languages such as German where the former interpretation would use the word *sein* (to be) and the latter *werden* (to become). The ambiguity in English therefore stems from the periphrastic marking of the passive with the verb *be*.

interesting theoretical question for linguists. For simplicity, most studies have assumed that aspect is unambiguous. While some have mentioned the topic as problem during annotation (cf. Croft et al. [2016], Friedrich and Palmer [2014]), there have been no studies to date looking aspect ambiguity [Friedrich et al., 2023], despite this being a not too uncommon occurrence in language (see 5.1). Some, such as Van Gysel et al. [2021] (see 4.1.1), have got around this problem by classifying aspect in a lattice structure, allowing for more coarse grained categories when necessary. While perhaps doable in practice, this is in unsatisfactory solution to describe several semantically distinct interpretations of a particular utterance.

2.3.1 Contexts frequently triggering aspectual ambiguity in English

While possessing an inventory of tools to specify aspectual features, there are many times in English where the aspectual interpretation of a verb phrase is less evident. This section aims to outline the most common contexts where some English verb phrases as ambiguous with respect to their aspectual reading.

Since, as mentioned above in 2.1.4, habituality operates on a different level to the majority of aspectual phenomena discussed so far, the question of when a habitual interpretation is possible is a slightly different one. The number of utterances where a habitual interpretation is feasible (if less likely) is large,[15] and indeed, [Dahl, 1985] finds that, when overtly marked, in many languages habituality is expressed by the simplest, least marked verb form (as in the English present simple), however that this is rarely unequivocal and shares several uses. This further supports the hypothesis of the high general feasibility of a habitual interpretation, and I therefore chose not to cover it in table 2.3.1. For example, sentence 23 exhibits the wide applicability of a habitual reading, even if it is not the most likely interpretation.

(23) They also said this court did not give the lawyers for the defense due procedure. *give*

RELATED WORK - Coercion and underspecification integrated: The state-event ambiguity of aspectual verbs (past simple); Automatic prediction of aspectual class of verbs in context (passive stative / dynamic)

15 This is supported by the findings of the manual annotation (see 5.1.2), where 24.8-42.7% of sentence-verb pairs had a possible habitual interpretation.

Context	Ambiguity	Example
repeatable events	[single event/iterative]	Peter knocked on the door [once/three times].
passive / past participle	[stative/dynamic]	The bottles were filled [with juice/in two hours].
past simple / future simple	[holistic/ongoing]	The rocks fell down [yesterday/slowly].
motion verbs with ambiguous prepositions	[telic/atelic]	Anna walked through the park [then turned left/for hours].

Table 4: Some common contexts triggering aspectual ambiguity in English.

There are other cases such as 24, where the aspect is less clear. Here we have a verb which would usually have a clear stative interpretation, however with a more "holistic" interpretation, i.e. the event is portioned into several bounded events (what HERWEG [1991] calls "pofective").

(24) Robert was in New York twice this year.

While not directly encoding dynamicity, telicity or iterativity, the choice of perfective/imperfective in Slavic languages is strongly influenced by boundedness and whether an event is a single situation or repeated [Wiemer and Seržant, 2017]. While the latter equates to iterativity, boundedness is also highly correlated with telicity and stativity (?), and hence the (obligatory) binary choice of aspect forces one of the two possible interpretations in each case. Apresyan [2024] also finds that L2 Russian speakers are more likely to choose the incorrect aspect if their L1 does not differentiate between these two interpretations in a particular case, corroborating intuition.

2.3.2 Coercion or underspecification?

21 and 22 are an example of how, unless further specified, a verb phrase in a sentence can tend towards a particular aspectual interpretation, while further specification leads to a different interpretation.

In cases where a verb can have several aspect readings depending on the context, a question that currently remains unclear is whether all verbs have an inherent aspect class, which is *overwritten* by the features of the context is occurs in (temporal adverbs, other verb phrases forcing a particular class such as *to tend to* etc.), or whether some are simply underspecified in the mental lexicon,

and the class is *determined* by these circumstantial features [Gerwien and Herweg, 2017]. It is also often the case that a verb phrase has a particular aspectual interpretation in the overwhelming majority of cases, however it can also be used in a different aspectual context without this sounding strange.

The former view, as put forward by Moens and Steedman [1988] and known as *coercion*, is used as an assumption in many previous works (cf. de Swart [2019]) with little evidence of its uniform validity [Gerwien and Herweg, 2017]. It is therefore an open question how both humans and language models deal with this aspectual conflict. The following study will aim to shed some light on the behaviour of the latter in such situations, and perhaps provide some insight into possible mechanisms of the former.

3 Related Work

3.1 Aspect classification

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] This is presumably due to the fact that it is a high-level semantic task, whose relevance to downstream applications is perhaps not as immediately obvious as other similarly complex tasks. However, there have been some works in recent years looking at this area and studying how well current models deal with the phenomenon.

3.1.1 Rules-based aspect class prediction

Siegel and McKeown [2000] develop a rules-based aspect classification of verbs using co-occurrence information from a corpus. The aspect class they aim to predict is an inherent class of a verb (i.e. lexical aspect), and hence they do not take into account the sentence context, which often has an effect on the aspect of the verb considered. This is one of the issues which makes the aforementioned distinction between lexical and grammatical aspect so difficult in practice. The aspect classification scheme they use was that of Moens and Steedman [1988], a 5-way class distinction building on Vendler [1957].

Interestingly they use their results draw the following linguistic conclusions: HERE.

Egg et al. [2019] do something cool too.

TALK ABOUT ASp-Ambig - be we have a different type of ambiguity

Chen et al. [2021] do something.

3.1.2 (L)LMs and aspect

Metheniti et al. [2022] are bros and it WORKS.

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

Katinskaia and Yangarber [2024]

3.2 Aspectual ambiguity

3.3 Available datasets

The available datasets are relatively sparse.

One particularly interesting example is English-Czech InterCorp (GET CITATION C$^{\check{}}$ermákandRosen, 2012; Rosen and Vav$^{\check{}}$rín, 2012), which leverages the fact mentioned in 2.2 that Slavic languages such as Czech have two forms of each verb, each assigned to a different aspectual reading depending on the context. This makes it possible to extract aspectual information from a Czech translation of an English sentence, assuming an accurate translation, to use as further training data.

3.4 Formal representations of aspect

In this section I will briefly mention some of the attempts to formalise aspect

3.5 Related areas of work

RED also identified fine-grained aspect claasses (or *Aspectual image schemata* as they termed it) which often hard to distinguish between at first glance, however this is a

They used this to create a semantic map

The approach is a different one to mine since the goal was to find groupings of semantically similar aspect classes, whereas the classes presented in table 2.3.1 are contexts (syntactic or otherwise) which lead to an ambiguous aspect reading and does not necessarily say anything about the semantics of the aspect classes themselves.

3.5.1 Situation entities

Situation entities classification is the task of identifying different types of situations, which exist at a clausal level. The task is comes more from the tradition of discourse analysis, since it is important for discourse representation theory (DRT) to know for example which new referents are introduced to a discourse and or also to analyse temporal relationships. Works usually use the original 8 types introduced by Smith [2003]: events, states, generalizing sentences, generic sentences, facts, propositions, questions and imperatives. While

4 Methods

This section will give an overview of the main research questions for this project and how I intend to investigate them. CHECK TENSE IN THIS SECTION

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.

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

The classification schema I decided on was Uniform Meaning Representation (UMR) [Van Gysel et al., 2021], for several reasons: On (SHOULD THIS BE CAPITALISED??) the one hand the lattice was designed for ease of annotation and usability and thus does away with a lot of the theoretical baggage of other classification systems (such as the insistence on lexical aspect types). On the other hand the schema is part of a larger framework, and hence a classification system using these labels has a practical application. Furthermore, UMR provides a lattice for annotation (see 4) meaning the level of the annotation classes can be ANGEPASST to the needs of the individual contexts.

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].

Aspect in UMR

UMR describes the following 5 coarse-grained aspect classes: (descriptions taken from Van Gysel et al. [2021][1]), also depicted in figure 4:

- **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 5.

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 4, 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 4. This allows for annotation both at a

1 For more detailed descriptions see GUIDELINES LINK!!!!

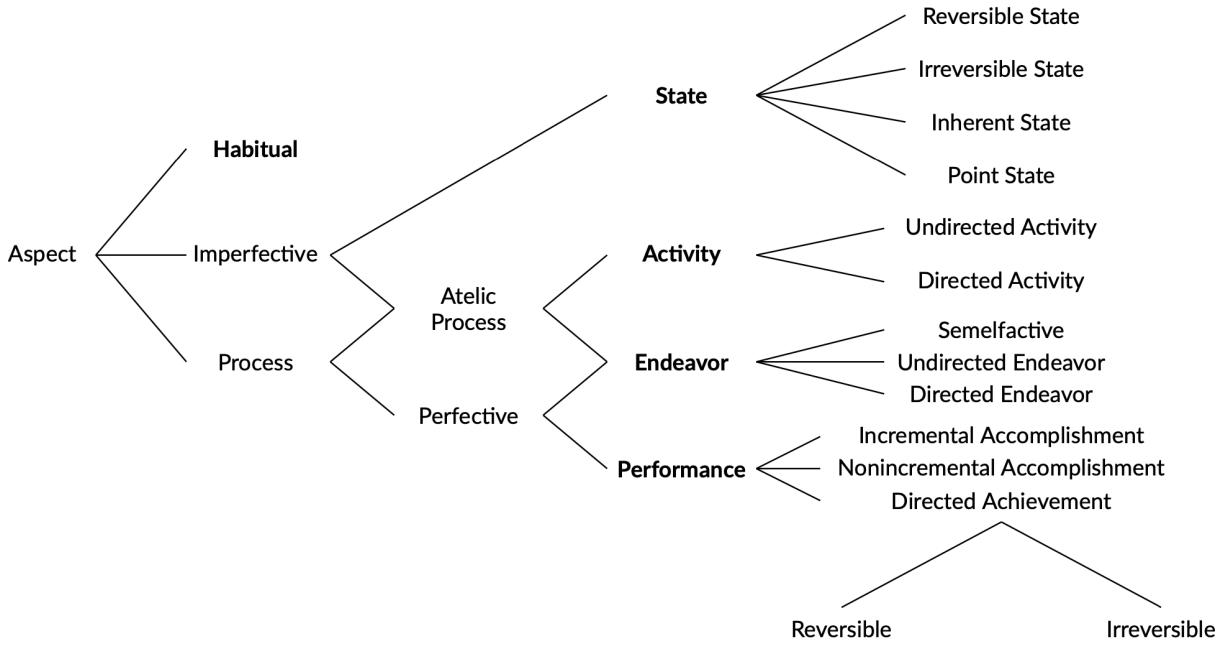


Figure 4: UMR aspect classification lattice [Jens Van Gysel and Xue, 2022]

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

Table 5: Comparison of aspectual classes. Adapted and extended from Egg, Prepens, and Roberts [2019].

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.

Difficulties with aspect in UMR

Slightly different to other frameworks (one category can encompass imperfective and perfective?) - is this true

Habitual as aspect

They are not based on classic parameters such as telicity, punctuality etc.

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 Research questions

RQ1: Can we train a system to identify aspect classes?

The first question I aimed to look at is whether we can train a system to automatically classify verbs in context as one of several aspect classes. This has been done before (see 3.1), however this is the first time that a large language model has been used for this task (CHECK THIS). If this model works well, it will be possible to use it to look at some linguistic questions, such as whether particular prefixes is Slavic languages tend towards certain aspect classes, or carrying out a typological analysis of aspectual ambiguity (RQ3).

RQ2: Can we automatically identify aspectually ambiguous verbs/sentences and which classes they tend towards?

Once I have a model which can accurately predict verbal aspect in context, the next question will be the more complex task of predicting aspectual ambiguity, which as far as I am aware, has not been done before. At this stage it is also an interesting question to examine for verb phrases how they are classified without context. This could SHED SOME LIGHT on inherent aspectual readings of verb phrases and thus ANSWER THE QUESTION of whether there are some verb phrases which are perhaps "underspecified" in the model's latent space (see 2.3.2).

RQ3: How does aspectual ambiguity compare across languages? Can we study the typology of aspectualisation?

Finally, I will attempt to investigate and compare aspectual ambiguity across languages, looking at whether languages with more explicit marking for aspect have less aspect ambiguity than those expressing aspect more implicitly. I plan to answer this both by looking at the ambiguity classification of verbs in context and also of the top *n* verbs in the language without context.

4.3 Project outline

Step 1: Fine-tune LLM for aspect classification

In order to use language models to investigate aspect, it is necessary to first understand how much current language models know about aspectual phenomena. To this end, I will fine-tune a large language model on a small dataset (since no large datasets are available for the aspect classification scheme I choose, UMR [Van Gysel et al., 2021], or indeed in general for an aspect classification task), testing on a hold-out set. This step of fine-tuning a *large* language model is necessary due to the scarcity of data available, which is not sufficient for fine-tuning a smaller model.

Step 2: Use fine-tuned LLM to annotate larger dataset

The LLM fine-tuned in step 1 can then be used to annotate a larger dataset. Having a larger dataset allows a great more possibilities for analysis than the dataset used for fine-tuning the larger model. For instance, this larger dataset can be used in a sort of knowledge distillation (KD) process to train a smaller (possibly multilingual) model, which is useful for latent space analysis or comparison across languages. Though it must be noted that there is a danger of error propagation at this step.

Step 3: Compare aspectual systems cross-lingually

Once we have training data of good enough quality, we can use this to train a multilingual model and use this to compare between languages. Here we can test the model in different contexts and use this to gain insights about the aspect systems of different languages, both in their own right and in comparison with one another, for example, to see if prevalence of aspectual ambiguity changes depending on how overtly a language marks for aspect, or if certain prefixes in a language mark a particular aspect class. Furthermore,

EQUATION FOR LANGUAGE LEVEL ENTROPY

Step 4: Fine-tune LLM for aspect ambiguity detection

In order to look at aspect ambiguity specifically, I will also fine-tune an LLM in a similar way to above to identify whether a verb in context has an ambiguous aspectual reading or not. However, since there does not yet exist a dataset annotated for aspectual ambiguity, this part requires human annotation. I propose to solve this by requiring annotators to label sentence-verb pairs with any labels they see plausible. This makes it possible to have both a human class annotation and a derived ambiguity annotation for datapoints labelled with several labels. Neural networks are known for being overly confident in their predictions, and I will investigate whether low certainty of the smaller aspect *classification* model (i.e. similar logit values across classes) correlates with the output of the model specially trained for ambiguity recognition.

4.4 Use of approximative LMs as sources of linguistic knowledge?? (Reversing the NLP pipeline)

Currently the two main approaches used to develop and validate linguistic hypotheses are through corpora and through introspection, the former being championed by empiricists and the latter by the Chomskyan rationalist tradition [McEnery and Wilson, 2001]. It is clear that corpora, including those used to train LLMs, can only contain a fraction of the famously infinite set of possible grammatical sentences in a language, and this has led Chomsky to decry corpus linguistics as seeking to model language *performance* rather than *competence* [McEnery and Wilson, 2001]. Native speaker introspection, on the other hand, while able to judge the grammaticality of any

sentence, is clearly highly subjective and biased. Language models, however, are productive and are able to generalize across their corpora to produce, with some sophistication, sentences not seen before in training, thus blurring the line between the traditional Chomskyan distinction between competence and performance. CHANGE THIS WORDING SO ITS NOT WORD FOR WORD AND CITE IT MAYBE??

4.5 Where do LLMs fail? (CHANGE - expectation management)

WAS KANN ICH WISSEN? WAS DARF ICH HOFFEN? KANT It is easy to get

COPY FROM PAPER?

In light of the current hype surrounding deep learning it is important to highlight the limitations of such techniques and what language models *cannot* do.

It is well-known that training the LMs discussed in this paper requires a large amount of data and computing resources. While pre-trained models mean that LMs trained on relatively large amounts of data are now available for general use, for less well-resourced languages this is a problem, and their performance suffers drastically. While this does not rule out the use of LMs on such languages (see [Kholodna et al., 2024]), it certainly limits the applicability of some of the uses highlighted in this article, such as for annotation or acceptability analysis.

Furthermore, it must also be noted that LMs take on any biases present in the training data, meaning the language they approximate should be treated with caution. Examining these biases, as has often been done before, can, however, be an area of study in its own right and can produce valuable data for sociolinguistics. However, it must also be noted that it cannot be guaranteed that characteristics of a model's latent space can be transferred to a more general linguistic space, since human linguistic competence and LM competence differ in some aspects. Further research is therefore needed in this area.

Finally, since LMs are trained to minimize error on one variety of a language, they are less well-suited to study linguistic variation, whether geographical or temporal. This makes their use less suitable for languages without an accepted standard variant, such as Swiss German. In these

cases, however, an approach using word embeddings such as Hamilton et al. [2016] could still be useful. CHANGE WORDING HERE

5 Dataset Creation

RENAME THIS CHAPTER DATASET CREATION??????

5.1 Dataset creation - should this go inside experiments?

In order to carry out further computational analysis of the topic it is helpful to have an annotated dataset. Since no extensive 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 combined with human annotators [Goel et al., 2023] or on their own [He et al., 2023, Yu et al., 2023, Gilardi et al., 2023], which is what I opted for.

5.1.1 Dataset annotation with LLMs

Large Language Models' use for annotation data has been demonstrated GIVE SOME EXAMPLES.

Törnberg [2024] formulated a set of best practices when using LLMs as text annotators, which I aim to be guided by. The paper provides guidelines for those wishing to use LLMs as text annotators, in the absence of labelled data. This situation differs from the "paradigmatic" one described in the paper by the fact that there already exist annotation guidelines and limited labelled data, hence the iterative systematic coding procedure described in the paper is not necessary. Nevertheless Törnberg [2024] provides a well-needed framework for this relatively new approach.

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

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 efficint 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.

During the course of this study, Llama 3 was released to the public CITE LLAMA 3, promising to have better reasoning skills and be better at following instructions effectively, so I therefore tried and compared both models in comparable settings.

Training data

As the only currently existing data containing UMR aspect classes, I used the example UMR dataset[1] provided by the creators of UMR. It contains HOW MANY sentences in 6 languages

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

Class	No. examples in UMR data	No. examples in upsampling data	**Total**
Performance	124	1	125
State	55	0	55
Activity	36	2	38
Endeavour	17	14	31
Habitual	2	31	33
Total	234	48	282

Table 6: Aspect classes of annotated verbal events in the UMR dataset.

(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 concluding that habituals and related aspect classes have a low frequency in actual use, and hence I manually added some datapoints to improve the balance. The added sentences were either found in existing online datasets and simly 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 very improbable 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 (see for example 5.1.1).

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 to focus on verbal events. In

```
(s1p / publication-91
   :ARG1 (s1l / landslide-01
      :ARG3 (s1a / and
         :op1 (s1d / die-01
            :ARG1 (s1p3 / person :quant 200)
            :aspect state)
         :op2 (s1f / fear-01
            :ARG1 (s1m / miss-01
               :ARG1 (s1p2 / person :quant 1500)
               :aspect state)
            :aspect state)
         :aspect process)
   :place (s1c / country :wiki "Philippines"
      :name (s1n / name :op1 "Philippines"))))
```

Figure 5: UMR graph of *200 dead, 1,500 feared missing in Philippines landslide.*

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

Prompt engineering is a new but important field in NLP, and studies have shown the importance of good prompts when dealing with LLMs [Kaddour et al., 2023, Hsieh et al., 2023, Sahoo et al., 2024]. Törnberg [2024] recommends an iterative prompt engineering process of

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.

Prompt type	Llama 2 7B	Llama 3 8B
No definitions	0.198	0.479
Normal	0.748	0.769
Long	0.688	0.740

Table 7: F1 score on test set from fine-tuning Llama 2 7B and 3 8B on different types of prompt

Class	Precision	Recall	F1-Score	Support
Performance	0.90	0.56	0.69	16
State	0.00	0.00	0.00	2
Activity	0.88	0.58	0.70	12
Endeavour	0.50	0.50	0.50	4
Habitual	0.76	1.00	0.86	37
Accuracy			0.77	71
Macro Avg.	0.61	0.53	0.55	71
Weighted Avg.	0.77	0.77	0.75	71

Table 8: Results on test set from fine-tuning Llama 2 on UMR aspect classes without upsampling. (UPDATE FORMATINGG!!!!!)

followed by the following question:

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

The results show that without an explanation of the classes, the model fails, achieving roughly random accuracy. WHICH EPOCH ARE THESE RESULTS FORMATINGG

Interestingly the results also show that the extended prompt (WHERE TO LINK THIS) also exhibited poorer performance. This could be due to the fact that it contained more redundant information which is

Quantitative analysis of results

The models seemed to learn the aspect classes relatively well, and it is clear that in general, manual upsampling improved results across the board, as is to be expected. However, the comparison of this result to the one on the model trained without upsampled data should be taken with a grain of salt, since the testing set also included some upsampled data (as there were not enough of some classes - mostly habituals - to make it a fair test), which are by design more prototypical and

Class	Precision	Recall	F1-Score	Support
Performance	0.78	0.88	0.82	16
State	1.00	0.75	0.86	8
Activity	1.00	0.56	0.71	9
Endeavour	0.89	0.62	0.73	13
Habitual	0.81	0.97	0.88	39
Accuracy			0.84	85
Macro Avg.	0.90	0.75	0.80	85
Weighted Avg.	0.85	0.84	0.83	85

Table 9: Results on test set from fine-tuning Llama 2 on UMR aspect classes with upsampling. (UPDATE FORMATINGG!!!!!)

thus "easier" than the natural data. Nevertheless, this result taken as a standalone one shows that the model was capable of recognising occurrences of the five UMR aspect classes with relatively high accuracy.

5.1.2 (Manual!!) Ambiguity annotation

Since there are no existing datasets for aspectual ambiguity, I decided to annotate the training data I created for the Llama model. This means that I end up with a database of ambiguous sentences which can later be used for testing. This is a relatively hard task for annotation since it is rather open-ended (with $\sum_{i=1}^{5} \binom{5}{i} = 31$ different possible annotations for a verb-sentence pair), and the interpretation of aspect is often rather a subjective process.

The annotation was done by an English native-speaker with a background in philosophy and myself. The annotators were given a detailed description of the UMR classes and asked to assign a class to each sentence-verb pair. If several readings were possible, the annotators were asked to write down all possibilities, indicating an aspectually ambiguous sentence. The annotators were also given the possibility to indicate when they were unsure on a particular sentence. See A.1.2 for the exact annotation guidelines.

For example, the following sentence-verb pair was annotated as belonging to `State` and `Activity` classes and hence has an ambiguous aspectual reading:

(25) The first footage from the devastated village showed a sea of mud covering what had been lush green valley farmland. **covering**

	Annotator 1		Annotator 2	
	Absolute freq.	Proportion	Absolute freq.	Proportion
Agreement with UMR[2]	202	0.706	202	0.706
Several labels (ambiguous)	71	0.248	71	0.248
Unsure	28	0.098	28	0.098

Table 10: Analysis of annotated data. GET FOOTNOTE TO WORK HERE!!! AND UPDATE FOR ANNOTATOR 2

5.1.2.1 (Manual!!) Annotation results

The results of the annotations can be seen in table 5.1.2.1.

It was interesting to note that the majority of occurrences of the habitual class (30 out of 34 or 88% AND WHAT FOR ANNOTATOR 2) were also annotated with other labels, empirically motivating the theory that habituality is located in a different level of aspectual distinction than the other classes. This is also reflected in the UMR aspect lattice (see figure 4), where habituals are on the second highest level of the tree.

Almost all (WHAT NUMBER) the sentence-verb pairs marked with several verb classes by annotator 1 were also marked as ambiguous by annotator 2, however only WHAT PERCENTAGE of those classified as ambiguous by the latter were seen as such by annotator 1. This causes a problem with

(DIFFERENT SETUP NOW)This presents a problem in how to design the experiment set-up. Since we want to encourage the model to learn prototypical examples of aspect classes to make sure it is not overly confident on ambiguous datapoints, it would make sense to remove the datapoints labelled as having more than one class from training set. However, this would lead to the habitual class being almost non-existant in the dataset. One solution would be to leave them in labelled as habituals, but this would lead to WHAT WOULD IT LEAD TO

I therefore chose to keep datapoints which are

5.1.2.2 Consequences for aspectology

Habitual on a different level

5.1.3 Larger dataset

In order to fine-tune the smaller model, a larger dataset is needed than the 209 sentences in the UMR example dataset. The only requirements for this larger dataset are that it is representative, clean and has enough exmaples to train the smaller model. It would also be an advantage if it is a sentence-aligned multilingual corpus, since this would make it possible to test the performance of a multilingual model in other languages where there are no labelled data from the Llama model.

WHY DID WE TAKE OUT HABITUAL? AND NOT ITERATIVE??

6 Experiments

Discussing the results of experiments. PUT DATASET INSIDE HERE

6.1 Aspect classification

6.1.1 BERT fine-tuning

Due to their size (and the subsequent large amount of data stored in the parameters), it is hard to carry out probing on LLMs, and, due the fact that they appeared recently, there has been less time to develop probing techniques, hence I decided to train a smaller model of the BERT [Devlin et al., 2019] family. These require more data to fine-tune than LLMs, and hence I used the fine-tuned Llama model to generate more training data in a technique known as knowledge distillation (KD). Knowledge distillation involves using a larger 'teacher' model to train a smaller 'student' model, often reaching the same or similar accuracy with a significantly smaller model. While not being able to probe the larger Llama model, smaller models are still an interesting artefact to consider and

Furthermore, the complexity of larger LMs (especially those trained for generation rather than sequence classification (CHECK THIS!!!!)) means that their embedding space is noisier than smaller models that have been fine-tuned. This makes it harder to do the sort of probing described in later in this chapter.

6.1.2 Aspect latent space

Figure 6 provides a visualization of the `[CLS]` token embedding of verb-sentence pairs in the training set, together with their aspect label, from which several interesting observations can be made. For instance, it is interesting to note the positioning of `habitual` instances between `state` and `activity`, which accurately captures their semantics as somewhere between the more generic, non-episodic state and activities, denoting "an event [that] has not necessarily ended and may be ongoing at Document Creation Time (DCT)" [Van Gysel et al., 2021].

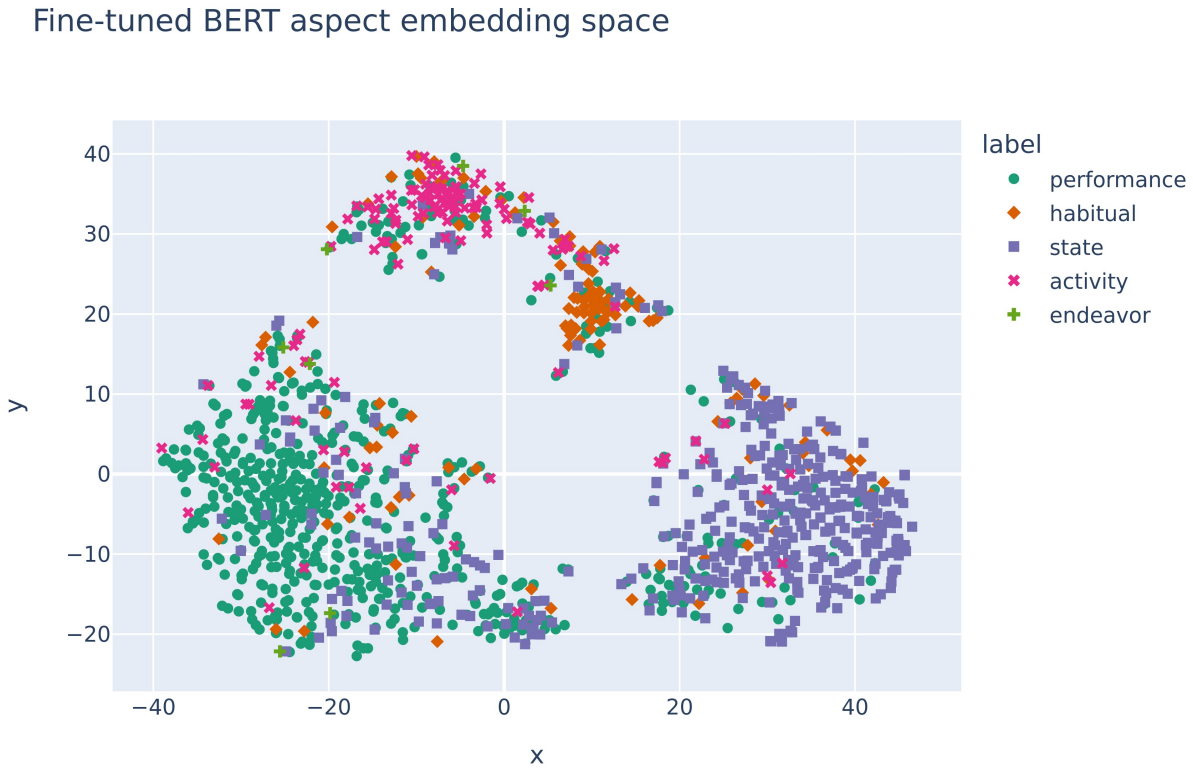


Figure 6: [CLS] embedding space of a BERT model fine-tuned on English verbs annotated for aspect in context, reduced to 2 dimensions by t-SNE.

6.1.3 Multilingual BERT fine-tuning

Since the LLM annotator Llama 3 can only be used in English, it is only possible to make training data in English. Luckily, however, there are multilingual models which can be trained with data using one language and then be used with languages other than that of the training data. I wanted to verify the performance of different models transferred to other languages and therefore needed a way of obtaining target labels for languages other than English.

I therefore trained a WHICH MODEL using WHICH DATASET annotated by the fine-tuned Llama 3 model and applied. The choice for this particular dataset was the availability of word-level alignments, a rarity in parallel corpora (IS THIS TRUE), meaning once a verb has been identified in the English sentence and given a particular aspect class, it is easy to find the corresponding verb in the other language. I only included sentences where the corresponding word in the other language is also a verb, since I wish to focus on verb phrases rather than general event classification.

In order to check the performance of this

Get results in English and results in French

ENTER RESULTS

Hence, it is clear that the result is worse than in English, as is to be expected, however still significantly higher than chance.

6.1.3.1 Russian aspect system

As explained in 2.2

6.1.3.2 Verb of motion clustering

Unlike in English, almost all Russian imperfective verbs of motion have both a telic and an atelic counterpart. The former is used to describe a motion with a destination, the latter for one without.

(26) Ja *SHel* v SHkolu. (I was walking to school.)

(27) Ja *hodil* po parku. (I was walking around the park.)

In UMR, events with no end goal are PROTOTYPISED? by activity, whereas telic events reaching completion are annotated with `performance`. We can verify that the model also makes this distinction HOW.

this empirically using the

Using this we can

6.1.3.3 Prefix clustering

An interesting application of this fine-tuned model is the

6.2 Aspectual ambiguity

6.2.1 Sentence-level ambiguity

DOES ENTROPY CORRELATE WITH AMBIGUITY PREDICTION?? It is well known that LMs are often too confident of their predictions, even when wrong INSERT CITATION HERE

The first question that is interesting to ask (CHANGE DIESE FORMULIERUNG) is whether there is a correlation between the uncertainty of the aspect classification model and the output of the aspect ambiguity model, i.e. does a (supposed) ambiguous aspect reading of a sentence-verb pair correlate with uncertainty in the former's output. In order to quantify I take the concept of entropy and apply it to this case with the following formula:

$$H_{aspect} = -\sum_{i=1}^{\#AspClass} p(x_i)log(p(x_i))$$

In this way, higher uncertainty (i.e. a more balanced probability across all classes) leads to a higher H_{aspect} value. It must be noted that this value is not comparable with models outputting a different number of classes (such as the traditional Vendlerian classification with 4), or indeed

with different aspect classification systems (IS THIS TRUE??), however it serves the purpose for use to compare between languages (REWORD).

Using this value it is possible to calculate a correlation coefficient. The measure used was the point-biserial correlation coefficient, a metric mathematically equivalent to Pearson's correlation coefficient, however specialised for the case of one binary and one continuous variable. It is calculated thus:

$$r_{pb} = \frac{\overline{X}_1 - \overline{X}_0}{s}\sqrt{\frac{n_1 n_0}{n^2}}$$

where:

- $\overline{X}_1$ is the mean of the continuous variable for the group where the binary variable is 1
- $\overline{X}_1$ is the mean of the continuous variable for the group where the binary variable is 0
- s is the standard deviation of the continuous variable
- n_1 is the number of observations in the group where the binary variable is 1
- n_0 is the number of observations in the group where the binary variable is 0
- n is the total number of observations

6.2.2 Verb-level ambiguity

6.2.3 Language-level ambiguity: Cross-linguistic comparison

7 Discussion

Discussing the results.

8 Conclusion

Verbal aspect is a prominent linguistic phenomenon.

In this thesis I have investigated the performance of current tools for aspect classification, using a classification schema taken from a larger meaning representation framework: UMR.

I exhibited some of the uses of this fine-tuned model for linguistics, such as Slavic prefix clustering with respect to aspect, and telicity detection in verbs of motion. Furthermore, I empirically validated the typological hypothesis that Slavic languages WHAT, using the language-level entropy

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

A.1.2 Annotation guidelines

Aspect Annotation Guidelines

This task is about aspect. Aspect is usually grouped as part of a larger linguistic system including tense and mood; however, it is distinct from both. While tense describes where in time an event takes place, aspect is the "lens" through which the event is viewed. For example, the events in both (1) and (2) take place in the past, but they differ through their aspect, meaning different parts of the event are emphasised.

(1) Anna was walking to the park.
(2) Anna walked to the park.

In (1) the continuous aspect is used, meaning the emphasis is on the act of walking to the park, whereas in (2) the emphasis is on the end of the event, i.e. the fact that Anna reached the park. This is why (3) makes sense, but (4) does not.

(3) Anna was walking to the park when she saw Ben. ✅
(4) Anna walked to the park when she saw Ben. ❌

Your task is to classify the verb phrases in the context of the following sentences as one of 5 aspect classes. This 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. This is also used for modal verbs (*want*, *need* etc.).

I am a doctor. – The glass is shattered. – They have a cat. – He's lying on the bed.[1]

The **Habitual** value is annotated on events that occur usually or often.

I usually wake up at 7. – I go to work by bike.

The **Activity** value indicates a process where it is not clear whether the event has come to an end. This also covers events in the present tense.

He was writing his paper yesterday. – She was phoning someone when I saw her. – He is singing. – They started to laugh. – She kept on playing the violin.

Endeavor is used for processes that end without reaching completion/termination (i.e. an end-point inherent in the process itself).

They mowed the lawn for 30 minutes. – We were walking until dusk.

Performance is used for processes that reach a completed result state.

He denied any wrongdoing. – We reached the summit in 4 hours.

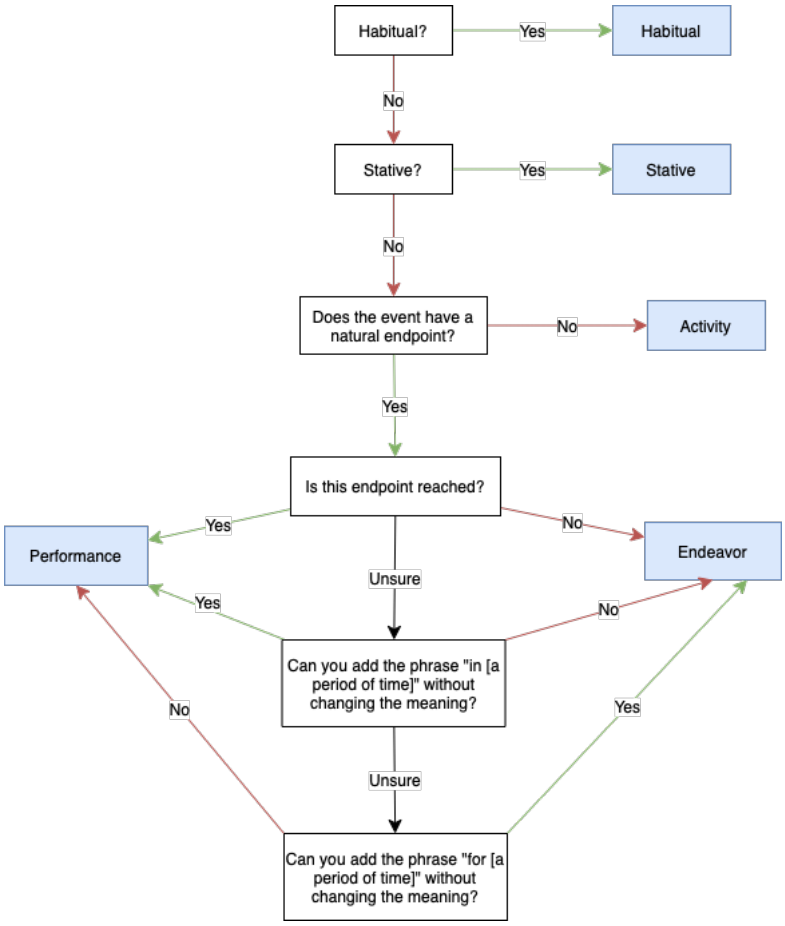
Some sentences have several possible interpretations. For example: "Let the streets be **filled** with song" could be both a state and an endeavor. This is an important part of the annotation. In this case, please enter all classes which are plausible, separated by a comma.

[1] This example is ambiguous and could also refer to an activity since it is a so-called "inactive action".

For convenience, please just enter the *first letter* of the correct class corresponding to the sentence, as in the table shown. See below for an example.

State	S
Habitual	H
Activity	A
Endeavor	E
Performance	P

For further guidance, see the official UMR guidelines or the following flowchart:



If you are unsure, please enter which of the 5 classes you think fits best, along with a 'Y' in the 'Unsure?' column.

Example:

Sentence	Class	Unsure
He was writing his paper yesterday.	A	
He's lying on the bed.	A,S	
This sentence is hard to figure out.	S	Y

Bibliography

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

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.

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.

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.

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.

N Freund. Recent change in the use of stative verbs in the progressive form in british english : I'm loving it. 2016. URL https://www.reading.ac.uk/elal/-/media/Project/UoR-Main/Schools-Departments/elal/LSWP/LSWP-7/elal_7_Freund.pdf?la=en&hash=C1403C6C7805F3F1CDA6E1F5E5CC4BEE.

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

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.

Anna Kibort. Aspect, 2008. URL http://www.grammaticalfeatures.net/features/aspect.html. Grammatical Features, 7 January 2008.

Östen Dahl and Viveka Velupillai. Perfective/imperfective aspect (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo, 2013. doi: 10.5281/zenodo.7385533. URL https://doi.org/10.5281/zenodo.7385533.

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.

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.

D.R. Dowty. *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Studies in Linguistics and Philosophy. Springer Netherlands, 2012. ISBN 9789400994737. URL https://link.springer.com/book/10.1007/978-94-009-9473-7.

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

Nora Boneh and Edit Doron. 33816 Modal and Temporal Aspects of Habituality. In *Lexical Semantics, Syntax, and Event Structure*. Oxford University Press, 02 2010. ISBN 9780199544325. doi: 10.1093/acprof:oso/9780199544325.003.0016. URL https://doi.org/10.1093/acprof:oso/9780199544325.003.0016.

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.

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=.

P.F. Kipka. *Slavic Aspect and Its Implications*. MIT working papers in linguistics. Massachusetts Institute of Technology, Department of Linguistics and Philosophy, 1990. URL https://dspace.mit.edu/handle/1721.1/13649.

Stephen Dickey. *Prefixation in the Rise of Slavic Aspect*, pages 85–102. 01 2017. ISBN 978-88-6453-697-2. doi: 10.36253/978-88-6453-698-9.07.

Rick Derksen. Etymological dictionary of the slavic inherited lexicon. 2008. URL https://brill.com/display/title/12607?language=en.

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.

Eric V. Siegel and Kathleen R. McKeown. Learning methods to combine linguistic indicators:improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–627, 2000. URL https://aclanthology.org/J00-4004.

William Croft, Pavlina Pešková, and Michael Regan. Annotation of causal and aspectual structure of events in RED: a preliminary report. In Martha Palmer, Ed Hovy, Teruko Mitamura, and Tim O'Gorman, editors, *Proceedings of the Fourth Workshop on Events*, pages 8–17, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1002. URL https://aclanthology.org/W16-1002.

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 Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. Designing a uniform meaning representation for natural language processing. *KI - Künstliche 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.

MICHAEL HERWEG. A Critical Examination of Two Classical Approaches to Aspect. *Journal of Semantics*, 8(4):363–402, 11 1991. ISSN 0167-5133. doi: 10.1093/jos/8.4.363. URL https://doi.org/10.1093/jos/8.4.363.

Björn Wiemer and Ilja Seržant. Diachrony and typology of slavic aspect: What does morphology tell us? 05 2017. doi: 10.5281/zenodo.823246.

Valentina Apresyan. Errors in foreign language acquisition as a multifaceted phenomenon: the case of russian aspect. *Russian Linguistics*, 48, 01 2024. doi: 10.1007/s11185-023-09287-8.

Johannes Gerwien and Michael Herweg. Aspectual class (under-)specification in the generation of motion event representations – a project outline. *Heidelberg Papers on Language and Cognition*, Bd. 1:2017, 2017. doi: 10.11588/HUPLC.2017.0.37820. URL http://journals.ub.uni-heidelberg.de/index.php/huplc/article/view/37820.

Henriëtte de Swart. *10. Mismatches and coercion*, pages 321–349. De Gruyter Mouton, Berlin, Boston, 2019. ISBN 9783110626391. doi: doi:10.1515/9783110626391-010. URL https://doi.org/10.1515/9783110626391-010.

Daniel Chen, Martha Palmer, and Meagan Vigus. AutoAspect: Automatic annotation of tense and aspect for uniform meaning representations. In Claire Bonial and Nianwen Xue, editors, *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 36–45, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.law-1.4. URL https://aclanthology.org/2021.law-1.4.

Eleni Metheniti, Tim Van De Cruys, and Nabil Hathout. About time: Do transformers learn temporal verbal aspect? In Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus, editors, *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 88–101, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.cmcl-1.10. URL https://aclanthology.org/2022.cmcl-1.10.

Anisia Katinskaia and Roman Yangarber. Probing the category of verbal aspect in transformer language models, 2024.

Carlota S. Smith. *Modes of Discourse: The Local Structure of Texts*. Cambridge Studies in Linguistics. Cambridge University Press, 2003.

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.

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.

Tony McEnery and Andrew Wilson. *Corpus Linguistics: An Introduction*. Edinburgh University Press, 2001. ISBN 9780748611652. URL http://www.jstor.org/stable/10.3366/j.ctvxcrjmp.

Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages, 2024.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1141. URL https://aclanthology.org/P16-1141.

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.

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.

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.

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.

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.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models, 2023.

Cho-Jui Hsieh, Si Si, Felix X. Yu, and Inderjit S. Dhillon. Automatic engineering of long prompts, 2023.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

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.