

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

Zusammenfassung auf Deutsch

Contents

List of Figures	VI
List of Tables	VII
1 Introduction	1
2 Aspectology: An Introduction	3
2.1 A (short) phenomenology of aspect	3
2.1.1 Lexical and grammatical aspect	4
2.1.2 Lexical aspect parameters	5
2.1.3 Grammatical aspect parameters	7
2.1.4 Boundedness	8
2.1.5 Some attempts at event classification	9
2.2 Aspect in Slavic languages	12
2.3 Aspectual ambiguity	14
2.3.1 Contexts frequently triggering aspectual ambiguity in English	15
2.3.2 Coercion or underspecification?	17
3 Related Work	18
3.1 Aspect classification and ambiguity	18
3.1.1 (L)LMs and aspect	20
3.2 Available datasets	21
3.3 Related areas of work	21
4 Methods	23
4.1 Aspect classification schema	23
4.1.1 UMR	23
4.2 Research questions	26
4.3 Project outline	27
4.4 Use of LMs as sources of linguistic knowledge: Reversing the NLP pipeline	28

Contents

5	Experiments and Results	31
5.1	Dataset creation	31
5.1.1	Dataset annotation with LLMs	31
5.1.2	Manual ambiguity annotation	36
5.1.3	Larger dataset	38
5.2	Aspect classification	38
5.2.1	Smaller LM fine-tuning	38
5.2.1.1	Aspect latent space	39
5.2.2	Multilingual BERT fine-tuning	39
5.2.2.1	Telicity classification of motion verbs	41
5.2.2.2	A look at Russian verbal prefixes	41
5.3	Aspectual ambiguity	44
5.3.1	Sentence-level ambiguity	44
5.3.2	Verb-level ambiguity (coercion / underspecification?)	45
5.3.3	Language-level ambiguity: Cross-linguistic comparison	45
6	Discussion	50
7	Conclusion	51
A	Appendix	52
A.1	Experiments with ChatGPT	52
A.2	Temporal reasoning	52
A.3	Long LLM prompt	52
A.4	Annotation guidelines	52
A.5	Language-level entropy	55
	Bibliography	57

List of Figures

1	Visual representation of time, situation and event, along with tense, lexical and grammatical aspect	4
2	Comrie's classification of aspectual oppositions	7
3	[Moens and Steedman, 1988] event types	13
4	Semantic map of common aspectual ambiguities	19
5	UMR aspect classification lattice [Jens Van Gysel and Xue, 2022]	25
6	UMR graph of the sentence "200 dead, 1,500 feared missing in Philippines landslide." in PENMAN notation.	34
7	BERT aspect latent space	40
8	Look at this graph	43
9	Look at this graph	48
10	DESCRIPTION. The grey bars represent the mean entropy of each language family.	49

List of Tables

1	Representation of (12),(13),(14) and (15) with regards to the reference time referred to by the utterance and inherent run-time of the event. Concept adapted from van Hout [2016].	10
2	Classification of Vendlerian event types by binary aspectual parameters [Smith, 1991].	11
3	Some example verb phrases given in Vendler [1957] for the classification of events.	12
4	Some common contexts triggering aspectual ambiguity in English.	16
5	(Approximate) Comparison of aspectual classes. Adapted and extended from Egg, Prepens, and Roberts [2019].	25
6	Aspect classes of annotated verbal events in the UMR dataset.	33
7	F1 score on test set from fine-tuning Llama 2 7B and 3 8B on different types of prompt after 1000 training steps.	35
8	Results on test set from fine-tuning Llama 2 on UMR aspect classes without upsampling.	35
9	Results on test set from fine-tuning Llama 2 on UMR aspect classes with upsampling.	36
10	Analysis of annotated data.	37
11	Model performance on English and French test sets after training on English training set.	41
12	Model output of 3 example Russian sentences with motion verbs.	41
13	CHANGE THIS	47
14	Language-level Entropy on TED2013 dataset	55
15	Language-level Entropy on TED2020 dataset	56

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 field 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 made in *Aspect* [Comrie, 1976] between the deictic nature of tense and the focus on the "internal temporal constituency of a situation" of aspect. In other words: "tense¹ relates the time of the situation referred to, to some other time, usually to the moment of speaking", and thus by relating the time of the situation to the time of the utterance it is deictic [Comrie, 1976]. Aspect on the other hand gives a *situation-internal* description of the events in that situation, such 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].

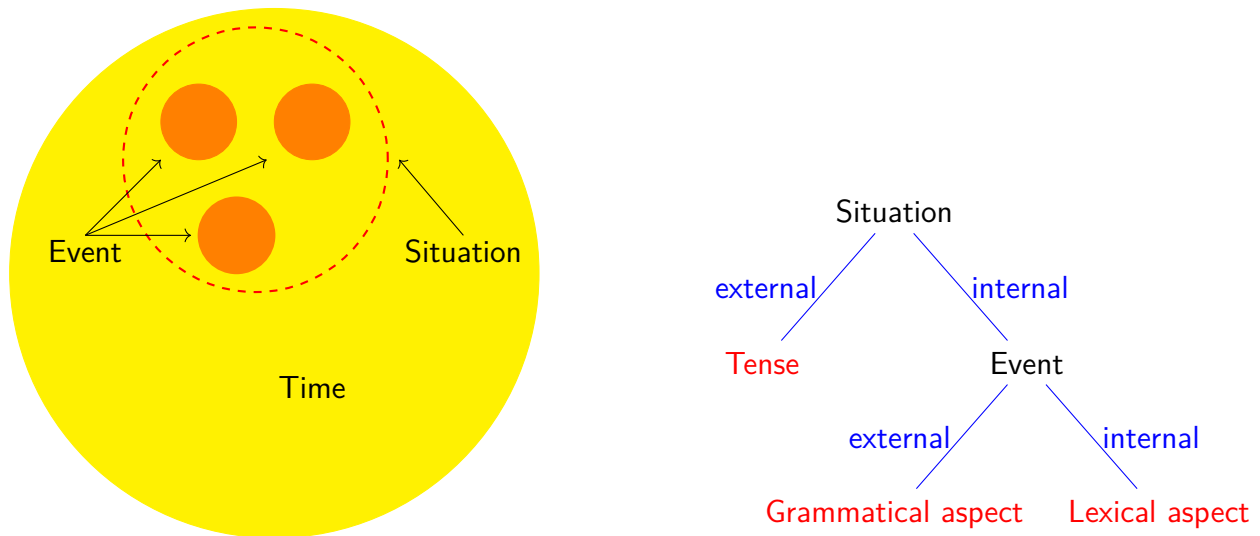


Figure 1: A visual representation of the relationship between time, situation and event (left), along with the categorisation of the three temporal phenomena discussed (right). Here each situation is a (non-empty) set of events ($\text{situation} \subseteq \text{events}$), and events (and by derivation situations) are anchored in time. In this (very simple) model, timeless generic 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, aspect³ 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 in a situation context, such as its habituality or ongoing nature, which is not an inherent feature of the events itself but can be seen rather as

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.

an external lens imposed on a (usually) verbal phrase.⁴ In English, one example of this lens is the progressive, which is formed by the verb *be* + *gerund*, as in 1.

- (1) Sue was running.

Figure 1 provides a visual representation of the concepts introduced in this section for clarity.

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 for reasons of pragmatism I will not strictly uphold the theoretical differentiation between these two in the experimental part of this study.

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 telicity⁵ is whether the verb phrase admits

4 To simplify matters, in this study I will focus on verbal events, though others such as Van Gysel et al. [2021] take a broader definition of event, including nominal events.

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 takes a different definition, defining 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 important parameter of lexical aspect 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.

The counterpart of states, *dynamic* events, in the other hand, are used to describe situations that change or where action is present.

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])

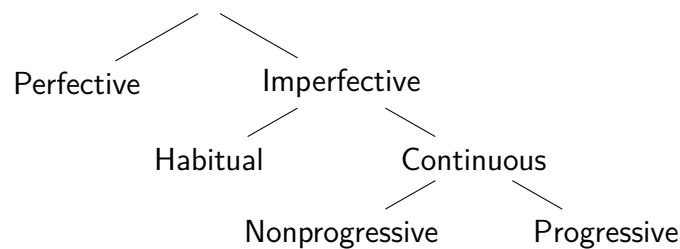


Figure 2: Comrie's classification of aspectual oppositions. Reproduced from Comrie [1976].

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,

(Im)Perfectivity

The main aspectual distinction made by Slavic languages and, as can be seen in figure 2, one of the main distinctions made in theoretical aspectology generally is that between perfective and imperfective. 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.

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 2.1.3.

- (9) Cherry trees bloom in April.

While usually still coming under the broad umbrella of aspectual phenomena,⁷ habituality is generally considered to be located at a different level to other aspect features. Dahl [1985] notes that the prevalence of habituels 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 using periphrastic means.

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 10 has no temporal bound on it (i.e. it is unclear when or if the event was ended), in 11 it 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.

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

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

Boundedness must be distinguished from telicity⁸ in order to avoid the so-called 'Imperfective Paradox' highlighted by Dowty [2012], here verbalised by Zucchi [2020]:

⁷ 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".

⁸ 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.

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 (12) entails (13), but (14) doesn't entail (15) in the examples below:

- (12) The man was running.
- (13) The man ran.
- (14) The man was building a house.
- (15) 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,⁹ and this further serves to show the dangers of 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.

2.1.5 Some attempts at event classification

Vendler [1957]

It was the seminal paper *Verbs and Times* [Vendler, 1957] of philosopher Zeno Vendler which initiated the discussion on inner aspect in the linguistic tradition.¹⁰ Vendler begins his discussion in the paper with the following premise:

⁹ I.e. by definition of telicity as whether an event 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 (12) and (13) are bounded, only (14) and (15) 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.

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.

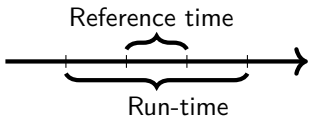
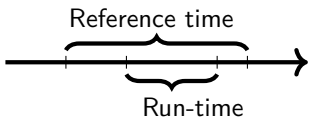
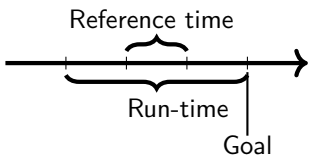
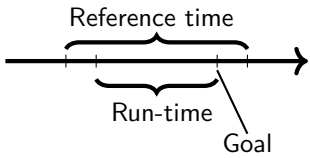
Sentence	Representation	Telicity	Boundedness
The man was running.	 <p>A horizontal timeline with an arrow pointing right. A bracket labeled 'Reference time' spans a portion of the timeline. Inside this bracket, another bracket labeled 'Run-time' spans a sub-interval. The 'Run-time' interval is unbounded, extending to the right edge of the 'Reference time' interval.</p>	atelic	unbounded
The man ran.	 <p>A horizontal timeline with an arrow pointing right. A bracket labeled 'Reference time' spans a portion of the timeline. Inside this bracket, another bracket labeled 'Run-time' spans a sub-interval. The 'Run-time' interval is bounded, ending before the right edge of the 'Reference time' interval.</p>	atelic	bounded
The man was building a house.	 <p>A horizontal timeline with an arrow pointing right. A bracket labeled 'Reference time' spans a portion of the timeline. Inside this bracket, another bracket labeled 'Run-time' spans a sub-interval. The 'Run-time' interval is unbounded, extending to the right edge of the 'Reference time' interval. A vertical line segment labeled 'Goal' points to the end of the 'Run-time' interval.</p>	telic	unbounded
The man built a house.	 <p>A horizontal timeline with an arrow pointing right. A bracket labeled 'Reference time' spans a portion of the timeline. Inside this bracket, another bracket labeled 'Run-time' spans a sub-interval. The 'Run-time' interval is bounded, ending before the right edge of the 'Reference time' interval. A vertical line segment labeled 'Goal' points to the end of the 'Run-time' interval.</p>	telic	bounded

Table 1: Representation of (12),(13),(14) and (15) with regards to the reference time referred to by the utterance and inherent run-time of the event. Concept adapted from van Hout [2016].

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 3), 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:

- **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 semelfactives¹²)

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].

11 The issue of anglocentrism is one which has plagued many a linguistic theory throughout the years, with work such as Chomsky's generative grammar [Chomsky, 1965] or the speech act theory paradigm developed by Searle, Austin, and Grice [Searle, 1969, Austin, 1962, Grice, 1975] being criticised for their too heavy focus on English and anglophone norms, and hence the lack of applicability to other languages [Levisen, 2019].

12 A semelfactive is a type of verb that denotes an action that is typically instantaneous and easily repeatable since the event ends returning to its initial state. Examples include "sneeze," "blink", or "knock" [Smith, 1991, Filip, 2012]

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.

Vendler states in his introduction that verbs "presuppose" certain time schemata and hence assigns a category to each verb (as in table 3). However it must also be stated that the true profile of a verb phrase such as those listed in table 3 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 table 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.

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 or 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 3 provides a visualisation of these different event and state types.

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].¹³ Verbs in

¹³ Tomelleri [2010] also notes that Georgian and Ossetian exhibit some similar properties with preverbs (mostly of spatial origin) used to denote aspectual meaning.

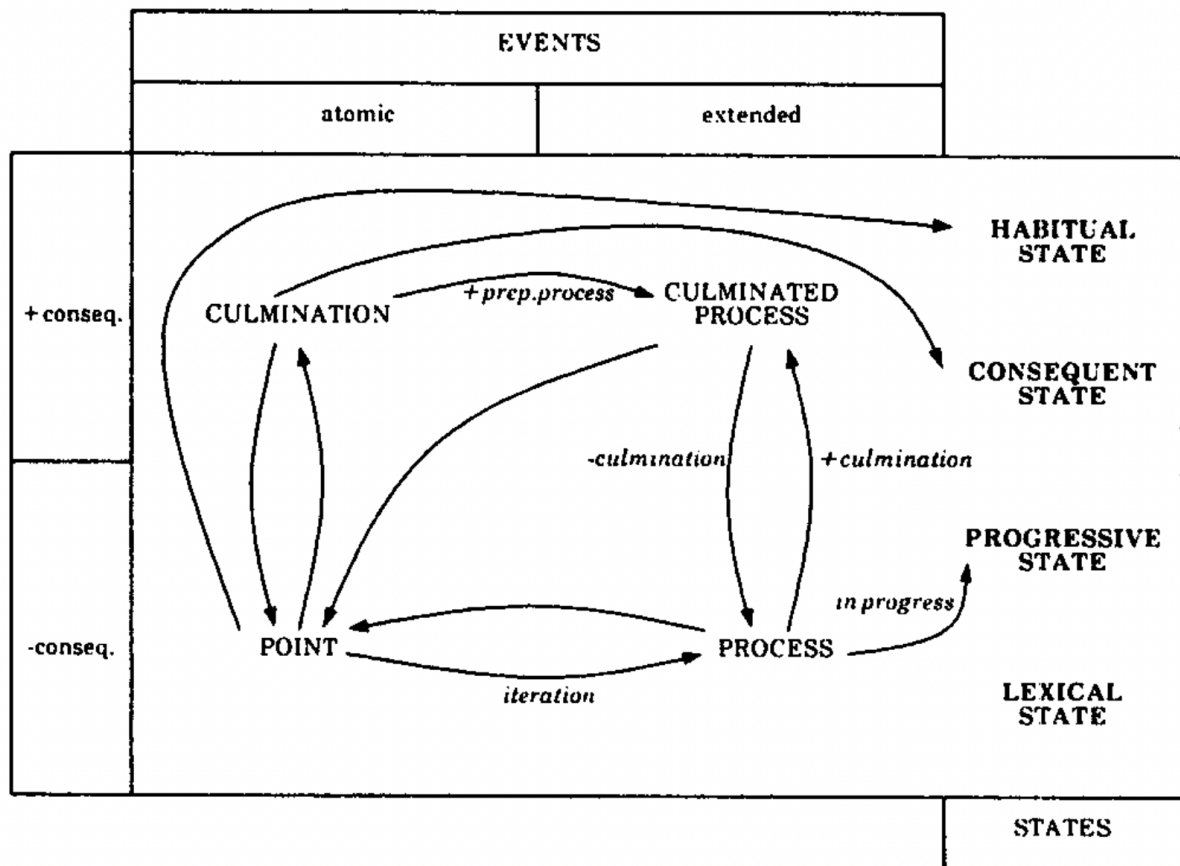


Figure 3: A visualisation of event and state types with their corresponding transitions [Moens and Steedman, 1988]

Slavic languages, with few exceptions,¹⁴ 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 ètu knig-u.
I read.PERF-FUT this book-ACC
'I will read this book.'
- (17) Ja budu čita-t' ètu knig-u.
I will read.IMPF-INF this book-ACC
'I will read (be reading) this book.'

¹⁴ 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])

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 sentence using the imperfective form denotes that the event has since been undone, although the window-opening event was perfective [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?'

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 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 (in this case 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.¹⁵ 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 used by someone narrating a story orally as in 22.

- (21) He walks to school.

- (22) So 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 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 an 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 are ambiguous with respect to their aspectual reading.

¹⁵ 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*, which is also used as the copula with the past participle.

One example was already introduced in 20 with the identical encoding of the past simple passive and a copular clause with a past participle adjective. Some other examples are semelfactives, which can often be interpreted as being either a repeated or a single event, or motion verbs with a certain set of prepositions which are ambiguous with regard to the telicity of the resulting verb phrase (such as "through", "across" etc.). It is also the case that the past and future simple are sometimes ambiguous in English with regard to their outer aspect (i.e. whether "cover" in the sentence "They covered their faces" refers to the outcome of the action or the action itself). Table 4 summarises these cases.

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].
motion verbs with ambiguous prepositions	[telic/atelic]	Anna walked through the park [then turned left/for hours].
past simple / future simple	[holistic/ongoing]	The rocks fell down [yesterday/slowly].

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

Since, as mentioned above in 2.1.3, 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,¹⁶ 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 4. 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.

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 "po-fective").

¹⁶ This is supported by the findings of the manual annotation (see 5.1.2), where 24.8-42.7%, depending on the annotator, of sentence-verb pairs had a possible habitual interpretation.

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

While Slavic languages do not directly encode dynamicity, telicity, or iterativity, the choice between perfective and imperfective aspects is strongly influenced by boundedness and whether an event is a single occurrence or repeated [Wiemer and Seržant, 2017]. Both aspect parameters express two distinct readings, meaning speakers are forced to choose a particular aspectual type. 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 (such as temporal adverbs or 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]. However, some studies from psycholinguistics, such as Lukasek et al. [2017] CHECK IF THIS IS RIGHT, have suggested the presence of underspecification in certain cases. 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

Despite significant efforts to model lexical and grammatical aspects in computational linguistics, this area has received comparatively very little attention from the natural language processing (NLP)¹ part of the field [Friedrich et al., 2023]. This is presumably due to the fact that, on the one hand, it is a high-level semantic task, whose relevance to downstream applications is perhaps not as immediately obvious as other similarly complex tasks, and on the other hand, it is a complex phenomenon which is sometimes hard to capture accurately and faithfully. 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 Aspect classification and ambiguity

There have been numerous works in computational linguistics over the last 20 years, using statistical and rule-based techniques to look at aspect. Some of the more prominent examples are listed here:

- Siegel and McKeown [2000] develop a rules-based aspect classification of verbs using certain linguistic indicators as features in a corpus to train decision tree, genetic programming and logistic regression models. 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 contextual reading, 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 to gain linguistic insights, such as the fact that stative verbs are characterised by their high verb frequency.
- Friedrich and Palmer [2014] train a 3-way random forest classifier (predicting DYNAMIC, STATIVE or BOTH), crucially for verbs *in context*. They note that there are many cases where the verbs themselves have an ambiguous aspectual reading which is resolved by the

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

context, but that equally there are situations where this ambiguity persists. However, they do not go on to further investigate this fact or the situations where this occurs, or to look at other types of aspectual ambiguity.

- Egg et al. [2019] train several logistic regression classifiers to predict the aspectual class of German verbs in context using a more fine-grained class system, achieving 71.2% accuracy on their 6-way classification. They note the phenomenon of *coercion* and aim to annotate the aspectual class of the element before coercion (therefore in these cases excluding this coercing context). They also stress the importance of dealing with aspectual ambiguity and give the example of the systematic ambiguity of so-called *degree achievements* (from Kennedy and Levin [2008]) such as "den Weg kehren" (Eng. *sweep the path*), which can have an unbounded reading or an extended change reading.
- Croft et al. [2016] propose an annotation scheme for the aspectual structure of events, intended to be integrated into a larger event description framework: Richer Event Descriptions (RED) [Styler et al., 2014]. In doing so, they group together aspectual classes that often get mixed up with each other and use these pairings to create a semantic map (see 4). These more coarse-grained classes are then used for a second level of annotation, where (if possible) these ambiguities are resolved into one of the two finer-grained categories in the pair. However, this is a different type of ambiguity as the one dealt with in this study, where the focus is less on semantic similarity (and the ensuing annotator uncertainty), but more on *true* ambiguity (i.e. where several interpretations are equally possible and valid).

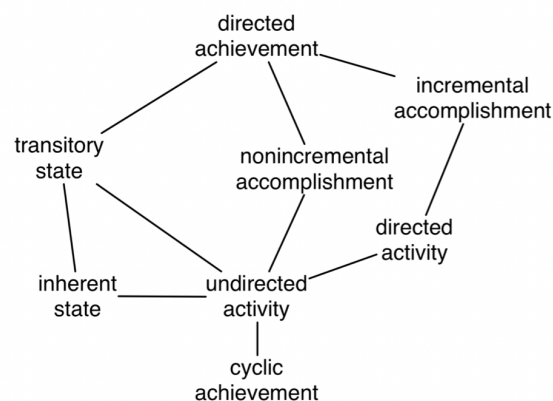


Figure 4: Croft et al.'s semantic map of common aspectual ambiguities Croft et al. [2016]

- Chen et al. [2021] build a rules-based system (or what they call "annotation tool") to predict UMR aspect class, by converting the aspect annotation guidelines into computable

steps. Out of 235 gold events distributed across four gold files, the model missed 56 events, successfully identifying 179, and among these 179 identified events, the model accurately labeled 112, resulting in an accuracy rate of 62.57%. This is the only work so far to my knowledge using the same aspect classification schema as this study, UMR (see 4.1.1).

- One other particularly interesting example is Friedrich and Gateva [2017], who use an English-Czech parallel corpus to leverage 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 for their linear regression model. They find that incorporating this "silver-standard" data into their training leads to significant increase in the F_1 score, especially when predicting *ATELIC* examples.

3.1.1 (L)LMs and aspect

There have also been attempts to use more modern NLP techniques such as language models for aspect classification and studies probing their aspectual knowledge.

- Metheniti et al. [2022] carry out a range of experiments with several transformer models (such as BERT, RoBERTa and XLNet) to try and understand to what extent they encode aspectual knowledge. They fine-tune the models on separate classification tasks for telicity and duration, as well as carrying out a qualitative analysis, and come to the conclusion that transformer models *do* understand aspectual phenomena to a sufficient extent, and they find that this knowledge relatively spread across the layers. However, they do not further probe how this information is encoded or which features the model uses to make its decision.
- Conversely, however, Katinskaia and Yangarber [2024] find that in Russian BERT models [Kuratov and Arkhipov, 2019] aspect information is stored in layers higher up. Katinskaia and Yangarber [2024] also find that their model has high uncertainty when predicting aspect in "alternative contexts" (i.e. where both the imperfective and the perfective verb is possible, depending on the implicit context). On the one hand, this is an encouraging find, as it suggests a correlation between machine behaviour and human behaviour, but on the other hand it is not particularly surprising, since "alternative contexts" are by definition contexts

with no tokens containing aspectual information, which are usually used by the model to make a decision.

- Rezaee et al. [2021] look at using generative and discriminative transformers for situation entity classification, which is, however, a slightly different task (see 3.3).

Friedrich et al. [2023] notes however that there have still been very little (if any) work looking at the probing or explainability of LMs for this task.

3.2 Available datasets

Although not small in number or size, the available datasets are relatively sparse since there is a very wide range of aspectual phenomena (/schemata) which can be annotated for. For example while some are annotated for aspectual phenomena such as habituality [Mathew and Katz, 2009, Ikuta et al., 2014] and others purely for telicity [Friedrich and Gateva, 2017, Kober et al., 2020, Metheniti et al., 2022], some (generally smaller) datasets are also annotated for a more general schema such as Vendler’s classification [Zarcone and Lenci, 2008, Hermes et al., 2015, Zellers and Choi, 2017] or Universal Dependencies (UD) aspect classes [Kondratyuk and Straka, 2019]. So while there is a relatively wide range of datasets, they are all annotated for slightly different tasks and with different guidelines, which makes matters a little more challenging.

For aspect ambiguity the only currently existing annotated dataset of which I am aware is Friedrich and Palmer [2014], which is however only annotated for ambiguity between DYNAMIC and STATIVE. I therefore had to create my own (see 5.1.2).

3.3 Related areas of work

Situation entities

Situation entity classification is the task of identifying different types of situations, which exist at a clausal level. The task 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 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 incorporating concepts from linguistic aspect, the focus is more on a discourse level and the classification schemata are thus different accordingly. Some works in this area include Palmer et al. [2007], Friedrich et al. [2016], Becker et al. [2017], Friedrich [2017] and Dai and Huang [2018].

Applications: Temporal reasoning

An application of the tasks looked at in this chapter and a related area of work is the more general area of temporal reasoning. While this is evidently closer to the area of situation entities introduced in 3.3, the more linguistically-informed side of aspect also has a role to play. For example, it is clear that the two following sentence, differing only in their aspect, refer to two very different events.

(25) The driver of this vehicle drinks wine.

(26) The driver of this vehicle is drinking wine.

Costa and Branco [2012] analyse the importance of aspect for temporal relation classification, and Derczynski and Gaizauskas [2015] employ a tense and aspect framework to sequence events in text. Allen and UzZaman [2012] explores various aspects of tense and aspect in event processing. Similarly, Kamp and Reyle [1993] discusses the integration of tense and aspect information within discourse representation theory (DRT). Furthermore, Kober et al. [2019] develop a dataset to evaluate the temporal and aspectual entailment capabilities of models.

More broadly, it is clear that understanding verbal aspect is crucial for a variety of more general downstream tasks, including machine translation and question answering systems. See Friedrich et al. [2023] for a more in-depth look at the applications of this area.

4 Methods

This chapter will give an overview of the main research questions for this project and how I intend to investigate them.

4.1 Aspect classification schema

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. The classification schema I decided on was the schema designed as part of Uniform Meaning Representation (UMR) [Van Gysel et al., 2021], for several reasons. On the one hand, the schema 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, UMR provides a lattice for annotation (see 5) meaning the level of the annotation classes can be adjusted to the needs of the annotation context, thus giving more flexibility. Furthermore, the schema is part of a larger framework, and hence a classification system using these labels has a practical application too, since, if it works well, it can be later integrated into a larger UMR parsing system.

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].

ADD STUFF HERE ABOUT HOW UMR WORKS GENERALLY - SENTENCES AS GRAPHS (NODES + EDGES) - SENTENCE LEVEL and DOC LEVEL - EXAMPLE GRAPH

Aspect in UMR

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

- **state** - stative events, i.e. no change takes place over the course of the event
- **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 Time¹ (DCT)
- **endeavour** - a process that ends without reaching its result state (i.e., termination)
- **performance** - a process that ends reaching its result state

How these relate to other aspectual classification schemata can be seen 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 (in almost all cases) a paradigmatic example of an outer aspect, rather than one inherent in the verb event itself. Therefore, as can be seen in figure 5, 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 is scalability, which entails learnability for annotators. As already mentioned, 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. Furthermore, this conflation of two phenomena can also be an advantage, since it shows the relationship between classes of each type

¹ Since UMR was mostly designed for written text, they use the creation time of the document as a reference point. In spoken language this could be equated with time of utterance, though slightly different, since speech is an extended act, whereas document publication is punctual.

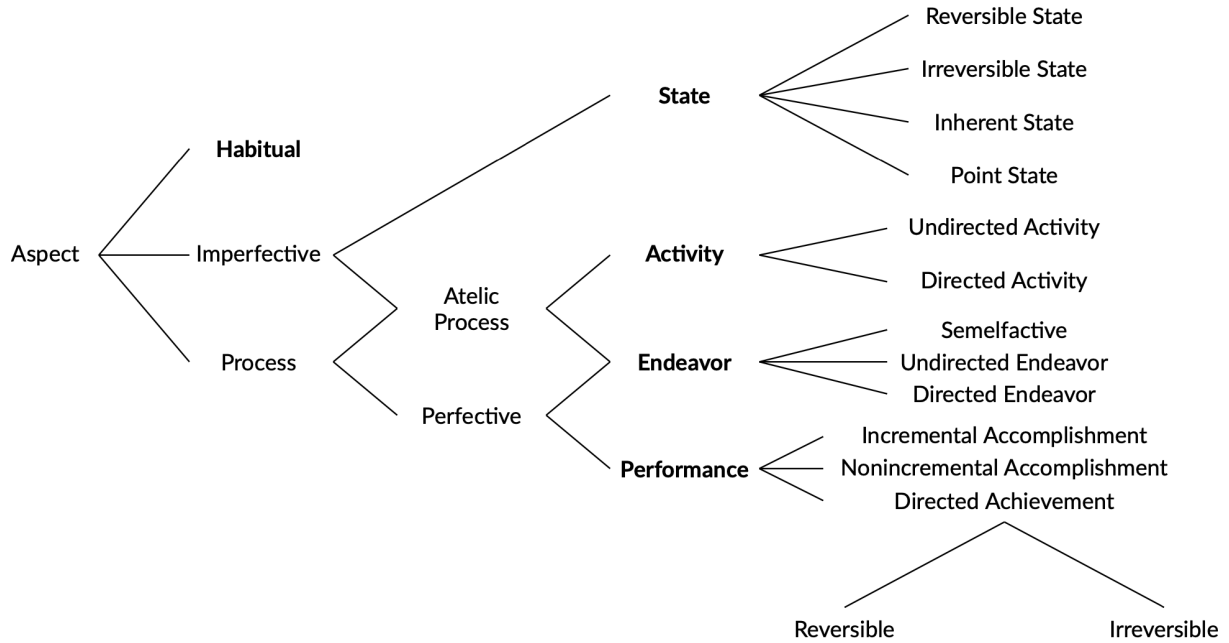


Figure 5: UMR aspect classification lattice [Jens Van Gysel and Xue, 2022]

Vendler [1957]	Moens and Steedman [1988]	Egg [2005]	Egg, Prepens, and Roberts [2019]		Van Gysel et al. [2021]
state	state (habitual state)	stative predicate (CHECK THIS)	stative		state
activity	process	process predicate	dynamic	unbounded	activity
accomplishment	culminated process	intergressive predicate change predicate		bounded	endeavour
achievement	point	intergressive predicate			performance
	culmination	change predicate			endeavour

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

(i.e. that *irreversible states* are imperfective and *directed achievements* perfective etc.), subsuming them all into *one* semantic parameter space concerning aspect.

Difficulties with aspect in UMR

The UMR aspect system does come with some difficulties. Firstly, the classification scheme is notably different from those that came before it, since it is neither based on Vendler’s classification, nor on typical binary aspect parameters, but rather seeks to create a typologically-informed aspect classification, taking into account explicit aspect encoding systems from a variety of languages. For example, the highest distinction made by Comrie [1976] was between PERFECTIVE and IMPERFEC-

TIVE, while in UMR there are classes such as ENDEAVOR which can be both. The independence from the classic aspect parameters such as TELICITY or DURATIVITY also means the datasets annotated for one of these parameters cannot be easily used.

Secondly, UMR and its annotation guidelines are relatively new, hence sometimes unclear and are subject to change. This means that there are currently some minor inconsistencies between the annotated data available and the guidelines are sometimes unclear or problematic. I hope that as the project grows and expands to other languages, the annotation guidelines will be revised and refined and the inconsistencies fewer. At time of writing,² the version 1.0 of the annotation guidelines has not yet been released.

4.2 Research questions

This section describes the research questions which the following study aims to answer.

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

The first question I aim 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, to the best of my knowledge, that a *large* language model³ has been used for this task. If this model works well, it will be possible to use it to look at some linguistic questions, such as whether particular prefixes in 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. At this stage it is also an interesting question to examine how verb phrases are classified without context. This could shed some

² June 2024

³ Defined as having > 1 million parameters, compared to BERT's 345 million.

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 languages 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 lot more possibilities for analysis than the dataset used for fine-tuning the larger model. For instance, this larger dataset can be used in a 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.

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,⁴ 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 LMs as sources of linguistic knowledge: Reversing the NLP pipeline

One of the aims of this thesis is to highlight how language models can possibly be useful as sources of linguistic knowledge too. This idea is by no means new, and yet in the wider linguistics community outside computational linguistics, these tools seem to be little used.

⁴ Recall that Friedrich and Palmer [2014] only annotate ambiguity in the DYNAMIC/STATIVE parameter.

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 generalise 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.

While in the past, the NLP pipeline used to be made up of a host of linguistically informed steps (such as lemmatisation, dependency parsing and POS tagging), this changed with the advent of transformer models, which, some have argued, "rediscover" the classical NLP pipeline, albeit implicitly [Tenney et al., 2019]. Since these end-to-end models no longer require explicit linguistic encoding, one could argue that the importance of linguistic knowledge in NLP has dramatically decreased [Church and Liberman, 2021], though it is undoubtedly true that linguistics is very useful for some areas of NLP which are currently less deep-learning inclined, such as low-resource NLP or meaning representation. Whether or not linguistics has a future in NLP, in this study I wish to look at the other direction of information flow and show that the benefits of collaboration between linguistics and NLP are not a one-way street.

When can (L)LMs *not* help us?

However, 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 finding datasets of the necessary order of magnitude 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 and validity of some of the experiments described in this thesis to less well-resourced languages.

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 with added dimensions such as Hamilton et al. [2016] could still be useful.

5 Experiments and Results

5.1 Dataset creation

In order to carry out further computational analysis of the topic it is helpful to have an annotated dataset. Since no extensive datasets exist for UMR, I had to create my own. 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 LLMs 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].

5.1.1 Dataset annotation with LLMs

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. However, this differs from my situation, where 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

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, these models come with several issues, as Törnberg [2024] notes: it is unknown what the training data for the model is, leading to problems with transparency, and ChatGPT models have been shown to evolve over time, meaning reproducibility is hindered. For these reasons, and also in order to host the model locally for better control (rather than using an API), I chose to use a different model. Meta's

Llama 2 model [Touvron et al., 2023] seemed to offer a good balance between the apparent current trade-off between performance and scientific good practice, performing comparatively well in previous similar studies [Yuan et al., 2023, Li et al., 2023].

Due to hardware constraints, I had to use a technique to improve the efficiency of the model fine-tuning process since the smallest Llama model is 7 billion parameters. In this case I chose to use "parameter efficient fine-tuning" (PEFT) [Mangrulkar et al., 2022], which leverages the insight that LM fine-tuning usually only updates parameters at the end of the network. 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, 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 thesis, Llama 3 was released to the public [Meta LLaMA Team, 2024], promising to have better reasoning skills and be better at following instructions effectively, so I therefore also tried and compared this.

Training data

As the only currently existing data containing UMR aspect classes, I used the example UMR dataset¹ provided by the creators of UMR. It contains a total of 2022 sentences in 6 languages (Arapaho, Chinese, English, Kukama, Navajo and Sanapana) annotated according to the UMR schema. The texts are a mix of news, oral narration and interviews.

In order to extract the verbs in the sentences I used Stanford NLP's Stanza POS tagger [Qi et al., 2020], 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 habituais 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

¹ The UMR v1.0 dataset can be found here.

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.

were either found in existing online datasets and simply labelled with a UMR aspect class by hand, or they were both manually composed and then labelled.

Törnberg [2024] points out that, since it is unknown exactly which training data was used to train many LMs, one should exercise caution when evaluating their performance on a test set, since the model may have seen the test data before during pre-training. In this case, while it is impossible to rule out that the UMR dataset was used for Llama 2 or Llama 3 pre-training, 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) where these refer to an event.

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 my analysis the :aspect parameter occurred with a verb in the source sentence 74.8% of the time, meaning 282 of the 377 events 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 developing, testing and then improving the annotation guidelines together with LLM prompt until the desired


```

(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 6: UMR graph of the sentence "200 dead, 1,500 feared missing in Philippines landslide." in PENMAN notation.

quality is achieved. Since, as already mentioned, the annotation guidelines are already defined, the task consisted solely in finding an effective prompt. One of the issues to balance here, also underlined by Törnberg [2024], is the balance between length and detail, since a certain amount of information is necessary to carry out the task, yet LLMs seem to suffer when given a prompt which is too long (which is what I also found in my experiments: see 7). I therefore decided on an adapted version of the succinct class descriptions provided in Van Gysel et al. [2021], rather than the more extensive descriptions from the annotation guidelines.

The model was given the following instruction for each datapoint:

The State value corresponds to stative events: no change occurs during the event. The Habitual value is annotated on events that occur regularly. The Activity value indicates an event has not necessarily ended and may be ongoing at Document Creation Time. Endeavor is used for processes that end without reaching completion (i.e., termination), whereas Performance is used for processes that reach a completed result state.

followed by the following question:

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

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 after 1000 training steps.

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
ENDEAVOR	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.

The results on a hold-out test set show that without an explanation of the classes, the Llama 2 model fails, achieving roughly random accuracy, whereas the Llama 3 model performs significantly better but still well under the performance with the class descriptions.

I also experimented with a longer, more detailed prompt², and interestingly the results also show that the extended prompt also exhibited poorer performance. This could be due to the fact that the model has limited memory capacity and, since the prompt contained more redundant information, it "forgets" important parts of the description at the beginning.

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 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 thus "easier" than the natural data. Nevertheless, this result taken as a standalone one shows that the

² For the long version, see A.3.

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
ENDEAVOR	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.

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 within this classification framework, 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 a rather subjective process.³

The annotation was done by an English native-speaker with a background in philosophy, and me, also an English native speaker. 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. See A.4 for the exact annotation guidelines.

For example, the following sentence-verb pair was annotated by annotator 1 as belonging to `STATE` and `ACTIVITY` classes and hence has an ambiguous aspectual reading:

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

³ This is evidenced by the often low inter-annotator agreement. For example, Friedrich and Palmer [2014] found a Cohen’s observed unweighted κ of between 0.6 and 0.7 on a two-way classification task.

	Annotator 1		Annotator 2	
	Num.	%	Num.	%
Contains the gold-standard class	202	0.7063	203	0.7098
Ambiguous	71	0.248	122	0.248
Ambiguous excluding HABITUAL	41	0.1434	18	0.0629

Metric	Value
IAA	0.552
≥ 1 class in common	0.745
Ambiguity IAA	0.633

Table 10: Analysis of annotated data.

Manual annotation results

The results of the annotations can be seen in table 5.1.2.

It was interesting to note that the majority of occurrences of the HABITUAL class (30 out of 34 (88.2%) for annotator 1 and 104 out of 105 (99.0%) 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 5), where habituais are on the second-highest level of the tree.

Since I had two annotators who can choose between 1 and 5 classes for each datapoint, I used the following formula to calculate the inter annotator agreement (IAA):

$$IAA = \frac{1}{N} \sum_{i=1}^N \frac{|A_{i,1} \cap A_{i,2}|}{|A_{i,1} \cup A_{i,2}|}$$

where $A_{i,j}$ is the set of labels decided by annotator j for datapoint i . I also calculated what proportion of datapoints had at least one class shared by both annotators in the following way:

$$\geq 1 \text{ class in common} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{|A_{i,1} \cap A_{i,2}| \geq 1}$$

The vast majority (87.0%) of sentence-verb pairs marked with several verb classes by annotator 1 were also marked as ambiguous by annotator 2, however only 69.9% of those classified as ambiguous by the latter were seen as such by annotator 1. This indicates that annotator 2 had interpreted a broader definition of the classes, meaning there were more cases where several labels

were possible. However, this causes a problem when deciding which labels to use. One possibility would be to take the class as +AMBIGUOUS where both annotators agree on it being ambiguous, however this would lead to an even smaller dataset where the different class interpretations of both annotators are confounded, which could therefore perhaps harm performance. The other possibility is just to take the annotations of one of the annotators, and I settled for this option, taking those of annotator 1. This is because, since ambiguity is already hard to recognise and define, I wanted to ensure a more constricted definition of aspectual ambiguity, as seems statistically to be the case for annotator 1.

DO WE KEEP HABITUALS?????

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 clean and has enough examples to train the smaller model. It would also be an advantage if it is a word-aligned multilingual corpus, since this would make it possible to extract the corresponding verb in another language and assign it the label assigned to the English verb. Thus, we can test the performance of a multilingual model in other languages where there are no labelled data from the Llama model.

The requirement of the corpus being word-aligned narrows down the options considerably, and for ease of use I therefore chose one of the default corpora offered in NLTK [Bird et al., 2009]: COMTRANS, which primarily consists of news articles and legal documents.

5.2 Aspect classification

5.2.1 Smaller LM 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, as described

above, 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.

5.2.1.1 Aspect latent space

Figure 8 provides a visualisation of the [CLS] token embedding of verb-sentence pairs in the training set of a monolingual BERT model after fine-tuning, 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].

5.2.2 Multilingual BERT fine-tuning

Since the LLM annotator Llama 3 can only reliably 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 at inference stage. 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 several models in English using the COMTRANS dataset annotated by the fine-tuned Llama 3 model and tested them both on an English hold-out test set and on a French one. The results, shown in 11, demonstrate that the models were all able to perform inference on French aspect without ever having seen any French training data, but that there was a performance drop between the languages, which is to be expected. On all metrics both in English and French, the xlm-roberta-base model outperformed all other models by a significant amount, and I therefore chose to use this model in the following experiments.

Fine-tuned BERT aspect embedding space

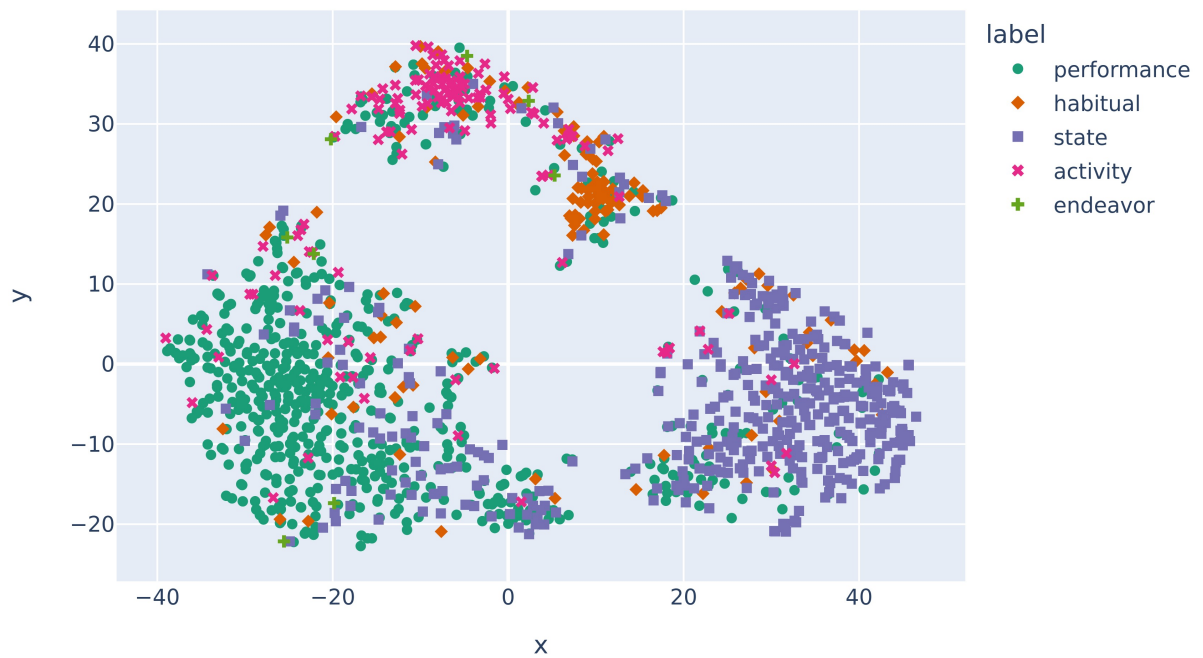


Figure 7: [CLS] embedding space of a BERT model fine-tuned on English verbs annotated for aspect in context, reduced to 2 dimensions by t-SNE.

Model	English		French	
	F1 (best)	Acc	F1	Acc
bert-base-multilingual-uncased	0.640	0.791	0.452	0.587
bert-base-multilingual-cased	0.647	0.790	0.450	0.572
xlm-roberta-base	0.665	0.813	0.522	0.651
xlm-roberta-large	0.647	0.797	0.447	0.629

Table 11: Model performance on English and French test sets after training on English training set.

5.2.2.1 Telicity classification of motion verbs

GERMAN? RUSSIAN? 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 28, the latter for one without 29.

(28) Ja *šel* v školu. (I was walking to school.)

(29) Ja *chodil* po parku. (I was walking around the park.)

Telic imperfective:	Ja <i>šel</i> v školu.	→	ENDEAVOUR	✓
Telic perfective:	Ja <i>pošel</i> v školu.	→	PERFORMANCE	✓
Atelic imperfective:	Ja <i>guljal</i> po parku.	→	ACTIVITY	✓
Atelic perfective:	Ja <i>poguljal</i> po parku.	→	ACTIVITY	?

Table 12: Model output of 3 example Russian sentences with motion verbs.

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

5.2.2.2 A look at Russian verbal prefixes

Another experiment I carried out with the fine-tuned multilingual model is to look at whether I can find empirical evidence that some verbal prefixes in Russian have a tendency towards

certain aspect classes due to their semantics, as has been suggested by theoretical works [Flier, 1975, Dickey, 2000], beyond the more trivially identified binary Slavic perfective/imperfective parameter.

In order to test the significance of my findings I use the statistical t-test, which is used to determine whether the mean of two groups is significantly different and by how much. The formula is the following:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where \bar{x}_1 and \bar{x}_2 are sample means, s_1^2 and s_2^2 are sample variances, and n_1 and n_2 are the sample sizes.

The probability of verbs with a prefix ($n = 3489; \sigma = 0.3298; \mu = 0.1425$) being classified as the STATE class was lower than verbs without a prefix ($n = 2473; \sigma = 0.4570; \mu = 0.3987$) by a statistically significant amount ($t = -25.1392; p < 1.3e - 132$). This was also the case for the ACTIVITY class, albeit by a much lesser degree ($t = -0.2841; p < 7.5e - 2$). As can be seen in 8, the opposite is also true for the PERFORMANCE class ($t = 19.4150$), also with very high significance ($p < 1.9e - 081$). The differences between the HABITUAL and ENDEAVOUR classes was not significant. This corroborates previous work that morphologically more complex, and thus usually semantically more complex, verbs are more likely to refer to a telic event (cf. English phrasal verbs transitioning from stative to dynamic, such as *sit* and *sit up* or *see* and *see in/off*) [Iacobini, 2009, Cappelle, 2007, Walková, 2017, Filip, 2004].

We can also look at a more specific example, taking the two prefixes *pro-* and *pere-*. While both meaning "through", they both emphasise different parts of the action. While *pere-* simply implies that the "inceptive" and "terminal" limits of the domain being crossed have been traversed, *pro-* puts emphasis on the action inside the domain and on the duration of this traversal [Flier, 1975]. This theoretical hypothesis is validated by the results here, with verbs beginning with *pro-* exhibiting being classed significantly more often as ACTIVITY ($t = 3.2400; p < 0.0014$) and less often as PERFORMANCE ($t = -2.5284; p < 0.0121$).

These findings provide empirical support for the notion that the semantics of verbal prefixes in Russian influence their aspectual classification, corroborating theoretical claims about their role in aspectual distinctions beyond the binary perfective/imperfective parameter. It must be noted that

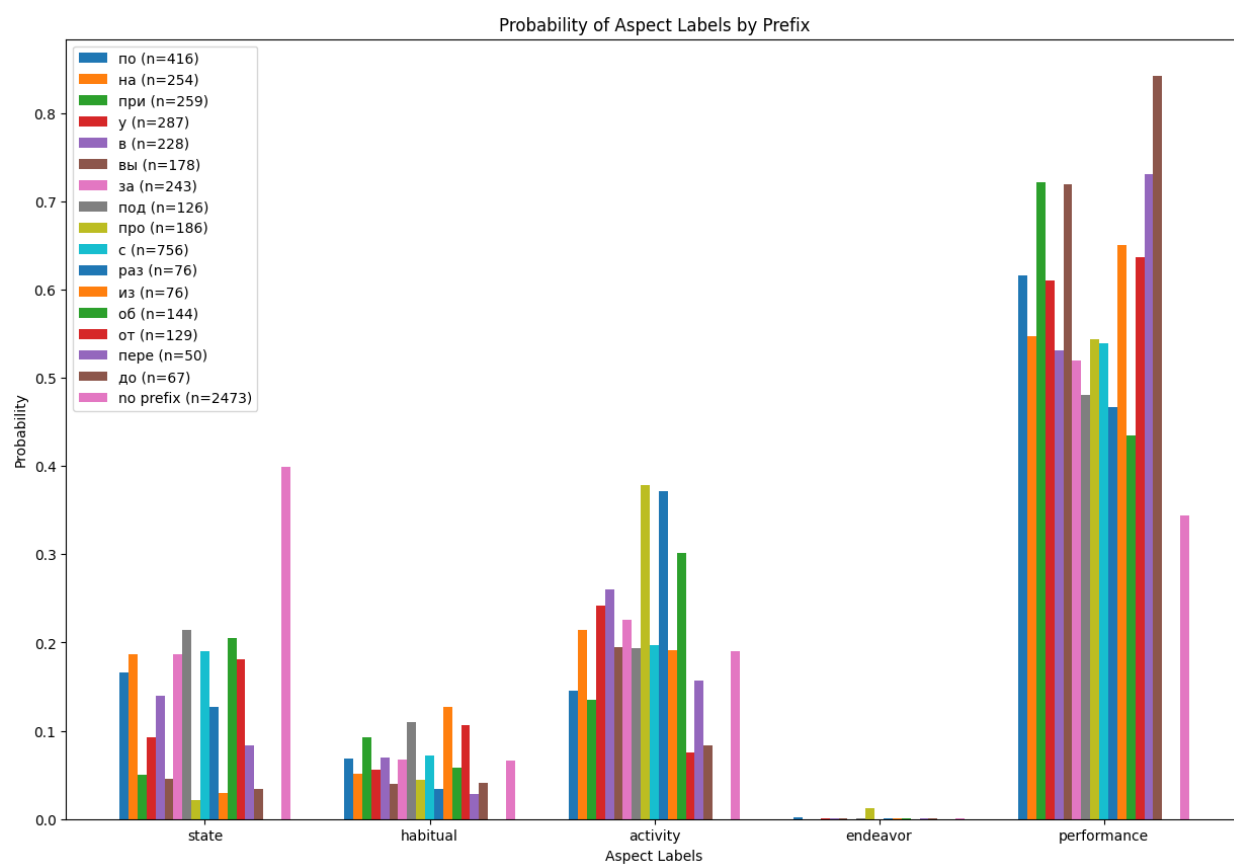


Figure 8: Look at this graph

this is a rather coarse-grained approach, which misses a lot of crucial information⁴, but the aim of this study is to discern and/or validate general tendencies in the behaviour of prefixes, rather than a more nuanced look at the verbs case by case, which lies more in the field of interest of theoretical linguistics.

5.3 Aspectual ambiguity

5.3.1 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 [Friedrich et al., 2023, ?] OTHER CITATIONS 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:

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

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:

⁴ For example that prefix groups can contain subgroups which act differently and cancel each other out on the macro scale, cf. Janssen and Borik [2012]

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

where:

- \bar{X}_1 is the mean of the continuous variable for the group where the binary variable is 1
- \bar{X}_0 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

I found that the model entropy value had a *medium*⁵ correlation of $\rho = 0.3465$ ($p < 0.0006$) with the upsampled ambiguity data derived from the human annotations. This is an encouraging result, especially the significance, however the correlation is not particularly high, and we should expect this value to be even lower for languages other than English, which sadly decreases the reliability of any conclusions taken from using the entropy value of this model. Furthermore, there was no correlation with the annotations of the fine-tuned LLM ($\rho = 0.0028; p < 0.85$), which is a disappointing result.

Nevertheless, while this correlation with human is probably too small to lead to truly meaningful results on the level of individual datapoints, with a large enough sample size we can still use this to find general tendencies on the macro level, such as at language level.

5.3.2 Verb-level ambiguity (coercion / underspecification?)

5.3.3 Language-level ambiguity: Cross-linguistic comparison

In order to compare between languages, it is first necessary to find a corpus available in all the languages I wish to compare, if not parallel then with at least similar texts (so called *comparable* corpora). To this end I decided to use TED2013 [Tiedemann, 2012] and TED2020 [Reimers and

5 According to Cohen's interpretation of Pearson's correlation [Cohen, 1988]

Gurevych, 2020], which are openly available corpora comprising of TED talks translated into different languages.

Since I am comparing verbal ambiguity, I first extracted the verbs from the sentences, as I did previously, and hence compiled a list of 10,000 verbs in context for each language. I then ran inference on these verbs alone (without context) and calculated the mean aspect entropy, defined in 1. TED2013 has 11 languages which are also supported by Stanza (for the verb extraction), and TED2020 has 32. The results for each language in TED2013 are shown in 9, and those from TED2020 are grouped by language family in 10. From both graphics it is clear that, with the exception of Greek and Chinese, Slavic languages have consistently lower language-level verbal aspect entropy values. This is to be expected considering the overt lexicomorphological marking of aspect discussed in 2.2.

The fact that Vietnamese had the highest entropy of the languages surveyed is unsurprising, given the fact that Vietnamese verbs do not mark for tense or aspect, but rather this information is given by the context.

Greek also has a perfective/imperfective system with two different stems [CHECK WORDING HERE](#).

Finnish - rich verb morphology (contrast to Vietnamese), similarly to Slavic languages

For the exact values, see A.5.

Do statistics test (t-test)!!

Verb		Entropy	PC	Verb		Entropy	PC
schlappte		0.00473	performance	schlenderte		0.05586	activity
schwebte		0.00712	activity	tauchte		0.05852	activity
trippelte		0.00784	activity	hetzte		0.06586	activity
schwankte		0.00836	activity	ritt		0.06784	habitual
schwamm		0.00837	activity	taumelte		0.07257	activity
pilgerte		0.00856	activity	trottete		0.08606	activity
trabte		0.00867	activity	wandelte		0.10565	activity
strömten		0.00877	activity	stolperte		0.11249	performance
joggte		0.00924	activity	ruderte		0.12711	activity
flanierte		0.00935	activity	rannte		0.13438	activity
tapste		0.00943	activity	tippelte		0.1351	performance
krabbelte		0.00964	activity	torkelte		0.13657	activity
watschelte		0.0097	performance	tigerte		0.16338	activity
latschte		0.00996	activity	skatete		0.17054	activity
rutschte		0.01028	activity	stieg		0.17735	endeavor
bummelte		0.01128	activity	humpelte		0.18405	activity
stromerten		0.01148	activity	radelte		0.2051	activity
hüpfte		0.01304	activity	ging		0.20655	activity
purzelte		0.01418	activity	stapfte		0.20949	endeavor
schlurfte		0.01443	activity	stapfte		0.20949	endeavor
wanderte		0.01623	activity	floh		0.26129	activity
kullerte		0.01686	activity	stolztierte		0.33834	activity
kroch		0.01697	activity	huschte		0.38448	habitual
hinkte		0.01769	activity	fuhr		0.39979	performance
schlich		0.01799	activity	raste		0.40537	activity
schweiften		0.018	activity	kletterte		0.44416	habitual
streunten		0.01959	activity	striefte		0.49204	activity
robbte		0.01969	activity	hastete		0.54362	activity
wieselte		0.0199	activity	stürmte		0.55842	activity
glitt		0.02027	activity	schrutt		0.5728	performance
flatterte		0.02123	activity	eierte		0.65506	activity
trampelte		0.02399	activity	segelte		0.67731	activity
trieb		0.0262	activity	sank		0.74246	performance
hoppelte		0.02664	performance	sauste		0.78209	activity
galoppierte		0.02702	activity	hopste		0.78673	endeavor
wankte		0.02801	activity	flog		0.84478	activity
spazierte		0.03137	activity	flitzte		0.85863	activity
marschierte		0.03802	activity	rollte		0.97166	endeavor
lief		0.03966	state	floss		0.99165	habitual
reiste		0.04599	activity	sprang		1.03842	endeavor
sprintete		0.05367	activity	eilte		1.27521	activity

Table 13: CHANGE THIS

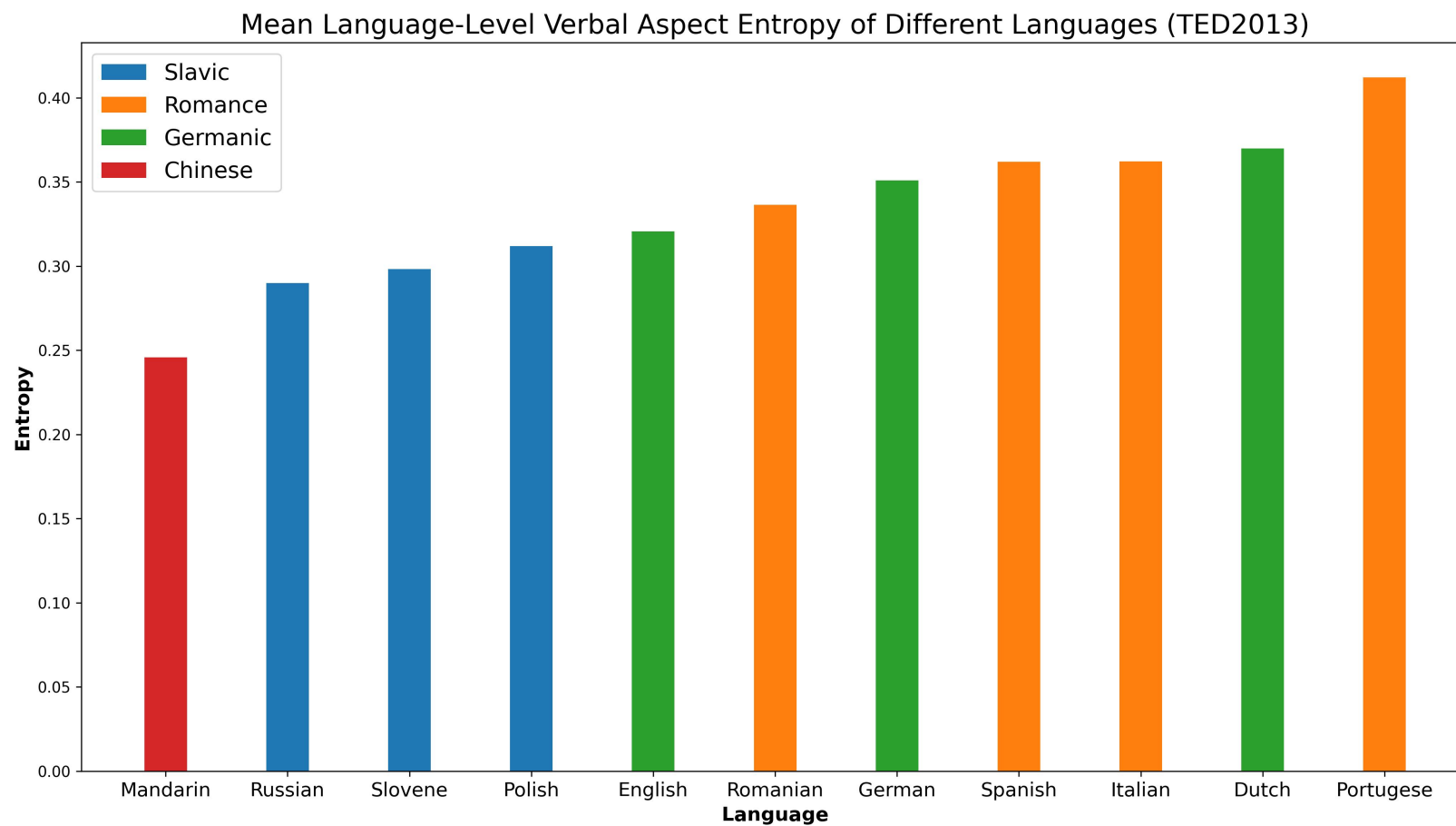


Figure 9: Look at this graph

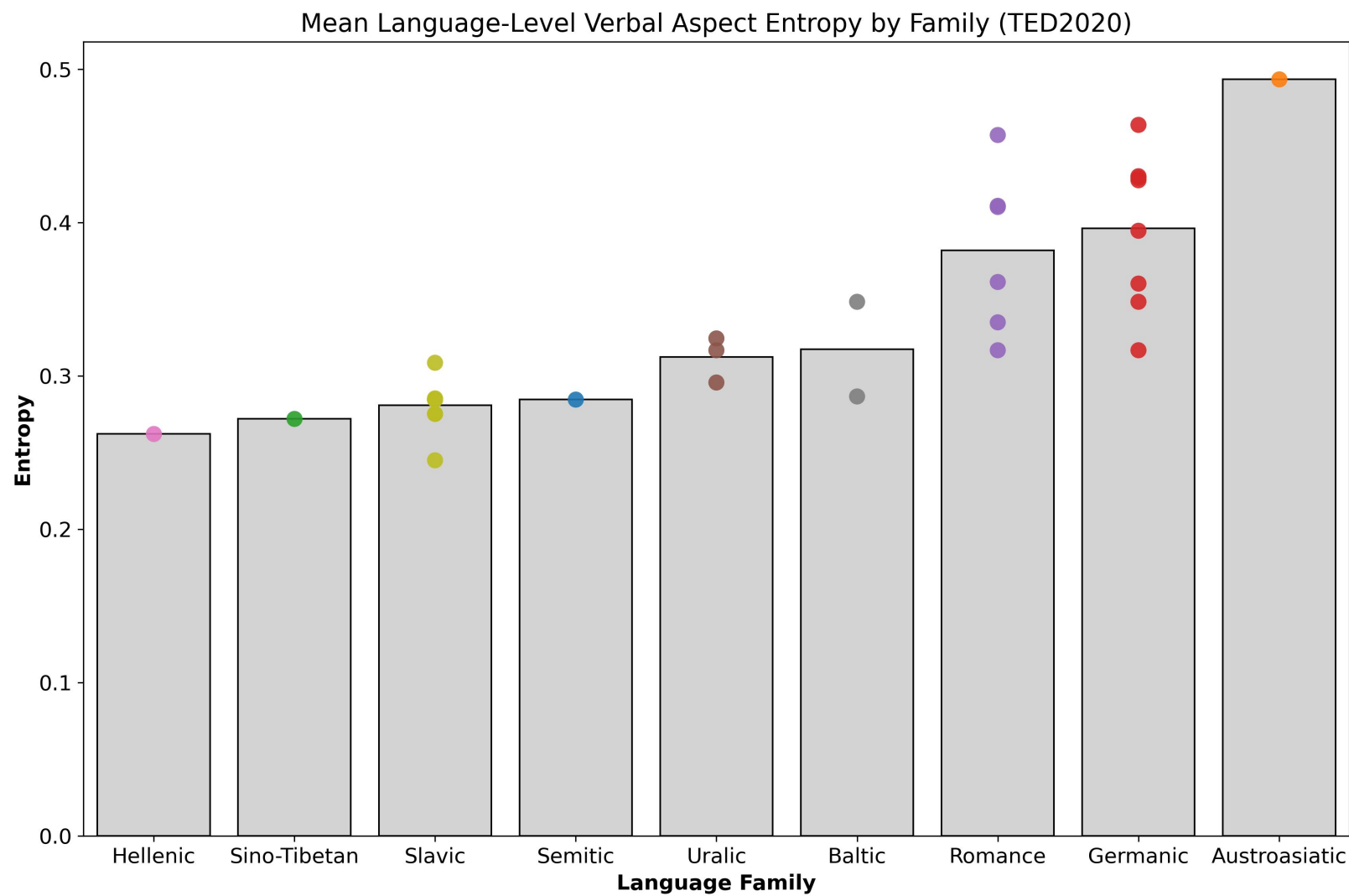


Figure 10: DESCRIPTION. The grey bars represent the mean entropy of each language family.

6 Discussion

Discussing the results.

It is clear that the model is far from perfect.

Pull together the aspect classification and aspect ambiguity models (why did I have two?)

7 Conclusion

Verbal aspect is a prominent linguistic phenomenon, but also highly complex.

Attempts to classify it are often either too coarse-grained ARE UNSATISFACTORY OR SMTH

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, how these tools deal with aspectual ambiguity, and what we can learn from this behaviour.

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

Multilingual model didnt work so well

AMBIGUITY DEPENDS ON THE SCHEMA USED!!

Come back to questions

Main take-aways?

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.2 Temporal reasoning

A.3 Long LLM prompt

Verbal aspect in language indicates how an action unfolds over time, emphasizing its internal structure, such as whether it is ongoing, completed, repeated, or momentary, distinct from tense which specifies when the action occurs relative to a reference point. The annotation distinguishes five base level aspectual values. The State value corresponds to stative events: no change occurs during the event. The Habitual value is annotated on events that occur regularly. The Activity value indicates an event with no inherent goal that has not necessarily ended and may be ongoing. Endeavor is used for processes which have an inherent end goal but which end without reaching completion (i.e., termination), whereas Performance is used for processes that reach a completed result state. Which class does "*{verb}*" belong to in this sentence: state, habitual, activity, endeavor or performance?

A.4 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.

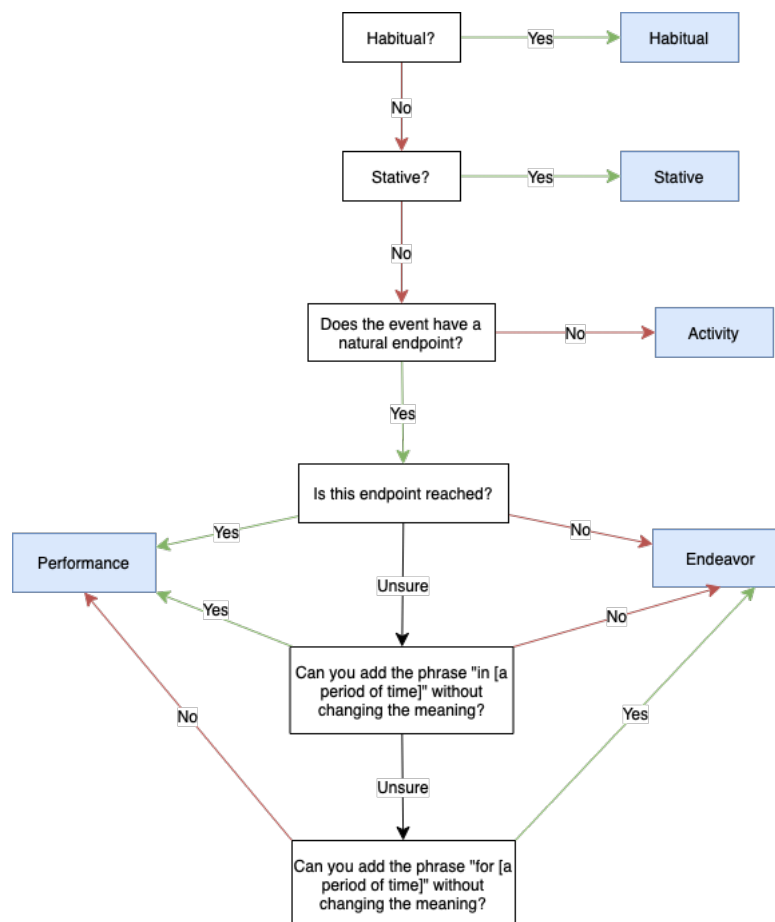
Some sentences have several possible interpretations. For example: “Let the streets be **filled** with song” could be both a state and an endeavor. This is an important part of the annotation. In this case, please enter all classes which are plausible, separated by a comma.

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

A.5 Language-level entropy

Table 14: Language-level Entropy on TED2013 dataset

Language	Mean Entropy	Variance
Chinese	0.24579	0.33647
English	0.32071	0.19312
German	0.35085	0.42130
Dutch	0.36985	0.41680
Italian	0.36227	0.42906
Romanian	0.33647	0.40408
Spanish	0.36194	0.43184
Portuguese	0.41209	0.43718
Polish	0.31180	0.41160
Slovene	0.29825	0.38113
Russian	0.28984	0.38741

Table 15: Language-level Entropy on TED2020 dataset

Language	Mean Entropy	Variance
Arabic	0.28453	0.38861
Vietnamese	0.49336	0.46466
Chinese	0.27207	0.34002
Danish	0.43019	0.42449
Dutch	0.36016	0.41209
English	0.31668	0.19319
German	0.34829	0.42011
Icelandic	0.39461	0.43853
Norwegian Bokmål	0.46367	0.46847
Norwegian Nynorsk	0.42910	0.46448
Swedish	0.42756	0.46448
Catalan	0.45707	0.48769
French	0.41091	0.44966
Italian	0.36121	0.42817
Portuguese	0.41012	0.43479
Romanian	0.33490	0.40521
Spanish	0.31668	0.43953
Estonian	0.31668	0.43953
Finnish	0.29564	0.36784
Hungarian	0.32440	0.39755
Greek	0.26212	0.36320
Latvian	0.28655	0.36055
Lithuanian	0.34829	0.42011
Belorussian	0.28453	0.38861
Bulgarian	0.28453	0.38861
Croatian	0.27519	0.37766
Czech	0.27519	0.37766
Polish	0.30852	0.40623
Russian	0.28529	0.38692
Slovakian	0.28453	0.38861
Slovene	0.28453	0.38861
Ukrainian	0.24490	0.35603

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>.
- 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>.
- 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 Wherry. 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 D ne 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>.
- 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>.
- 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.
- 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>.

- Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA, 1965.
- John R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, UK, 1969.
- J. L. Austin. *How to Do Things with Words*. Clarendon Press, Oxford, 1962.
- H.P. Grice. Logic and conversation. *Syntax and Semantics*, 3:41–58, 1975.
- 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.
- Hana Filip. *Lexical Aspect*, pages 721–751. 01 2012. doi: 10.1093/oxfordhb/9780195381979.013.0025.
- 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>.
- 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 Ilya 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>.
- Julia Lukassek, Anna Prystowska, Robin Hörnig, and Claudia Maienborn. The semantic processing of motion verbs: Coercion or underspecification? *Journal of Psycholinguistic Research*, 46, 08 2017. doi: 10.1007/s10936-016-9466-7.
- Christopher Kennedy and Beth Levin. Measure of change: The adjectival core of degree achievements. In *Adjectives and Adverbs*, 2008. URL <https://semantics.uchicago.edu/kennedy/docs/kl07-measure.pdf>.
- W. Styler, K. Crooks, T. O’Gorman, and M. Hamang. Richer Event Description (RED) Annotation Guidelines, 2014. Unpublished manuscript, University of Colorado at Boulder.
- 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>.

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>.

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.

Yuri Kuratov and Mikhail Arkhipov. Adaptation of deep bidirectional multilingual transformers for russian language, 2019.

Mehdi Rezaee, Kasma Darvish, Gaoussou Youssouf Kebe, and Francis Ferraro. Discriminative and generative transformer-based models for situation entity classification, 2021.

Thomas A. Mathew and E. Graham Katz. Supervised categorization of habitual and episodic sentences. In *Sixth Midwest Computational Linguistics Colloquium*, Bloomington, Indiana, 2009. Indiana University.

Rei Ikuta, Will Styler, Mariah Hamang, Tim O’Gorman, and Martha Palmer. Challenges of adding causation to richer event descriptions. In Teruko Mitamura, Eduard Hovy, and Martha Palmer, editors, *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 12–20, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-2903. URL <https://aclanthology.org/W14-2903>.

- Thomas Kober, Malihe Alikhani, Matthew Stone, and Mark Steedman. Aspectuality across genre: A distributional semantics approach. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4546–4562, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.401. URL <https://aclanthology.org/2020.coling-main.401>.
- Alessandra Zarcone and Alessandro Lenci. Computational models for event type classification in context. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/315_paper.pdf.
- Jürgen Hermes, Michael Richter, and Claes Neufeind. Automatic induction of german aspectual verb classes in a distributional framework. 09 2015.
- Rowan Zellers and Yejin Choi. Zero-shot activity recognition with verb attribute induction. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 946–958, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1099. URL <https://aclanthology.org/D17-1099>.
- Dan Kondratyuk and Milan Straka. 75 languages, 1 model: Parsing Universal Dependencies universally. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1279. URL <https://aclanthology.org/D19-1279>.
- Carlota S. Smith. *Modes of Discourse: The Local Structure of Texts*. Cambridge Studies in Linguistics. Cambridge University Press, 2003.
- Alexis Palmer, Elias Ponvert, Jason Baldridge, and Carlota Smith. A sequencing model for situation entity classification. In Annie Zaenen and Antal van den Bosch, editors, *Proceedings*

of the 45th Annual Meeting of the Association of Computational Linguistics, pages 896–903, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL <https://aclanthology.org/P07-1113>.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. Situation entity types: automatic classification of clause-level aspect. 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 1757–1768, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1166. URL <https://aclanthology.org/P16-1166>.

Maria Becker, Michael Staniek, Vivi Nastase, Alexis Palmer, and Anette Frank. Classifying semantic clause types: Modeling context and genre characteristics with recurrent neural networks and attention. In Nancy Ide, Aurélie Herbelot, and Lluís Màrquez, editors, *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 230–240, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-1027. URL <https://aclanthology.org/S17-1027>.

Annemarie Silke Friedrich. States, events, and generics: computational modeling of situation entity types. 2017.

Zeyu Dai and Ruihong Huang. Building context-aware clause representations for situation entity type classification. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3305–3315, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1368. URL <https://aclanthology.org/D18-1368>.

Francisco Costa and António Branco. Aspectual type and temporal relation classification. In Walter Daelemans, editor, *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 266–275, Avignon, France, April 2012. Association for Computational Linguistics. URL <https://aclanthology.org/E12-1027>.

Leon Derczynski and Robert J. Gaizauskas. Temporal relation classification using a model of tense and aspect. In *Recent Advances in Natural Language Processing*, 2015. URL <https://aclanthology.org/R15-1017.pdf>.

James F. Allen and Naushad UzZaman. Interpreting the temporal aspects of language. 2012. URL <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=f0b9a86a80a5bc09e448db250a97ade218e4d529>.

Hans Kamp and Uwe Reyle. *Tense and Aspect*, pages 483–689. Springer Netherlands, Dordrecht, 1993. ISBN 978-94-017-1616-1. doi: 10.1007/978-94-017-1616-1_6. URL https://doi.org/10.1007/978-94-017-1616-1_6.

Thomas Kober, Sander Bijl de Vroe, and Mark Steedman. Temporal and aspectual entailment. In Simon Dobnik, Stergios Chatzikyriakidis, and Vera Demberg, editors, *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 103–119, Gothenburg, Sweden, May 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-0409. URL <https://aclanthology.org/W19-0409>.

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>.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline, 2019.

Kenneth Church and Mark Liberman. The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence*, 4, April 2021. ISSN 2624-8212. doi: 10.3389/frai.2021.625341. URL <http://dx.doi.org/10.3389/frai.2021.625341>.

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>.

- 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.
- 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.

- Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. Back to the future: Towards explainable temporal reasoning with large language models, 2023.
- Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu lee Wang, Qing Li, and Xiaoqin Zhong. Label supervised llama finetuning, 2023. URL <https://arxiv.org/abs/2310.01208>.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>, 2022.
- Meta LLaMA Team. Introducing meta llama 3: The most capable openly available llm to date. <https://ai.meta.com/blog/meta-llama-3/>, 2024.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages, 2020. URL <https://arxiv.org/abs/2003.07082>.
- 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.
- Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- Michael S. Flier. Remarks on russian verbal prefixation. *The Slavic and East European Journal*, 19(2):218–229, 1975. ISSN 00376752. URL <http://www.jstor.org/stable/306777>.
- S.M. Dickey. *Parameters of Slavic Aspect: A Cognitive Approach*. Dissertations in Linguistics. Cambridge University Press, 2000. ISBN 9781575862361. URL <https://web.stanford.edu/group/cslipublications/cslipublications/site/1575862360.shtml>.

- Claudio Iacobini. Phrasal verbs between syntax and lexicon. *Italian Journal of Linguistics*, 21: 97–117, 01 2009.
- Bert Cappelle. *When “wee wretched words” wield weight: The impact of verbal particles on transitivity*, pages 41–54. 01 2007.
- Milada Walková. Particle verbs in english: Telicity or scalarity? *Linguistics*, 55(3):589–616, 2017. doi: doi:10.1515/ling-2017-0005. URL <https://doi.org/10.1515/ling-2017-0005>.
- Hana Filip. The telicity parameter revisited. *Semantics and Linguistic Theory*, 14, 09 2004. doi: 10.3765/salt.v0i0.2909.
- Maarten Janssen and Olga Borik. A database of russian verbal aspect. *Oslo Studies in Language*, 4, 2012. URL <http://ru.oslin.org/?action=aspect>.
- J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.
- Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.
- Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation, 2020. URL <https://arxiv.org/abs/2004.09813>.
- 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.