Adding Aspectual Information to Structured Meaning Representations

Claire Benét Post,[1] **Paul Bontempo,**[1] **August Milliken,**[1]
Nicholas Derby, Saksham Khatwani, Sumeyye Nabieva,
Alvin Po-Chun Chen, Karthik Sairam, Alexis Palmer

University of Colorado Boulder
{benet.post, paul.bontempo, august.milliken, alexis.palmer}@colorado.edu

Abstract

To fully capture the meaning of a sentence, semantic representations should encode *aspect*, which describes the internal temporal structure of events. In graph-based meaning representation frameworks such as Uniform Meaning Representations (UMR), aspect expresses how events unfold over time, including distinctions such as states, activities, and completed events. Despite its importance, aspect remains sparsely annotated across semantic meaning representation frameworks, hindering not only current manual annotation, but also the development of automatic systems capable of predicting aspectual information. In this paper, we introduce a new dataset of English sentences annotated with UMR aspect labels over Abstract Meaning Representation (AMR) graphs. We describe the annotation scheme and guidelines used to label eventive predicates according to the UMR aspect lattice, as well as the annotation pipeline used to ensure consistency and quality across annotators through a multi-step adjudication process. To demonstrate the utility of the dataset for future automation, we perform simple baseline experiments using three modeling approaches. Our results establish initial benchmarks for automatic UMR aspect prediction and provide a foundation for integrating aspect into semantic meaning representations more broadly.

Keywords: Aspect annotation, semantic meaning representations, aspectual generation benchmarks

1. Introduction

Semantic representations frequently center around capturing components of meaning related to the core *events* conveyed by individual natural language utterances. Nearly all meaning representation (MR) formats express the core predicates associated with those events, along with any arguments to those predicates. MRs differ quite substantially, though, when it comes to the expression of additional event information, such as tense, modality, aspect, or information structure. Languages also differ substantially in the degree to which they grammaticalize (or require the expression of) the same event-related information.

Aspect is a core component of Uniform Meaning Representation (UMR),[1] a graph-based semantic framework designed to represent meaning in a cross-linguistically applicable and computationally tractable way. UMR is an extension and modification of the Abstract Meaning Representation (AMR) framework (Banarescu et al., 2013), an MR which nicely suits the typological properties of English but which begins to strain when adapted for other languages. Like AMR, UMR is a graph-based framework encoded in Penman-style notation (Wein and Bonn, 2023). Unlike AMR, UMR was designed by and with linguistic typologists, in order to build an approach to annotation suitable for diverse languages (Van Gysel et al., 2021).

```
She is still writing her paper.
(w/ write-01
    :ARG0 (p/ person
        :ref-person 3rd
        :ref-number Singular)
    :ARG1 (p2/ paper
        :poss p
        :ref-number Singular)
    :mod (s/ still)
    :aspect Activity
    :modstr FullAff)
```

Figure 1: Example UMR graph with the eventive predicate `write`. This graph captures predicate argument structure, properties of the arguments, and event-related properties. We highlight parts of the graph relevant for the aspect label *Activity*.

Unlike tense, which encodes *when* an event occurs, aspect captures the *how*: the internal temporal structure, duration, and completedness of events (Comrie, 1976; Croft, 2012; Donatelli et al., 2018). It allows a semantic system to distinguish between, for example, habitual, ongoing processes, or completed achievements, enabling a more nuanced interpretation of event semantics.

In UMR, aspect is applied to all *eventive* elements (also known as *eventualities*) in a sentence. The central eventuality introduced by an utterance is typically the concept aligned with the main finite

[1] Equal contribution.

[1] https://umr4nlp.github.io/web/

Proceedings of The Seventh International Workshop on Designing Meaning Representations (DMR 2026) @ LREC 2026, pages 20–36

verb, as seen in Figure 1. Eventualities in UMR refer to the full predication, encompassing the verb and its arguments (Donatelli et al., 2018; Kingsbury and Palmer, 2003). UMR defines a particular inventory of aspectual categories, following Croft (2022), aligned with other well-established event typologies, and including (among others) states, activities, accomplishments, achievements, and processes (Bach, 1986). In UMR the aspectual categories are organized into a lattice that supports both coarse- and fine-grained aspectual distinctions. Unlike surface grammatical cues, like those in auxiliaries or verb morphology, UMR aspect is a semantic feature. It abstracts away from morphosyntactic form to represent covert event structure and is intended to generalize across typologically diverse languages (Van Gysel et al., 2021).

Annotating aspect is no simple feat. Theoretical debates span decades, including disagreements about the universality of aspectual categories, the granularity of classifications, and their interaction with tense and modality (Reichenbach, 1947; Vendler, 1957; Comrie, 1976; Langacker, 2011; Dowty, 1986; Hinrichs, 1986; Moens and Steedman, 1988; Klein, 2013; Chang et al., 2022; Partee, 2011; Croft, 2012). While there are a number of corpora with aspect annotations, and several computational models (Friedrich et al., 2023), there is no unified approach to annotation or modeling (among others: Pustejovsky et al., 2003; Derczynski, 2017; Pustejovsky et al., 2017; **?**; Friedrich et al., 2016; Mostafazadeh et al., 2016; Laparra et al., 2018; O'Gorman et al., 2016; Gantt et al., 2022).

From a typological perspective, some languages encode aspect more saliently than others, further complicating annotation for multilingual or cross-linguistic frameworks. For example, American Sign Language and Mandarin Chinese prioritize aspectual distinctions over tense (Li and Thompson, 1989; McDonald, 1982), while Hindi includes a dedicated aspect morpheme separate from tense or mood (Van Olphen, 1975). In contrast, many Indo-European languages conflate aspect and tense morphologically, often obscuring the underlying semantic distinctions. The categories in and structure of the UMR aspect lattice are flexible enough to precisely encode aspectual distinctions as seen in each of these languages.

Given these complexities, manual aspect annotation is time-consuming, error-prone, and highly sensitive to annotator interpretation. Yet its inclusion in UMR is a cornerstone to achieving a more expressive, cross-linguistic meaning representation system. UMR builds on earlier formalisms such as Abstract Meaning Representation (AMR) (Banarescu et al., 2013), where aspect was initially introduced to support event-based reasoning but was never fully adopted into standard annotation guidelines. Donatelli et al. (2018) formalize aspect in AMR, laying out annotation principles and aligning event types with lexical frames.

Despite its importance for accurately representing event semantics, aspect remains underannotated in existing UMR resources. This scarcity limits both the scale and consistency of manual UMR annotation and hinders development of automatic parsers capable of reliably predicting aspect.

To address this bottleneck, we present a new dataset of English sentences manually annotated with UMR aspect labels. The study presented here is part of an ongoing effort to build a large-scale English UMR dataset by converting existing AMR graphs to UMR representations for the same sentences Bonn et al. (2023). The conversion is a multistep process, combining automated processing and manual intervention. The AMR graphs which are the foundation of that conversion effort do not include any aspect annotations, and the new dataset is intended to support development of automated aspect annotation to fill in aspect values for the remaining AMR graphs. Under our approach, annotators are provided with AMR-derived graphs for these same sentences sentence and asked to label eventualities according to the UMR aspect lattice (see Section 4.1 for corpus details).

Our annotation pipeline combines structured annotator training with multiple rounds of independent annotation, group adjudication sessions, and expert consultation to improve the guidelines and resolve difficult annotation decisions, resulting in a high-quality resource intended to support future UMR modeling efforts. To validate dataset quality, we additionally report baseline experiments spanning three modeling families: (1) a rule-based approach, (2) an embedding-based classifier, and (3) a large language model (LLM) prompting approach.

We target two complementary objectives:

- **Task 1: Data annotation.** We construct a new gold-standard dataset of 1473 carefully validated sentences labeled according to the UMR aspect annotation scheme.
- **Task 2: Baseline modeling.** We present standard data splits and initial performance benchmarks for automated UMR aspect labeling.

This is the first dataset designed to support supervised learning for UMR aspect labeling.[2]

2. Related Work

The semantics of aspect has been a long-standing topic of debate in linguistic theory. Seminal works

[2]All data and labels available at: https://github.com/clairepost/UMR_Aspect_Data.git

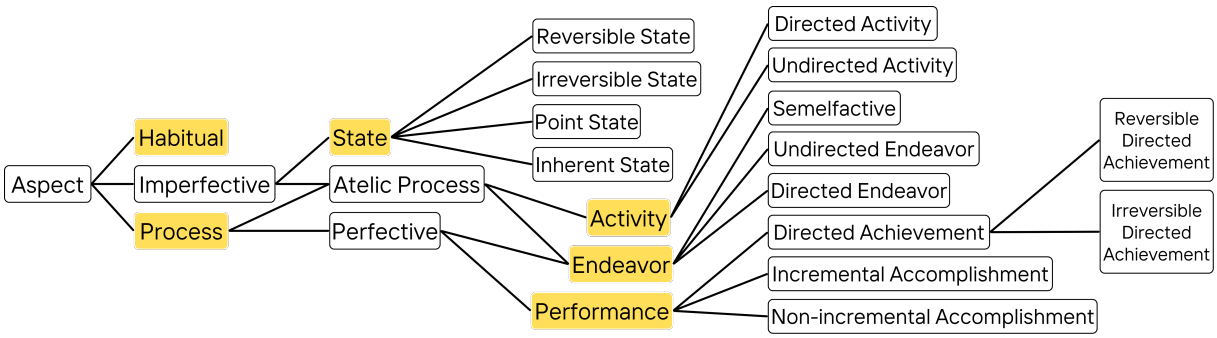

Figure 2: UMR aspect lattice with aspectual values utilized in our English annotation highlighted in yellow.

by Reichenbach (1947), Vendler (1957), and Comrie (1976) lay the foundation for distinguishing between types of eventualities–states, achievements, activities, accomplishments–based on their temporal and structural properties. Dowty (1986) and Langacker (2011) further explore the interaction between aspect, argument structure, and lexical semantics. These formalisms inform how events are modeled in UMR today.

Later developments such as Hinrichs' interval-based models (1986), Moens and Steedman's narrative structure theory (1988), and Klein's temporal logic (2013) introduce more formal ways to encode event structure and its temporal entailments. These insights highlight the need for meaning representation frameworks like UMR to go beyond grammatical tense and directly encode aspectual distinctions based on semantic content.

Aspect annotation. Aspect has been incorporated into previous semantic annotation and event modeling efforts, particularly in temporal information extraction. TimeML (Pustejovsky et al., 2003) and its follow-up projects such as the TempEval competitions (Derczynski, 2017) include annotation for aspect, typically via shallow textual cues. There are several datasets developed with robust manual aspect annotation that consider sentential context and event structure, such as DIASPORA (Kober et al., 2020) and the Universal Decompositional Semantics dataset (Gantt et al., 2022); these datasets employ coarse-grained aspect classes, rather than the more extensive and typologically-broad lattice provided by the UMR framework. Other recent work seeks to automate aspect classification using linguistic features (**?**), discourse roles (Friedrich et al., 2016), and LSTM-based models that integrate context (Mostafazadeh et al., 2016; Laparra et al., 2018).

While effective to some degree, these systems often operate over flat text or shallow syntactic representations. They do not handle the rich predicate-argument structures or graph-based semantics found in UMR and AMR. Moreover, they treat aspect as a kind of downstream feature, rather than an integral part of event structure representation.

Donatelli et al. (2018) develop an approach for adding feature-based tense and aspect information to AMR graphs. For aspect, the scheme encodes four crucial semantic features: `+/-stable`, `+/-ongoing`, `+/-complete`, and `+/-habitual`. Many aspectual categories can be derived from the combination of semantic feature values. Under the UMR scheme, however, aspectual categories are to be labeled directly rather than broken down into features.

Automatic aspect annotation for UMR. Due to the small amount of available UMR data, prior work has focused primarily on methods for generating UMR graphs without supervised training. Chun and Xue (2024) propose a multi-step strategy for converting AMR graphs into UMR graphs, including the addition of aspect. The approach derives aspect from `Tense` and `VerbForm` features output by UDPipe v2 (Straka, 2018). Similarly, Sun et al. (2024) experiment with few-shot and Think-Aloud prompting on LLMs to generate Chinese UMR graphs, including aspect labels.

AutoAspect, which directly targets UMR aspect, proposes a branching rule-based approach specifically for classifying UMR aspects in English UMR graphs (Chen et al., 2021). AutoAspect delivers high precision for two aspectual categories and variable results for others. We aim to build a larger aspect-labeled dataset to support a broader range of learning approaches. We further note that significant refinements have been made to the UMR dataset since this system was published, as seen in (Bonn et al., 2024a).

3. Annotation Scheme

3.1. UMR Aspect Lattice

This paper is concerned with the aspect annotation of English sentences, as highlighted in Figure 2. UMR organizes aspectual categories within an as-

pect lattice, a hierarchical structure that captures relationships between coarse and fine-grained labels. This design allows annotations to represent general distinctions while remaining compatible with more specific readings when additional linguistic information is available.

This structure is particularly useful for cross-linguistic semantic annotation. English often relies on relatively coarse-grained aspectual distinctions, while other languages encode finer aspectual contrasts directly in their grammar. The lattice enables UMR to support consistent representations across typologically diverse languages. This was a major motivation for our work, and we hope in the future to expand aspect annotation to more languages.

The aspectual categories chosen for English annotation include a set of base-level distinctions–*State*, *Performance*, *Endeavor*, *Activity*, and *Habitual*, and *Habitual*–as well as a more coarse-grained value for event nominals and other underspecified events, *Process*.

3.2. Aspect Types

State. This value corresponds to stative events, indicating that no change occurs during the event, as prescribed by (Vendler, 1967). It includes predicate nominals, predicate locations, and thetic (presentational) possession.

[1] *The cat loves milk*.

```
(l/ love-01
    :ARG0 (c/ cat
        :ref-number Singular)
    :ARG1 (m/ milk)
    :aspect State
    :modstr FullAff)
```

More specifically, in English, the *State* value encompasses modal verbs (e.g., "The cat **needs** to eat.") and events under the scope of ability modals (e.g., "The cat is **able** to eat."). UMR classifies *inactive actions*, as defined by (Croft, 2012), as stative. This includes posture verbs (e.g., "The cat **hangs** on the windowsill."), perception verbs (e.g., "The cat **sees** milk."), mental activities (e.g., "The cat **thinks** about jazz."), verbs of operation (e.g., "The cat is **working** on catching mice."). The *State* value, in English, is also an umbrella that covers inherent states (e.g., "The cat **is** black."), reversible states (e.g., "The cat is **hungry**."), irreversible states (e.g., "The glass **is shattered**."), and point states (e.g., "When it **is** 12:30pm, feed the cat.").

Performance. This category covers events that reach a result state, such as achievements that have some instantaneous binary change, accomplishments where there is a run-up process before the change, or when the event reaches a result state that has a natural endpoint.

[2] *The cat walked along the fence in 2 minutes.*

```
(w/ walk-01
    :ARG0 (c/ cat
        :ref-number Singular)
    :ARG2 (a/ along
        :op1 (f/ fence))
    :duration (t/ temporal-quantity
        :unit (m/ minute)
        :quant 2)
    :aspect Performance
    :modstr FullAff)
```

For instance, completive markers (e.g., "The cat finished **climbing** *up the tree*.") and container adverbials (e.g., "The cats **scampered** along the fence *in 10 seconds*.") are both indicators that an event has reached a distinct result state. Note that the temporal expression "in two minutes" appears in the graph under the attribute `:duration`.

Endeavor. *Endeavor* is used for processes that end within the time window in question, but do *not* reach a particular result state: e.g., compare graph [2] to graph [3] below.

[3] *The cat walked along the fence.*

```
(w/ walk-01
    :ARG0 (c/ cat
        :ref-number Singular)
    :ARG2 (a/ along
        :op1 (f/ fence))
    :aspect Endeavor
    :modstr FullAff)
```

The *Endeavor* value is often mistaken for *Performance* and vice versa. Often, in English predicates have explicit aspectual marking to be considered an *Endeavor*. Terminative aspectual markers, like "stop" in English, and durative adverbials (e.g., "The cat **ate** kibble *for thirty seconds*.") are both strong indicators for *Endeavor*.

Activity. The *Activity* aspect covers processes that do not start or end during the time window in question. They can be ongoing with respect to present or past time (e.g., “The cat **was playing** the piano.”) For an example graph, see Figure 1.

Identifying *Activity* is difficult because it is largely dependent upon context, document creation time, and real world knowledge. However, there are some grammatical clues that can help. For example, if the event is in the present progressive (e.g., "The cat **is playing** the piano."), it is typically annotated with *Activity*. Inceptive and continuative aspectual markers may also imply that an event has not ended (e.g., "The cat **started playing** the piano." and "The cat **kept on playing** the piano."). In UMR, iterative events are labeled *Undirected Activity*; they fit under the *Activity* umbrella for us.

Habitual. The *Habitual* aspectual sense is usually straightforward to identify. It covers things that happen repeatedly or regularly. In English, adverbials such as “used to” and “always” often (but not always) modify the verb in habitual events.

[4] *The cat eats kibble.*

```
(w/ eat-01
    :ARG0 (c/ cat
        :ref-number Singular)
    :ARG1 (k/ kibble)
    :aspect Habitual
    :modstr FullAff)
```

In English, *Habitual* is often expressed through simple present tense, while habituals in the past are often indicated with "used to" (e.g., "The cat **used to eat** kibble.").

Process. Of the labels we use for English, *Process* is the most coarse-grained. It describes an ongoing event where the beginning or end is uncertain or unspecified. The most common use of *Process* in our dataset is as the default label for event nominalizations (e.g., "The cat denied **wrongdoing**.").

[5] *After the game, the cat slept.*

```
(s/ sleep-01
    :ARG0 (c/ cat
        :ref-number Singular)
    :temporal (a/ after
        :op1 (g/ game
            :aspect Process))
    :aspect State
    :modstr FullAff)
```

Graph [5] shows another category of events typically annotated with *Process* in UMR. Here, the **game** event is packaged in a referring expression, and the prepositional phrase appears under the attribute *:temporal*. We take a similar approach for underived nominals, nominalizations, and gerunds.

None. For our annotations, we additionally allow annotators to apply the label *NONE* for predicates that they deem to be non-eventive, even though the appear in the graph as a semantic predicate. This happens frequently for adjectival and adverbial concepts, which do not participate in eventualities. These show up as predicates because they often receive automatic mappings into FrameNet predicates in AMR (Baker et al., 1998); these are then carried over into the UMR versions of the graphs.

3.3. Comparison to Other Aspect Annotation Schemata

Aspect annotation has been widely studied in linguistics, and UMR is a recent schema that builds on prior theoretical work. In particular, UMR follows approaches such as Croft (2012), which emphasize that aspectual interpretation depends on multiple factors and that a single event may admit multiple plausible aspectual readings depending on context. This perspective informed our adjudication process when resolving difficult cases.

Because our task is motivated by downstream NLP and machine learning applications, our annotation process follows principles outlined by Pustejovsky et al. (2017). We prioritize consistency in label assignment in order to maximize the learnable signal in the dataset, even when this means limiting the number of annotated examples. Nevertheless, as shown in Figure 2, the selected aspect labels differ in granularity, which introduces variation in specificity across the dataset.

Prior work such as the DIASPORA dataset (Kober et al., 2020) also explores aspect annotation but employs a more coarse-grained 3-label schema (*state*, *telic*, *atelic*) to reduce label overlap and maintain uniform granularity. Our dataset instead adopts the richer UMR aspect inventory while remaining compatible with prior work. To illustrate this compatibility, we developed a mapping between our schema and the DIASPORA labels and automatically applied it to a subset of our data, manually evaluating the resulting label assignments. We found that the two schemata are broadly compatible, excepting the case of event nominals, which take *process* labels in UMR but are without a consistent equivalent in DIASPORA, requiring manual adjudication rather than automated label mapping.

4. Building the Corpus

4.1. Data

Our dataset is sourced from the UMR 2.0 Dataset (Bonn et al., 2025) which contains roughly 30,000 UMR graphs in different stages of conversion from AMR graphs (Knight et al., 2020; Bonn et al., 2020).[3] Some UMR 2.0 graphs have aspect annotations from previous work; we only annotate graphs that do not yet have aspect labels. To ensure broad coverage for training and evaluation, we select four corpora from the dataset to annotate:

1. The Little Prince corpus, a set of sentences from the English translation of *The Little Prince* by Antoine de Saint-Expupéry.

2. The Minecraft corpus, a set of dialogues and corresponding grounding data from a collaborative structure-building task in Minecraft (Narayan-Chen et al., 2019).

[3]We keep all graphs in the original format for aspect labeling; we make no structural modifications or revisions to the data.

3. The BOLT DF corpus, which contains English language forum posts crawled as part of the DARPA BOLT project.

4. The Weblog corpus, comprised of weblog posts and online news articles.

A detailed summary of aspect label statistics, by corpus, for the existing UMR dataset can be found in Appendix A as Table 5.

4.2. Annotation

Annotation proceeded in two phases: in PHASE 1, a team of 8 annotators worked in pairs to label each event marked in the graphs. In PHASE 2, a smaller group focused on adjudicating decision ties with expert consultation, while ensuring consistency with previous annotations.

4.2.1. Phase 1: Initial Annotation

Given the complexity of aspect annotation and its theoretical underpinnings, we first focused on helping all members of the team develop a strong understanding of the UMR aspect schema, before proceeding to the bulk annotation of PHASE 1. Our goals were to ensure that each annotator contributed quality data and that rules were applied consistently between annotators. To this end, we conducted 8 weekly training sessions throughout PHASE 1.

Annotation guidelines and training materials were built from existing UMR resources and developed into task-specific training materials.[4] Each week, team members presented on different topics from these materials and discussed example annotations as a group to clarify issues.[5]

We conducted an initial practice task in which each team member annotated up to 50 predicates from the Pear Story corpus (Bonn et al., 2023). This dataset was selected for its short, visually grounded sentences, which aided learning and facilitated discussion. We reviewed inter-annotator agreement and recurring errors on the practice task before proceeding to bulk annotation.

Label-wise accuracy measured over the Pear Story annotations showed that the categories *State* and *Performance* were quite reliably and consistently identified, while minority classes like *Endeavor* and *Habitual* were less consistent. These findings prompted us to hold a focused error correction and continued training session, in which we reviewed common sources of confusion, such as distinctions between *State* vs. *Performance* and *Performance* vs. *Endeavor*.

We next moved into full-scale annotation. Each corpus was assigned to two annotators for independent labeling, resulting in two first-pass labels per event per corpus. Each numbered predicate within the AMR graphs was annotated with one of the six UMR aspect labels or marked with *NONE* if the predicate was deemed non-eventive. Table 1 shows the distribution of aspect labels for the completed dataset. The first-pass bulk annotation lasted approximately 6 weeks.

4.2.2. Phase 2: Tie-breaking and Adjudication

Following PHASE 1, all events with conflicting aspect labels were routed to a tie-breaking process. Each sentence and its annotations were reviewed by a third annotator to make a final determination. If the adjudicating annotator disagreed with both original labels, the sentence went to an additional adjudication step, up to a maximum of 5 total annotation rounds. All intermediate labels are preserved and ranked in our final dataset, along with the final adjudicated labels.

In the next step, a team of two annotators reviewed all data for consistency. Together, the tie-breaking and consistency adjudication lasted about 8 weeks. This review process ensured each adjudicated aspect label: (i) follows our annotation guidelines, and (ii) is consistent with other instances in the dataset. To confirm consistency of a given label, we compare difficult cases to sentences with the same event *and* to events with the same label.

The duration and complexity of this annotation process indicates the corresponding complexity of aspect itself; even with months of discussion, some instances in the dataset remained ambiguous. For particularly complex disagreements—such as differentiating *Endeavor* from *Performance*—we consulted directly with external experts to align the annotations with their interpretations.

4.3. Inter-Annotator Agreement

We report agreement metrics across the various phases of our annotation and adjudication process, compared against our finalized gold-level labels in Table 1. Overall, these metrics indicate sufficient consensus for further analysis using our dataset.

Per-class Krippendorff's alpha. These values, computed for each aspect label class against all other classes combined, measure the agreement across annotators from zero (random chance) to 1 (perfect agreement), with a standard significance threshold of 0.67 (Krippendorff, 2004).

[4] See Appendix B for more details.

[5] Resources from these meetings are on GitHub: https://github.com/clairepost/UMR_Aspect_Data.git.

Per-class		
Label	$K\alpha*$	**Count**
None	0.735	514
State	0.706	385
Performance	0.716	360
Process	**0.547**	100
Habitual	0.704	56
Activity	**0.502**	40
Endeavor	**0.559**	18
Total		**1,473**
First-pass		
Percent agreement		74.1%
Cohen's κ		0.656

Table 1: Inter-annotator agreement metrics; boldface numbers indicate low agreement. *Per-class Krippendorff's alpha computed as one-vs-rest.

Gold label	Error rate	Error Label
Activity	6.7%	Performance
Endeavor	12.2%	Performance
Habitual	6.8%	Performance
None	6.0%	State
Performance	3.9%	None
Process	9.1%	None
State	4.8%	None

Table 2: Final gold aspect labels with the error rate compared to the final adjudication. 'Error Label' is the most common incorrect annotation.

First-pass metrics. Percent agreement represents the proportion of events where both first-pass annotators provided the same label for an event, though this does not necessarily mean that such events were exempt from adjudication later. Cohen's κ measures agreement on the same scale as Krippendorff's α, but strictly between two annotators, so we report the κ score for the two first-pass annotators, averaged across all 4 corpora.

5. Annotation Challenges

Table 5 reports the most frequent erroneous labels applied during our annotation process, with the associated gold-standard label in the left column. The results illustrate that certain types of events are particularly difficult to disambiguate, even for experts. In Table 3 we present representative examples of particularly challenging annotation cases. Many of these issues stem from the limited contextual information available to annotators, as sentences were presented individually rather than as part of complete documents. In naturally occurring discourse, surrounding context typically resolves such ambiguities, and aspectual interpretation is no exception.

For instance, in Table 3 example (a), the event suggests a *Process* label in the sentence context, since the speaker specifies no explicit number of blocks, and the event lacks an inherent endpoint. However, the wider discourse context (blocksworld video game) includes a particular number of red blocks available to the players, and a finite grid on which to place them, which instead motivates a *Performance* label.

Similarly, in example (b), contextual cues could determine whether an event is an *Endeavor* or a *Performance*, depending on whether we understand "clap" as a process without change of state (hands end up as they started), or a complete event that reaches a natural conclusion (hands start apart and end together in a single motion). Without surrounding document context, we cannot say for sure which type of clap this sentence describes.

To account for multiple plausible interpretations, our adjudication schema allows for the identification of a secondary label that reflects a reasonable alternative reading, even when a primary label is selected based on the most likely interpretation. For events without such ambiguity such as example (c), we provide only one adjudicated label.

6. Automatic Annotation

One goal of this annotation effort is to build training and evaluation data for automatic aspect classification. Toward that end, we establish standard data splits and evaluate the performance of several simple baseline models, leaving development of more sophisticated models for near future work.

6.1. Data splits

We establish standardized splits of our final dataset in a 70/15/15 train/val/test ratio, as illustrated in Table 7, found in Appendix C. We split the data on sentence boundaries, rather than by event, to ensure that no contextual information is shared across splits. This is necessary because a single sentence may include multiple events, and sentential context seen in training for one event could unfairly inform test performance on another. We stratified each split to ensure consistent and balanced class distribution.[6] To accomplish this, we first identify a dominant aspect class for each sentence, by counting the most frequent label across all events per sentence. For sentences with just one event or whose events all have different labels, we consider the first event's label to be dominant. We perform an initial 70/30 split of the sentences, keeping the label distribution of both consistent with those of the overall dataset. We do the same for the secondary

[6] Exact data splits are available on our GitHub.

Sentence	Event	Main label	2nd label
(a) <Architect> now above that place red blocks on the grid.	Place	Performance	Process
(b) "Clap your hands, one against the other," the conceited man now directed him.	Clap	Endeavor	Performance
(c) Greed is when you are wealthy and lobby your representatives for special tax breaks because you are over 60 years of age.	Age	State	N/A

Table 3: Selected examples with ambiguous aspect; some annotated with secondary labels.

division on the 30% split, producing comparable validation and test splits.

6.2. Baseline models

To establish a performance baseline for this task on this data set, we test on two types of models: 1) LLMs (both closed and open-source) under a simple prompting paradigm; and 2) a simple feedforward neural architecture.

LLM Prompting. We experiment with multiple LLMs in a prompting paradigm to evaluate their capability for aspect classification without fine-tuning. LLM performance on structured prediction tasks has been shown to vary drastically based on slight changes to prompt structure (Lu et al., 2022), and other work suggests that LLMs lack meta-linguistic reasoning capability (Bonn et al., 2024b); we examine the ability of LLMs to identify covert aspectual information from a sentence, as well as produce a baseline against which to compare other neural approaches in the future. Although finding the optimal prompt for this task is intractable, we first ran a preliminary search across different prompt strategies on a validation set to determine if any of them boosts aspect prediction performance significantly.[7] We found minimal differences between prompt styles, and proceeded to the test phase. Our tests compare `Llama-3.1-8b-instruct` (Grattafiori et al., 2024) and `GPT-5mini`. We experiment with few-shot in-context learning using 3 examples per label (21 examples per prompt).

Feedforward Classifier. We investigate the ability of LLM encoder layers to capture representations that may be useful for aspect classification based on the hypothesis that contextual embeddings encode a broad range of linguistic phenomena (Arora et al., 2024). We do this by combining the token embeddings of the input sentence with a simple feedforward neural classifier to produce a label prediction from the text alone via supervised training.

[7] Details of the prompt-tuning experiments available in Appendix C.

To evaluate the usefulness of LLM embeddings out-of-the-box, we pass the natural language sentence through the encoder block of Llama-3.1 8B and average the resulting token embeddings to generate a sentence vector with standardized dimensions, then use a simple feedforward classifier head to produce a label prediction. We use the same averaged embedding as input for each event in the sentence. We train a fully-connected feedforward network to predict one of the seven aspect labels using that embedding as input. The results from this method serve as a useful benchmark for evaluating more complex strategies in future work.

6.3. Automatic Modeling Results

Table 4 displays the accuracies and F1-scores across the two LLMs and the feedforward neural classifier. We report weighted F1, average precision, and average recall; to address the imbalanced label distribution in the data, we also report macro F1. We evaluate all methods on the same stratified test set. The table also shows reported results for AutoAspect (Chen et al., 2021), a rule-based approach to UMR aspect classification. Note that the reported results use a different test set, so these serve as a general reference for rule-based approaches rather than a direct comparison. Finally, we show annotation agreement scores as an upper bound for the task.

LLM Prompting. The dataset's significant imbalance, with *State* and *NONE* labels being dominant and many classes being rare, heavily influences the results. Weighted F1 scores are skewed by the majority class, while lower macro F1 scores accurately reflect poor performance across most categories. GPT-5mini outperforms Llama across all metrics, which we attribute to architectural updates, including advanced knowledge distillation. GPT-5mini performance is relatively indifferent to in-context examples, while Llama's performance actually decreases with the addition of in-context learning, suggesting that more comprehensive training is needed for aspect prediction.

Type	Model	Acc.	Macro F1	Wtd. F1	Precision	Recall
Upper Bound	Single Human Annotator	0.84	0.76	0.84	0.77	0.82
LLM	LLaMA-3.1-8B-Instruct (zero-shot)	0.31	0.19	0.27	0.29	0.24
	LLaMA-3.1-8B-Instruct (3-shot)	0.25	0.16	0.22	0.32	0.21
	GPT-5mini (zero-shot)	**0.56**	**0.49**	0.56	**0.69**	**0.49**
	GPT-5mini (3-shot)	**0.56**	0.46	**0.60**	0.49	0.46
Neural	Feedforward MLP	0.45	0.27	0.44	0.29	0.32
Symbolic	AutoAspect[†]	0.39	0.23	0.40	—	—

Table 4: Baseline results on the test split (254 events, 72 sentences). Human performance reflects first-pass annotator accuracy against adjudicated gold labels as an upper bound on performance, against which to measure automated method results. Precision and Recall are macro averages across classes.

Feedforward Classifier. Neural classification using Llama embeddings results in middling performance, coming short in all three evaluation metrics compared to LLM prompting methods. Although sentence embeddings have been seen to capture semantic information in other tasks, these results demonstrate that embeddings alone are insufficient for capturing aspectual information.

Human Baseline. The human baseline scores show agreement between one annotator's first-pass labels for each of the events in the test set, compared against the final adjudicated labels. The success of the human baseline over the automated methods supports two conclusions: (i), the complexity of the aspect annotation task, and (ii), the need for automated methods which better utilize the sentential context and/or the inherent graphical nature of event-argument structures. We are currently developing aspectual classification models that learn from the graph structure as well as the surface form of the utterance.

7. Conclusion and Future Work

In this work, we introduce a new, carefully-annotated dataset of 1473 English sentences annotated with aspect labels within the UMR framework, achieving good agreement between annotators. We describe (and release) the annotation scheme and guidelines and detail our multi-stage annotation and adjudication process. Analysis of annotator disagreements follow expected patterns with respect to the confusability of same label pairs, motivating us to allow (and preserve) multiple labels per instance.

On the modeling side, we establish straightforward baselines for automated aspect classification using rule-based methods, embedding-based classifiers, and large language model prompting approaches. These results provide initial benchmarks for automatic UMR aspect classification, and we expect to see significant increase in model performance when we turn to more sophisticated architectures. The guidelines, dataset, stratified data splits, and initial benchmarks together lay a foundation for studying aspect in structured semantic representations and will support future work on automated UMR parsing and cross-linguistic semantic annotation.

Ethical Considerations

This work builds on existing publicly available corpora that were previously released for research purposes. Our dataset adds aspectual annotations to sentences drawn from these sources in accordance with their respective licenses. Annotation was conducted by trained researchers who are authors of this paper.

Because the dataset focuses on English sentences, it representatively only reflects information about English-language corpora and does not directly capture aspectual distinctions present in other languages. Future work will expand this annotation framework to additional languages in order to support broader cross-linguistic semantic analysis.

We do not anticipate significant risks of misuse for this dataset. However, some of the examples in our corpus were pulled from online message fora without censoring, and may contain offensive, explicit, or harmful language. The resource is intended to support research in semantic representation and natural language processing.

Limitations

We attempted to reimplement the AutoAspect rules-based classifier (Chen et al., 2021) on our novel set of annotated UMR graphs in order to compare its performance against the neural approaches as a benchmark. AutoAspect focuses on a structured set of rules which closely followed the UMR annotation guidelines and decision lattice to predict labels in a wholly deterministic method, without machine learning. However, due to dependency

issues with the semantic parser in the original AutoAspect codebase, we are unable to report this benchmark on our dataset, and instead provide the AutoAspect classifier's performance on the dataset with which it was published, as a reference for rule-based approaches in general.

Acknowledgments

This work is supported by a grant from the CNS Division of National Science Foundation (Award Number: NSF_2213805) entitled "Building a Broad Infrastructure for Uniform Meaning Representations." We extend our sincere gratitude to Julia Bonn for her invaluable insights and suggestions on the adjudication of aspectual decisions, as well as her learning materials on UMR. Thanks also to Bill Croft for his insights into edge cases on the UMR aspect lattice and other helpful advice given. Lastly, we express our appreciation to the reviewers for their helpful comments and feedback.

8. Bibliographical References

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14638–14663, Bangkok, Thailand. Association for Computational Linguistics.

Emmon Bach. 1986. The algebra of events. *Linguistics and philosophy*, pages 5–16.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet Project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Julia Bonn, Matthew J Buchholz, Jayeol Chun, Andrew Cowell, William Croft, Lukas Denk, Sijia Ge, Jan Hajic, Kenneth Lai, James H Martin, et al. 2024a. Building a broad infrastructure for uniform meaning representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2537–2547.

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

Julia Bonn, Martha Palmer, Zheng Cai, and Kristin Wright-Bettner. 2020. Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4883–4892, Marseille, France. European Language Resources Association.

Julia Bonn, Harish Tayyar Madabushi, Jena D. Hwang, and Claire Bonial. 2024b. Adjudicating LLMs as PropBank adjudicators. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024*, pages 112–123, Torino, Italia. ELRA and ICCL.

Nancy Chang, Daniel Gildea, and Srini Narayanan. 2022. A dynamic model of aspectual composition. In *Proceedings of the twentieth annual conference of the cognitive science society*, pages 226–231. Routledge.

Daniel Chen, Martha Palmer, and Meagan Vigus. 2021. AutoAspect: Automatic Annotation of Tense and Aspect for Uniform Meaning Representations. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 36–45, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jayeol Chun and Nianwen Xue. 2024. Uniform meaning representation parsing as a pipelined approach. In *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 40–52, Bangkok, Thailand. Association for Computational Linguistics.

Bernard Comrie. 1976. *Aspect: An introduction to the study of verbal aspect and related problems*, volume 2. Cambridge University Press.

William Croft. 2012. *Verbs: Aspect and causal structure*. OUP Oxford.

William Croft. 2022. Constructions of the World's Languages. In *Morphosyntax*.

Leon RA Derczynski. 2017. *Automatically ordering events and times in text*. Springer.

Lucia Donatelli, Michael Regan, William Croft, and Nathan Schneider. 2018. Annotation of tense and aspect semantics for sentential AMR. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 96–108.

David R Dowty. 1986. The effects of aspectual class on the temporal structure of discourse: semantics or pragmatics? *Linguistics and philosophy*, pages 37–61.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768.

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

William Gantt, Lelia Glass, and Aaron Steven White. 2022. Decomposing and recomposing event structure. *Transactions of the Association for Computational Linguistics*, 10:17–34.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo

Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models.

Erhard Hinrichs. 1986. Temporal anaphora in discourses of English. *Linguistics and philosophy*, pages 63–82.

Paul Kingsbury and Martha Palmer. 2003. Propbank: the next level of treebank. In *Proceedings of Treebanks and lexical Theories*, volume 3. Citeseer.

Wolfgang Klein. 2013. *Time in language*. Routledge.

Kevin Knight, Bianca Badarau, Laura Baranescu, Claire Bonial, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Tim O'Gorman, Martha Palmer, Nathan Schneider, and Madalina Bardocz. 2020. Abstract meaning representation (AMR) annotation release 3.0.

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

Klaus Krippendorff. 2004. Measuring the Reliability of Qualitative Text Analysis Data. *Quality and Quantity*, 38:787–800.

Ronald W Langacker. 2011. Remarks on English aspect. In *Tense-aspect: Between semantics & pragmatics*, pages 265–304. John Benjamins Publishing Company.

Egoitz Laparra, Dongfang Xu, and Steven Bethard. 2018. From characters to time intervals: New paradigms for evaluation and neural parsing of time normalizations. *Transactions of the Association for Computational Linguistics*, 6:343–356.

Charles N Li and Sandra A Thompson. 1989. *Mandarin Chinese: A functional reference grammar*. Univ of California Press.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Betsy Hicks McDonald. 1982. *Aspects of the American Sign Language predicate system*. State University of New York at Buffalo.

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

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd workshop on computing news storylines (CNS 2016)*, pages 47–56.

Barbara H Partee. 2011. Nominal and Temporal Semantic Structure. In *Prague Linguistic Circle Papers: Travaux du cercle linguistique de Prague nouvelle série. Volume 3*, pages 91–108. John Benjamins Publishing Company.

James Pustejovsky, Harry Bunt, and Annie Zaenen. 2017. Designing annotation schemes: From theory to model. *Handbook of Linguistic Annotation*, pages 21–72.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Hans Reichenbach. 1947. *Elements of Symbolic Logic*. Macmillan, New York.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Haibo Sun, Nianwen Xue, Jin Zhao, Liulu Yue, Yao Sun, Keer Xu, and Jiawei Wu. 2024. Chinese UMR annotation: Can LLMs help? In *Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024*, pages 131–139, Torino, Italia. ELRA and ICCL.

Jens EL Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, et al. 2021. Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, 35(3):343–360.

Herman Van Olphen. 1975. Aspect, tense, and mood in the Hindi verb. *Indo-Iranian Journal*, 16(4):284–301.

Z Vendler. 1967. Linguistics in Philosophy Ithaca, NY: Cornell Univ.

Zeno Vendler. 1957. Verbs and times. *The philosophical review*, 66(2):143–160.

Shira Wein and Julia Bonn. 2023. Comparing UMR and cross-lingual adaptations of AMR. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 23–33, Nancy, France. Association for Computational Linguistics.

9. Language Resource References

Julia Bonn, Claire Bonial, Matt Buchholz, Hsiao-Jung Cheng, Alvin Chen, Ching-wen Chen, Andrew Cowell, William Croft, Lukas Denk, Ahmed Elsayed, Eva Fučíková, Federica Gamba, Carlos Gomez, Jan Hajič, Eva Hajičová, Jiří Havelka, Loden Havenmeier, Ath Kilgore, Veronika Kolářová, Lucie Kučová, Kenneth Lai, Bin Li, Jingyi Li, Markéta Lopatková, Marie MacGregor, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Skatje Myers, Michal Novák, Tim O'Gorman, Petr Pajas, Alexis Palmer, Martha Palmer, Jarmila Panevová, Claire Benét Post, James Pustejovsky, Petr Sgall, Jialin Song, Li Song, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Haibo Sun, Yao Sun, Rosa VallejosYopán, Jens VanGysel, Meagan Vigus, Kristin Wright-Bettner, Jiawei Wu, Nianwen Xue, Dan Xing, Keer Xu, Zhixing Xu, Liulu Yue, Daniel Zeman, Jin Zhao, Šárka Zikánová, and Zdeněk Žabokrtský. 2025. Uniform meaning representation 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Julia Bonn, Chen Ching-wen, James Andrew Cowell, William Croft, Lukas Denk, Jan Hajič, Kenneth Lai, Martha Palmer, Alexis Palmer, James Pustejovsky, Haibo Sun, Rosa Vallejos Yopán, Jens Van Gysel, Meagan Vigus, Nianwen Xue, and Jin Zhao. 2023. Uniform meaning representation. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

10. Appendix A - Data Statistics and Dataset Splits

For more information on aspect data statistics, Table 5 shows general information on aspect data that was labeled in the UMR data prior to this annotation project. Next, Table 6 shows the distributions and counts at the end of Phase 1 annotation and before Phase 2 adjudication during our project. Finally, Table 7 provides statistics for our dataset splits. Precise Sentence IDs for the splits are available on Github.[8]

11. Appendix B - Annotation

Training materials. These were developed mostly from existing UMR tutorial materials and supplemented with custom task-specific resources, including an explanatory slide deck[9] which summarizes the UMR guidelines[10] with added clarifications and examples.

Practice annotation. Table 8 shows the results from the Pear Story practice annotation task. Due to the different number of annotations each person performed, we report Gwet's AC1 as a measure for inter-annotator agreement (IAA) since this metric can be calculated for different numbers of labels. We report Fleiss' Kappa only for predicates that were labeled by all annotators. We find moderate-to-good IAA for the practice round, motivating the need for additional training that was conducted.

Annotation process diagram. Figure 3 illustrates the flow of data through the two phases of corpus building, including multiple rounds of tie-breaking. The 143 events listed as single-annotated in the first-pass were part of a teaching demonstration, but they did ultimately receive second annotations and were reviewed for consistency during adjudication; this detail was omitted from the diagram for visual simplicity.

12. Appendix C - Modeling

In this Appendix we provide additional information on the automatic aspect modeling design and results.

LLM Prompt Tuning. We try three strategies to gauge the impact of prompt structure on LLM performance. Initially, we manually draft a list of short definitions for each aspect class based on the experience gained from our annotator training sessions. In a second prompt attempt, we provide the initial prompt and instruct the model to generate a better prompt for our task, to investigate if the LLM's pretraining contains aspectual knowledge beyond our basic definitions, which resulted in a streamlined version with more general task instruction. Finally, to take advantage of extensive LLM

[8] https://github.com/clairepost/UMR_Aspect_Data.git

[9] Please see our GitHub for more information.

[10] https://github.com/umr4nlp/umr-guidelines/blob/master/guidelines.md

Aspect	Little Prince	Minecraft	BOLT DF	WB	Pear Story	Lorelei	UMR 1.0	Total
State	63	20	119	45	121	0	62	430
Habitual	1	0	17	4	28	0	2	52
Process	2	0	5	5	44	1	1	58
Activity	9	0	31	14	57	0	21	132
Performance	35	2	43	18	159	3	57	317
Endeavor	0	0	0	0	2	0	14	16
Total	**110**	**22**	**215**	**86**	**411**	**4**	**157**	**1005**

Table 5: Aspect label distribution from different existing UMR datasets before any additional annotation was done.

Aspect	**Little Prince**	**Minecraft**	**BOLT DF**	**WB**	**Existing Labels**	**Total**
State	172	14	101	14	430	731
Habitual	41	0	4	0	52	97
Process	31	0	38	8	58	135
Activity	15	2	10	3	132	162
Performance	163	43	69	32	317	624
Endeavor	15	0	4	0	16	35
None	158	49	121	77	-	405
Total	**595**	**108**	**347**	**134**	**1,005**	**2,189**
Fleiss' Kappa	0.78	0.82	0.45	0.40	-	-

Table 6: Label distribution by corpus and annotated aspect. We report Fleiss' Kappa between the two initial annotators and do not include disagreements in the reported total.

Split	**Sentences**	**Events**
Train	333	999
Dev	71	220
Test	72	254
Total	**476**	**1,473**

Table 7: Stratified 70/15/15 train/dev/test split, divided at the sentence level using dominant aspect label for stratification.

context windows, we try providing the UMR guidelines for aspect[11] in their entirety and instructing the model to predict a label. We find marginally higher validation accuracy with the second strategy, and employ it in testing. We provide the full prompt we used in testing in Table 9, as well as in our GitHub repository.

Category	Metric	Value
Accuracy	State	0.82
	Habitual	0.63
	Activity	0.52
	Performance	0.80
	Endeavor	0.11
	Overall Accuracy	0.74
	Perfect Accuracy	0.35
F1	Macro F1	0.49
	Weighted Macro F1	0.76
IAA	Fleiss' Kappa	0.55
	Gwet's AC1	0.66

Table 8: Practice Annotation Results: *Overall Accuracy* is the ratio of the total number of correct annotations over the total number of predicates annotated. *Perfect Accuracy* is the ratio of predicates that were correctly annotated by all annotators. No occurrences of *Process* aspects were present in the practice set.

[11] github.com/umr4nlp/umr-guidelines/blob/master/guidelines.md#part-3-3-1-Aspect

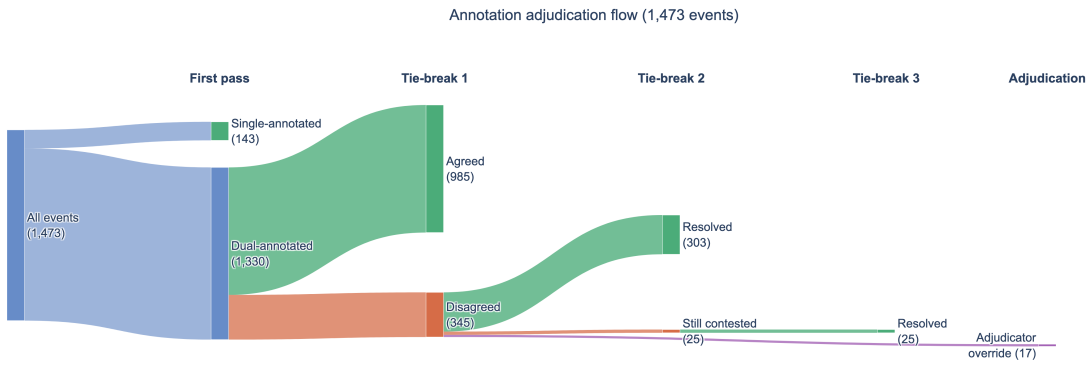

Figure 3: Flow diagram illustrating the annotation and adjudication phases, and how many labels were completed at each stage of the process.

Setting	Prompt
Experiment: Few-shot with definitions (3 examples per class, total 21 examples) **Model:** `gpt-5-mini` **Seed:** 42	**[SYSTEM TURN]** You are a UMR expert, required to predict the aspectual value of a predicate in a given sentence. The goal is to annotate the aspect of each predicate, which can be one of six distinct values: State, Habitual, Process, Activity, Endeavor, Performance, or a seventh 'None' option if you think the given predicate is not an event. Respond with only the label name and nothing else. Do not restate the sentence, predicate, or provide any reasoning. **[USER TURN]** **Definitions:** • state: stable condition or property (e.g. "knows", "believes") • habitual: recurring or generic action (e.g. "walks to school every day") • process: ongoing activity without clear endpoint (e.g. "is running") • activity: dynamic event (e.g. "built a house") • performance: bounded event with natural endpoint (e.g. "arrived") • endeavor: attempted action that may not fully complete (e.g. "tried to open") • none: no clear aspectual reading **Examples:** Sentence: "The new immigration law also permits the immigration service to provide " limited " biometric information on New Zealand citizens to neighboring countries ." | Predicate: "immigrate-01" -> none Sentence: "And as I had with me neither a mechanic nor any passengers , I set myself to attempt the difficult repairs all alone ." | Predicate: "attempt-01" -> endeavor Sentence: "long hours and lots of long nights." | Predicate: "long-03" -> none Sentence: "I am about to receive a visit from an admirer ! " he exclaimed from afar , when he first saw the little prince coming ." | Predicate: "come-01" -> activity Sentence: "However, I just have to say that I think it is ludicrous for John to accept any other explanation for his encounter except for what it is." | Predicate: "contrast-91" -> none Sentence: "" Yes ? " said the little prince , who did not understand what the conceited man was talking about ." | Predicate: "talk-01" -> activity Sentence: "At a glance I can distinguish China from Arizona ." | Predicate: "glance-01" -> endeavor Sentence: "" Yes ? " said the little prince , who did not understand what the conceited man was talking about ." | Predicate: "understand-01" -> state Sentence: "" What ! "" | Predicate: "say-91" -> performance Sentence: "" This man , " the little prince said to himself , " reasons a little like my poor tippler ... "" | Predicate: "reason-01" -> habitual Sentence: "Freedom of speech has never been understood to restrict the ability of people to sue other people for things like libel and slander." | Predicate: "restrict-01" -> state Sentence: "In the course of this life I have had a great many encounters with a great many people who have been concerned with matters of consequence ." | Predicate: "encounter-01" -> habitual Sentence: "And as I had with me neither a mechanic nor any passengers , I set myself to attempt the difficult repairs all alone ." | Predicate: "repair-01" -> endeavor Sentence: "Absurd as it might seem to me , a thousand miles from any human habitation and in danger of death , I took out of my pocket a sheet of paper and my fountain - pen ." | Predicate: "die-01" -> process Sentence: "" I admire you , " said the little prince , shrugging his shoulders slightly , " but what is there in that to interest you so much ? "" | Predicate: "shrug-01" -> activity Sentence: "" It is to raise in salute when people acclaim me ." | Predicate: "raise-01" -> habitual Sentence: "If I owned a flower , I could pluck that flower and take it away with me ." | Predicate: "possible-01" -> state Sentence: "This higher health spending is a function of different prices and different usage of medical care." | Predicate: "use-01" -> process Sentence: "[Builder puts down a green block at X:1 Y:2 Z:0]" | Predicate: "put-down-17" -> performance Sentence: "So I lived my life alone , without anyone that I could really talk to , until I had an accident with my plane in the Desert of Sahara , six years ago ." | Predicate: "talk-01" -> process Sentence: "Here you may see the best portrait that , later , I was able to make of him ." | Predicate: "make-01" -> performance **Classify:** Sentence: "My mom just retired." Predicate: "retire-01" Label: **[ASSISTANT TURN]**

Table 9: Few-shot prompt used for LLM aspect classification.