Decomposing and Recomposing Event Structure

William Gantt
University of Rochester

Lelia Glass
Georgia Institute of Technology

Aaron Steven White
University of Rochester

Abstract

We present an event structure ontology empirically derived from inferential properties annotated on sentence- and document-level semantic graphs. We induce this ontology jointly with semantic role, entity type, and event-event relation ontologies using a document-level generative model, identifying sets of types that align closely with previous theoretically-motivated taxonomies.

1 Introduction

Natural language provides myriad ways of communicating about complex events. For instance, one and the same event can be described at a coarse grain, using a single clause (1), or at a finer grain, using an entire document (2).

(1) The contractor built the house.

(2) They started by laying the house's foundation. They then framed the house before installing the plumbing. After that[...]

Further, descriptions of the same event at different granularities can be interleaved within the same document—e.g. (2) might well directly follow (1) as an elaboration on the house-building process.

Consequently, extracting knowledge about complex events from text involves determining the structure of the events being referred to: what their parts are, how those parts are laid out in time, who participates in them and how, etc. Determining this structure requires an event ontology whose elements are (associated with) event structure representations. A number of such ontologies and annotated corpora exist: FrameNet (Baker et al., 1998), VerbNet (Kipper Schuler, 2005), PropBank (Palmer et al., 2005), Abstract Meaning Representation (Banarescu et al., 2013), and Universal Conceptual Cognitive Annotation (Abend and Rappoport, 2013) among others discussed in §2.

Our aim in this paper is similar in spirit to this prior work, but it differs in approach: while many of the event structure representations posited by the above ontologies are empirically *inspired*, none are empirically *derived*. We present an event structure ontology that is empirically derived from a wide variety of theoretically informed semantic properties that are relevant to event structure and annotated on a set of genre-diverse documents. The properties on which our categories rest target (i) the substructure of an event—e.g. that the building described in (1) consists in a sequence of subevents resulting in the creation of some artifact; (ii) the superstructure in which an event takes part—e.g. that laying a house's foundation is part of building a house, alongside framing the house, installing the plumbing, etc.; (iii) the relationship between an event and its participants—e.g. that the contractor in (1) causes the building and the house is created as a result; and (iv) properties of the event's participants—e.g. that the contractor in (1) is animate while the house is not.

A subset of the properties we use to derive our event structure ontology are already annotated in the Universal Decompositional Semantics dataset (UDS; White et al., 2016, 2020), but a a range of key event structure properties remain to be captured. After motivating the need for these properties (§2), we develop annotation protocols (§3) within the Universal Decompositional Semantics framework. We validate our protocols (§4) and use them to collect annotations for the entire Universal Dependencies (Nivre et al., 2016) English Web Treebank (§5; Bies et al. 2012), resulting in the UDS-EventStructure dataset (UDS-E). To derive an event structure ontology from UDS-E and existing UDS annotations, we develop a document-level generative model that jointly induces event types, entity types, semantic role types, and event-event relation types (§6), and we compare these types to those found in existing event structure ontologies (§7). We make UDS-E and our derived ontologies available at decomp.io.

2 Background

Modern theoretical treatments of event structure tend to take as their starting point Vendler's (1957) seminal four-way classification. We briefly discuss this classification and elaborations thereon before turning to event structure ontologies developed for annotating corpora.[1] We then contrast these ontologies with the fully decompositional approach we take in this paper.

Theoretical Approaches Vendler categorizes event descriptions into four classes: *statives* (3), *activities* (4), *achievements* (5), and *accomplishments* (6). As theoretical constructs, these classes are used to explain both the distributional characteristics of event descriptions as well as inferences about how an event progresses over time.

(3) Jo was in the park.
stative = [+DUR, −DYN, −TEL]

(4) Jo ran around in the park.
activity = [+DUR, +DYN, −TEL]

(5) Jo arrived at the park.
achievement = [−DUR, +DYN, +TEL]

(6) Jo ran to the park.
accomplishment = [+DUR, +DYN, +TEL]

Work building on Vendler's discovered that these classes can be decomposed into the now well-accepted component properties in (7)–(9) (Kenny, 1963; Lakoff, 1965; Verkuyl, 1972; Bennett and Partee, 1978; Mourelatos, 1978; Dowty, 1979).

(7) DUR(ATIVITY): whether the event happens at an instant or extends over time

(8) DYN(AMICITY): whether the event involves change, broadly construed

(9) TEL(ICITY): whether the event culminates in a participant changing state or location, being created or destroyed, etc.

Later work further expanded these properties and, therefore, the possible classes. Expanding on DYN, Taylor (1977) suggests a distinction between dynamic predicates that refer to events with dynamic subevents—e.g. the individual strides in a running—and ones that do not—e.g. the gliding in (10) (see also Bach, 1986; Smith, 2003).

(10) The pelican glided through the air.

Dynamic events with dynamic subevents can be further distinguished based on whether the subevents are similar—e.g. the strides in a running—or dissimilar—e.g. the subevents in a house-building (Piñón, 1995). In the case where the subevents are similar and a participant itself has subparts—e.g. when the participant is a group—there may be a bijection from participant subparts to subevents. In (11), there is a smiling for each child that makes up the composite smiling—*smile* is *distributive*. In (12), the meeting presumably has some structure, but there is no bijection from members to subevents—*meet* is *collective* (see Champollion, 2010, for a review).

(11) {The children, Jo and Bo} smiled.

(12) {The committee, Jo and Bo} met.

Expanding on TEL, Dowty (1991) argues for a distinction among telics in which the culmination comes about incrementally (13) or abruptly (14) (see also Tenny, 1987; Krifka, 1989, 1992, 1998; Levin and Hovav, 1991; Rappaport Hovav and Levin, 1998, 2001; Croft, 2012).

(13) The gardener mowed the lawn.

(14) The climber summitted at 5pm.

This notion of incrementality is intimately tied up with the notion of DUR(ATIVITY). For instance, Moens and Steedman (1988) point out that certain event structures can be systematically transformed into others—e.g. whereas (14) describes the summitting as something that happens at an instant (and is thus abrupt), (15) describes it as a process that culminates in having reach the top of the mountain (see also Pustejovsky, 1995).

(15) The climber was summitting.

Such cases of *aspectual coercion* highlight the importance of grammatical factors in determining the structure of an event. More general contextual factors are also at play when determining event structure: *I ran* can describe a telic event—e.g. when it is known that I run the same distance or to the same place every day—or an atelic event—e.g. when the destination and/or distance is irrelevant in context (Dowty, 1979; Olsen, 1997). This context-sensitivity strongly suggests that annotating event structure is not simply a matter of building a type-level lexical resource and projecting its labels; actual text must be annotated.

Event Structure Ontologies Early, broad-coverage lexical resources, such as the Lexical Conceptual Structure lexicon (LCS; Dorr, 1993), attempt to directly encode an elaboration of the core Vendler classes in terms of a hand-engineered graph representation proposed by Jackendoff (1990). VerbNet (Kipper Schuler, 2005) further elaborates on LCS by building on

[1] The theoretical literature on event structure is truly vast. See Truswell 2019 for a collection of overview articles.

the fine-grained syntax-based classification of Levin (1993) and links her classes to LCS-like representations. More recent versions of VerbNet (v3.3+; Brown et al., 2018) update these representations to ones based on the Dynamic Event Model (Pustejovsky, 2013).

COLLIE-V, which expands the TRIPS lexicon and ontology (Ferguson and Allen 1998 *et seq*), takes a similar tack of producing hand-engineered event structures, combining this hand-engineering with a procedure for bootstrapping event structures (Allen et al., 2020). FrameNet also contains hand-engineered event structures, though they are significantly more fine-grained than those found in LCS or VerbNet (Baker et al., 1998).

VerbNet, COLLIE-V, and FrameNet are not directly annotated on text, though annotations for at least VerbNet and FrameNet can be obtained by using SemLink to project FrameNet and VerbNet annotations onto PropBank annotations (Palmer et al., 2005). PropBank frames have been enriched in a variety of other ways. One such enrichment can be found in Abstract Meaning Representation (AMR; Banarescu et al., 2013; Donatelli et al., 2018). Another can be found in Richer Event Descriptions (RED; O'Gorman et al., 2016), which annotates events and entities for factuality (whether an event actually happened or is hypothetical) and genericity (whether an event/entity is a particular or generic) as well as annotating for causal, temporal, sub-event, and co-reference relations between events (see also Chklovski and Pantel, 2004; Hovy et al., 2013).

Additional less fine-grained event ontologies exist in TimeBank (Pustejovsky et al., 2006), Universal Conceptual Cognitive Annotation (UCCA; Abend and Rappoport, 2013), and the Situation Entities dataset (SitEnt; Friedrich and Palmer, 2014b; Friedrich et al., 2016). Of these, the closest to capturing the standard Vendler classification and decompositions thereof is SitEnt. The original version of SitEnt annotates only for a state-event distinction (alongside related, non-event structural distinctions), but later elaborations further annotate for telicity (Friedrich and Gateva, 2017). Because of this close alignment to the standard Vendler classes, we use SitEnt annotations as part of validating our own annotation protocol in §3.

Universal Decompositional Semantics In contrast to the hand-engineered event structure ontologies discussed above, our aim is to derive event structure representations directly from semantic annotations. To do this, we extend the existing annotations in the Universal Decompositional Semantics dataset (UDS; White et al. 2016, 2020) with key annotations for the event structural distinctions discussed above. Our aim is not necessarily to reconstruct any of the ontologies discussed above, though we do find in §6 that our event type ontology approximates Vendler's.

UDS consists of two layers of annotations on top of the Universal Dependencies (UD) syntactic graphs in the English Web Treebank (EWT): (i) predicate-argument graphs with mappings into the syntactic graphs, derived using the PredPatt tool (White et al., 2016; Zhang et al., 2017); and (ii) crowd-sourced annotations for properties of events (on the *predicate nodes* of the predicate-argument graph), entities (on the *argument nodes*), and their relationship (on the *predicate-argument edges*). These properties are organized into three *predicate subspaces* with five properties:

- FACTUALITY (Rudinger et al., 2018)
  factual: did the event happen?
- GENERICITY (Govindarajan et al., 2019)
  kind: is the event generic?
  hypothetical: is the event hypothetical?
  dynamic: is the event dynamic or stative?
- TIME (Vashishtha et al., 2019)
  duration: how long did/will the event last?

Two *argument subspaces* with four properties:

- GENERICITY (Govindarajan et al., 2019)
  particular: is the entity a particular?
  kind: is the entity a kind?
  abstract: is the entity abstract or concrete?
- WORDSENSE (White et al., 2016)
  Which coarse entity types (WordNet supersense) does the entity have?

And one *predicate-argument subspace* with 16 properties (see White et al. 2016 for full list):

- PROTOROLES (Reisinger et al., 2015)
  instigation: did participant cause event?
  change of state: did participant change state during or as a consequence of event?
  change of location: did participant change location during event?
  existed {before, during, after} did participant exist {before, during, after} the event?

The properties cover a variety of event structural distinctions relevant to those discussed above—e.g. *dynamicity*, *telicity* (in the form of *change of* *state*, *change of location*, and *existed {before, during, after}*), and *durativity*. But they fail to capture other core distinctions.

3 Annotation Protocol

We annotate for the core event structural distinction not currently covered in UDS, breaking our annotation into three subprotocols. For all questions, annotators report confidence in their response to each question on a scale from 1 (*not at all confident*) to 5 (*totally confident*).

Event-subevent Annotators are presented with a sentence containing a single highlighted predicate followed by four questions about the internal structure of the event it describes. Question (1) asks whether the event described by the highlighted predicate has natural subparts. Question (2) asks whether the event has a natural endpoint.

The final questions depend on the response to (1). If an annotator responds that the highlighted predicate refers to an event that *has* natural parts, they are asked (i.a) whether the parts are similar to one another and (ii.a) how long each part lasts on average. If an annotator instead responds that the event referred to does *not* have natural parts, they are asked (i.b) whether the event is dynamic, using the same prompt as Govindarajan et al. (2019), and (ii.b) how long the event lasts.

All questions are binary except those concerning duration, for which answers range from *effectively no time at all* to *effectively forever* (see Figure 3 for list). Together, these questions target the three Vendler-inspired features (DYN, DUR, TEL), plus a fourth dimension for subtypes of dynamic predicates. In the context of UDS, they form a predicate node subspace, alongside FACTUALITY, GENERICITY, and TIME.

Event-event Annotators are presented with either a single sentence or a pair of adjacent sentences, with the two predicates of interest highlighted in distinct colors. For a predicate pair (p_1, p_2) describing an event pair (e_1, e_2), annotators are asked whether e_1 is a (potentially improper) mereological part of e_2, and vice versa. Both questions are binary: a positive response to both indicates that e_1 and e_2 are the same event; and a positive response to exactly one of the questions indicates proper parthood. In the context of UDS, these subevent properties form a novel predicate-predicate edge subspace whose edges potentially cross sentence boundaries.

This subprotocol targets generalized event coreference, identifying *constituency* in addition to strict identity. It also augments the information collected in the event-subevent protocol: insofar as a proper subevent relation holds between e_1 and e_2, we obtain additional fine-grained information about the subevents of the containing event—e.g. an explicit description of at least one subevent.

Event-entity The final subprotocol focuses on the relation between the event described by a predicate and its plural or conjoined arguments, asking whether the predicate is distributive or collective with respect to that argument. In the context of UDS, this property forms a predicate-argument edge subspace, alongside PROTOROLES.

4 Validation Experiments

We validate our annotation protocol (i) by assessing interannotator agreement (IAA) among both experts and crowd-sourced annotators for each subprotocol on a small sample of items drawn from existing annotated corpora (§4.1-4.2); and (ii) by comparing annotations generated using our protocol against existing annotations that cover (a subset of) the phenomena that ours does and are generated by highly trained annotators (§4.3).

4.1 Item Selection

For each of the three subprotocols, one of the authors selected 100 sentences for inclusion in the pilot for that subprotocol. This author did not consult with the other authors on their selection, so that annotation could be blind.

For the event-subevent subprotocol, the 100 sentences come from the portion of the MASC corpus (Ide et al., 2008) that Friedrich et al. (2016) annotate for eventivity (EVENT V. STATE) and that Friedrich and Gateva (2017) annotate for telicity (TELIC V. ATELIC). For the event-event subprotocol, the 100 sentences come from the portions of the Richer Event Descriptions corpus (RED; O'Gorman et al., 2016) that are annotated for event subpart relations. To our knowledge, no existing annotations cover distributivity, and so for our event-entity protocol, we select 100 sentences (distinct from those used for the event-subevent subprotocol) and compute IAA, but do not compare against existing annotations.

	Annotation	Count (%)	Example
Event-internal	Has natural parts	6,903 (23%)	The eighteen steps of the dance are done rhythmically
	Parts similar	4,498 (15%)	Israel resumed its policy of targeting militant leaders
	Parts dissimilar	2,158 (7%)	Fish are probably the easiest to take care of
	(Part duration)	(–)	(ordinal; see Figure 3)
	No natural parts	23,069 (77%)	It had better nutritional value
	Dynamic	13,903 (48%)	I would like to informally get together with you
	Not dynamic	8,839 (29%)	I assume this is 12:30 Central Time?
	(Full duration)	(–)	(ordinal; see Figure 3)
	Natural endpoint	6,031 (20%)	I will deliver it to you
	No natural endpoint	23,941 (80%)	If you know or work there could you enlighten me?
	total	29,984	(all event descriptions)
Event-event	P1, P2 identical	2,435 (6%)	All horses [...] are happy$_1$ & healthy$_2$ when they arrive
	P1, P2 disjoint	30,247 (80%)	I am often stopped$_1$ on the street and asked, 'Who does your hair ...I LOVE$_2$ it'
	P1 $\subset$ P2	1,832 (5%)	The office is shared with a foot doctor and it's very sterile$_1$ and medical feeling$_2$, which I liked
	P2 $\subset$ P1	3,029 (8%)	It is a very cruel death$_1$ with bodies dismembered$_2$
	total	37,719	(pairs of event descriptions w/ temporal overlap)
Ev-particip.	Distributive	4,812 (50%)	the **pics** turned out ok
	Collective	4,876 (50%)	**we** draw on our many faith traditions to arrive at a common conviction
	total	9,710	(event descriptions with plural arguments)

Table 1: Descriptive statistics and examples from Train and Dev data. Each item was annotated by a single annotator in Train; and by three annotators in Dev, of which this table reports the majority opinion.

4.2 Interannotator Agreement

We compute two forms of IAA: (i) IAA among expert annotators (the three authors); and (ii) IAA between experts and crowd-sourced annotators. In both cases, we use Krippendorff's α as our measure of (dis)agreement (Krippendorff, 2004). For the binary responses, we use the nominal form of α; for the ordinal responses, we use the ordinal.

Expert Annotators For each subprotocol, the three authors independently annotated the 100 sentences selected for that subprotocol.

Prior to analysis, we ridit score the confidence ratings by annotator to normalize them for differences in annotator scale use (see Govindarajan et al. 2019 for discussion of ridit scoring confidence ratings in a similar annotation protocol). This method maps ordinal labels to (0, 1) on the basis of the empirical CDF of each annotator's responses—with values closer to 0 implying lower confidence and those nearer 1 implying higher confidence. For questions that are dynamically revealed on the basis of the answer to the *natural parts* question—i.e. *part similarity*, *average part duration*, *dynamicity*, and *situation duration*—we use the average of the ridit scored confidence for *natural parts* and that question.

Figure 1 shows α when including only items that the expert annotators rated with a particular ridit scored confidence or higher. The agreement for the event-event protocol (mereology) is given in two forms: given that e_1 temporally contains e_2, (i) *directed*: the agreement on whether e_2 is a subevent of e_1; and (ii) *undirected*: the agreement on whether e_2 is a subevent of e_1 *and* whether e_1 is a subevent of e_2.

The error bars are computed by a nonparametric bootstrap over items. A threshold of 0.0 corresponds to computing α for all annotations, regardless of confidence; a threshold of $t > 0.0$ corresponds to computing α only for annotations associated with a ridit scored confidence of greater than t. When this thresholding results in less than $\frac{1}{3}$ of items having an annotation for at least two annotators, α is not plotted. This situation occurs only for questions that are revealed based on the answer to a previous question.

For *natural parts*, *telicity*, *mereology*, and *distributivity*, agreement is high, even without filtering any responses on the basis of confidence, and that agreement improves with confidence. For *part similarity*, *average part duration*, and *situation duration*, we see more middling, but still reasonable, agreement, though this agreement does not reliably increase with confidence. The fact that it does not increase may have to do with interactions between confidence on the *natural parts* question and its dependent questions that we do not capture by taking the mean of these two confidences.

Crowd-Sourced Annotators We recruit crowd-sourced annotators in two stages. First, we se-

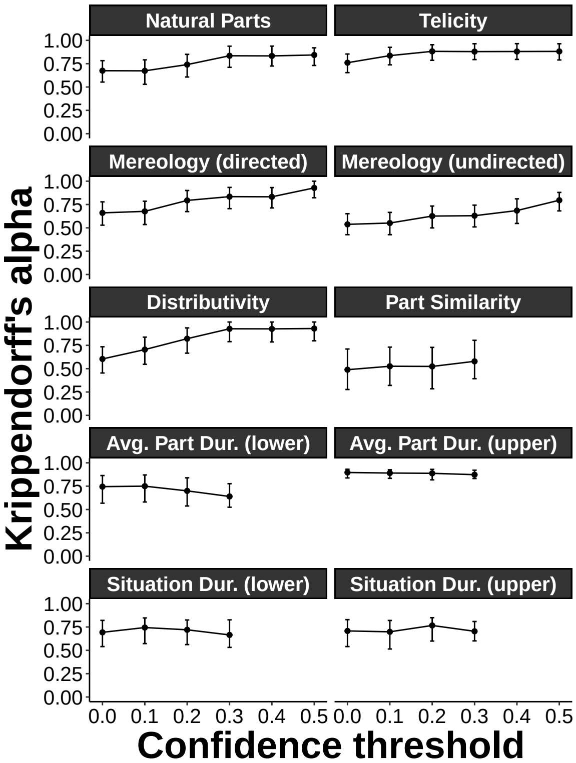

Figure 1: IAA among experts for each property, filtering annotations with ridit-scored confidence ratings below different thresholds. Confidence threshold 0.0 implies no filtering. Errors bars show 95% confidence internals computed by a nonparametric bootstrap.

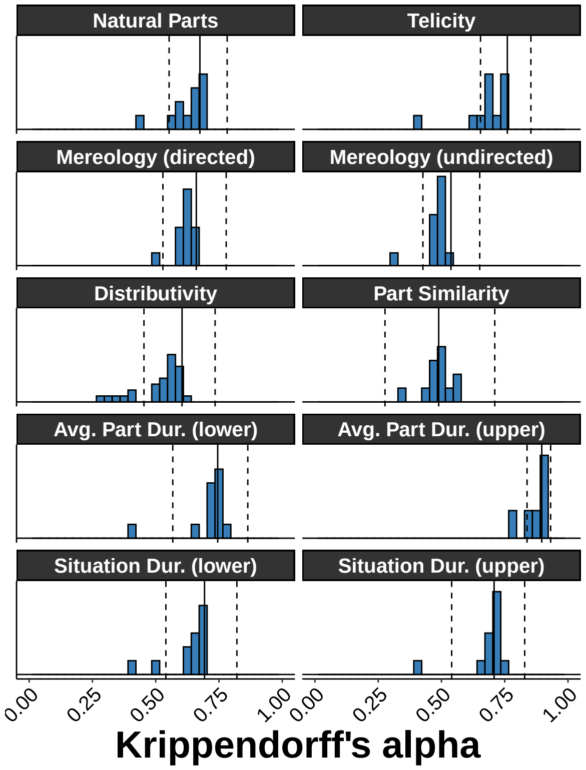

Figure 2: IAA between crowd-sourced annotators and experts for each property at confidence threshold 0 (no filtering). Solid vertical lines show IAA among experts; dashed vertical lines show 95% confidence interval for IAA among experts (see Figure 1).

lect a small set of items from the 100 we annotate in the expert annotation that have high agreement among experts to create a qualification task. Second, based on performance in this qualification task, we construct a pool of trusted annotators who are allowed to participate in pilot annotations for each of the three subprotocols.

Qualification For the qualification task, we selected eight of the sentences collected from MASC for the event-subevent subprotocol on which expert agreement was very high and which were diverse in the types of events described. We then obtained event-subevent annotations for these sentences from 400 workers on Amazon Mechanical Turk (AMT), and selected the top 200 among them on the basis of their agreement with expert responses on the same items. These workers were then permitted to participate in the pilot tasks.

Pilot We conducted one pilot for each subprotocol, using the items described in §4.1. Sentences were presented in lists of 10 per Human Intelligence Task (HIT) on AMT for the event-event and event-entity subprotocols and in lists of 5 per HIT for event-subevent. We collected annotations from 10 distinct workers for each sentence, and workers were permitted to annotate up to the full 100 sentences in each pilot. Thus, all pilots were guaranteed to include a minimum of 10 distinct workers (all workers do all HITs), up to a maximum of 100 for the subprotocols with 10 sentences per HIT or 200 for the subprotocol with 5 per HIT (each worker does one HIT). All top-200 workers from the qualification were allowed to participate.

Figure 2 shows IAA between all pilot annotators and experts for individual questions across the three pilots. More specifically, it shows the distribution of α scores by question for each annotator when IAA is computed among the three experts and that annotator only. Solid vertical lines show the expert-only IAA and dashed vertical lines show the 95% confidence interval.

4.3 Protocol Comparison

To further validate the event-event and event-subevent subprotocols, we evaluate how well our pilot data predicts the corresponding CONTAINS v. CONTAINS-SUBEVENT annotations from RED in the former case, as well as the EVENT v. STATE and TELIC v. ATELIC annotations from SitEnt in the latter. In both cases, we used the (ridit-scored)

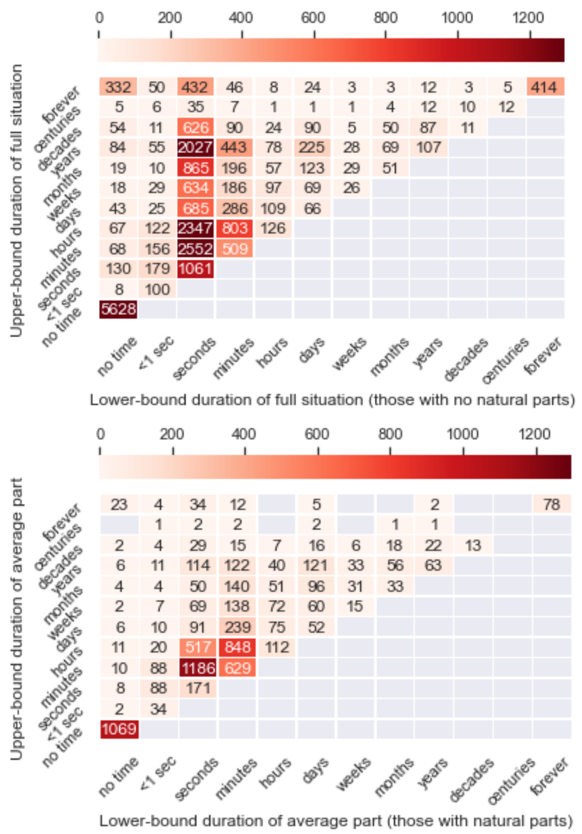

Figure 3: Lower- and upper-bound duration annotations, for the entire situation (those without natural parts) and for each part (those with natural parts).

confidence-weighted average response across annotators for a particular item as features in a simple SVM classifier with linear kernel. In a leave-one-out cross-validation on the binary classification task for RED, we achieve a micro-averaged F1 score of 0.79—exceeding the reported human F1 agreement for both the CONTAINS (0.640) and CONTAINS-SUBEVENT (0.258) annotations reported by O'Gorman et al. (2016).

For SitEnt, we evaluate on a three-way classification task for STATIVE, EVENTIVE-TELIC, and EVENTIVE-ATELIC, achieving a micro-averaged F1 of 0.68 using the same leave-one-out cross-validation. As Friedrich and Palmer (2014a) do not report interannotator agreement for this class breakdown, we further compute Krippendorff's alpha from their raw annotations and again find that agreement between our predicted annotations and the gold ones (0.48) slightly exceeds the interannotator agreement among humans (0.47).

These results not only suggest that our subprotocols effectively capture the relevant event structural phenomena, but that they may even serve as effective alternatives to these other protocols while not requiring any linguistic expertise.

5 Corpus Annotation

We collect crowd-sourced annotations for the entirety of UD-EWT. Predicate and argument spans are obtained from the PredPatt predicate-argument graphs for UD-EWT available in UDS1.0. The total number of items annotated for each subprotocol is presented in Table 1.

Event-subevent These annotations cover all predicates headed by verbs (as identified by UD POS tag), as well as copular constructions with nominal and adjectival complements. In the former case, only the verb token is highlighted in the task; in the latter, the highlighting spans from the copula to the complement head.

Event-event Pairs for the event-event subprotocol were drawn from the UDS-Time dataset, which features pairs of verbal predicates, either within the same sentence or in adjacent sentences, each annotated with its start- and endpoint relative to the other. We additionally included predicate-argument pairs in cases where the argument is annotated in UDS as having a WordNet supersense of EVENT, STATE, or PROCESS.

Event-entity For the event-entity subprotocol, we identify predicate-argument pairs in which the argument is either plural or conjoined. Plural arguments are identified by the UD NUMBER attribute, and conjoined ones by the `conj` dependency relation between an argument head and another noun. We consider only predicate-argument pairs with a UD dependency label of `nsubj`, `nsubjpass`, `dobj`, or `iobj`.

6 Event Structure Induction

Our goal in inducing event structural categories is to learn representations of those categories on the basis of annotated UDS graphs. We aim to learn four sets of interdependent ontologies grounded in UDS properties: event types, entity types, semantic role types, and event-event relation types. These ontologies are interdependent in that we assume a generative model that incorporates both sentence- and document-level structure.[2]

Document-level UDS As it stands, semantics edges in UDS1.0 only represent sentence-internal semantic relations. This constraint implies that annotations for cross-sentential semantic relations—a significant subset of our event-event

[2] See Ferraro and Van Durme 2016 for a related model that uses FrameNet's ontology, rather than inducing the ontology.

Initialize queue I ;
for *sentence* $s \in \mathcal{S}$ **do**
  Initialize queue J ;
  Enqueue $J \to I$;
  if $\text{length}(I) > W$ **then**
    Dequeue I
  for *predicate node* $v \in \text{predicates}(s)$ **do**
    Sample event type $t_{sv} \sim \text{Cat}\left(\boldsymbol{\theta}^{(\text{event})}\right)$;
    for *property* $p \in \mathcal{P}_{\text{event}}$ **do**
      for *annotator* $i \in \mathcal{A}^{(\text{event})}_{svp}$ **do**
        Sample $x^{(\text{event})}_{svpi} \sim f^i_p\left(\boldsymbol{\mu}^{(\text{event})}_{t_{sv}}\right)$
    Enqueue $\langle s, v\rangle \to J$;
    for *argument node* $v' \in \text{arguments}(s, v)$ **do**
      Sample ent. type $t_{sv'} \sim \text{Cat}\left(\boldsymbol{\theta}^{(\text{entity})}\right)$;
      for *property* $p \in \mathcal{P}_{\text{ent}}$ **do**
        for *annotator* $i \in \mathcal{A}^{(\text{ent})}_{sv'p}$ **do**
          Sample $x^{(\text{part})}_{sv'pi} \sim f^i_p\left(\boldsymbol{\mu}^{(\text{part})}_{t_{sv'}}\right)$
      if v' *is eventive* **then**
        Enqueue $\langle s, v'\rangle \to J$;
      Sample role type $r_{svv'} \sim \text{Cat}\left(\boldsymbol{\theta}^{(\text{role})}_{t_{sv}t_{sv'}}\right)$;
      for *property* $p \in \mathcal{P}_{\text{role}}$ **do**
        for *annotator* $i \in \mathcal{A}^{(\text{role})}_{svv'p}$ **do**
          Sample $x^{(\text{role})}_{svv'pi} \sim f^i_p\left(\boldsymbol{\mu}^{(\text{role})}_{r_{svv'}}\right)$
    for *index pair* $\langle s', v'\rangle \in \text{flatten}(I)$ **do**
      Sample rel. type $q \sim \text{Cat}\left(\boldsymbol{\theta}^{(\text{rel})}_{t_{sv}t_{s'v'}}\right)$;
      for *property* $p \in \mathcal{P}_{\text{rel}}$ **do**
        for *annotator* $i \in \mathcal{A}^{(\text{rel})}_{svs'v'p}$ **do**
          Sample $x^{(\text{rel})}_{svs'v'pi} \sim f^i_p\left(\boldsymbol{\mu}^{(\text{rel})}_q\right)$

Algorithm 1: Generative story of event structure induction model for a single document with sentence window W

annotations—cannot currently be represented in the graph structure. To remedy this, we extend UDS1.0 by adding *document edges* that connect semantics nodes either within a sentence or in two distinct sentences, and we associate our event-event annotations with their corresponding document edge. Because UDS1.0 does not have a notion of document edge, it does not contain Vashishtha et al.'s (2019) fine-grained temporal relation annotations. We additionally add those attributes to their corresponding document edges.

Generative Model Algorithm 1 gives the generative story for our event structure induction model. We assume some fixed numbers of event types $\mathcal{T}_{\text{event}}$, role types $\mathcal{R}_{\text{role}}$, entity types $\mathcal{T}_{\text{ent}}$, and relation types $\mathcal{R}_{\text{rel}}$.

Annotation Likelihoods The distribution f^a_p on the annotations themselves is implemented as a mixed model dependent on the property p being annotated with annotator random intercepts $\boldsymbol{R}$, where the random intercepts for annotator a are $\boldsymbol{\rho}_a \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{ann}})$ with unknown $\boldsymbol{\Sigma}_{\text{ann}}$. When p receives binary annotations, a simple logistic mixed model is assumed, where $f^a_p = \text{Bern}(\text{logit}^{-1}(\mu_{i_p} + \rho_{ai_p}))$ and i_p is the index corresponding to property p in the expected annotation $\boldsymbol{\mu}$. When p receives nominal annotations, $f^a_p = \text{Cat}(\text{softmax}(\boldsymbol{\mu}_{i_p} + \boldsymbol{\rho}_{ai_p}))$ and i_p is a set with cardinality of the number of nominal categories. And when p receives ordinal annotations, we follow (White et al., 2020) in using an ordinal (linked logit) mixed effects model where $\boldsymbol{\rho}_a$ defines the cutpoints between response values in the cumulative density function for annotator a:

$$\mathbb{P}(r_{ai_p} \leq j) = \text{logit}^{-1}(\mu_{i_p} - \rho_{ai_p})$$
$$f^a_p(r_{ai_p} = j) = \mathbb{P}(r_{ai_p} \leq j) - \mathbb{P}(r_{ai_p} \leq j - 1)$$

Conditional Properties For both the UDS-EventStructure and UDS-Protoroles protocols, certain annotations are conditioned on others, owing to the fact that whether some questions are asked at all depends upon annotator responses to previous ones. Following (White et al., 2017), we model the likelihoods for these properties using hurdle models (Agresti, 2014): for a given property, a Bernoulli distribution of the same form as above determines whether the property applies; if it does, the property value is determined using a second distribution of the appropriate type.

Temporal Relations Temporal relations annotations from UDS-Time consist of 4-tuples $(\overleftarrow{e_1}, \overleftarrow{e_2}, \overrightarrow{e_1}, \overrightarrow{e_2})$ of real values on the unit interval, representing start- and endpoints of two event-referring predicates or arguments, e_1 and e_2. Each tuple is normalized such that the earlier of $(\overleftarrow{e_1}, \overleftarrow{e_2})$ is always locked to the left end of the scale (0) and the later of $(\overrightarrow{e_1}, \overrightarrow{e_2})$ to the right end (1). The likelihood for these annotations must consider the different possible orderings of the two events. To do so, we first determine whether $\overleftarrow{e_1}$ is locked, $\overleftarrow{e_2}$ is, or both are, according to $\text{Cat}\left(\text{softmax}(\boldsymbol{\mu}_{\text{lock}\leftarrow} + \boldsymbol{\rho}_{ai_{\text{lock}\leftarrow}})\right)$. We do likewise for $\overrightarrow{e_1}$ and $\overrightarrow{e_2}$, using a separate distribution $\text{Cat}\left(\text{softmax}(\boldsymbol{\mu}_{\text{lock}\rightarrow} + \boldsymbol{\rho}_{ai_{\text{lock}\rightarrow}})\right)$. Finally, if the start point from one event and the endpoint from the other are free (i.e. not locked), we determine their relative ordering using a third distribution $\text{Cat}\left(\text{softmax}(\boldsymbol{\mu}_{\text{lock}\leftrightarrow} + \boldsymbol{\rho}_{ai_{\text{lock}\leftrightarrow}})\right)$.

Implementation We fit our model to the training data using expectation-maximization. We use loopy belief propagation to obtain the posteriors over event, entity, role, and relation types in the expectation step and the Adam optimizer to estimate the parameters of the distributions associated with each type in the maximization step. As a stopping criterion, we compute the evidence that the model assigns to the development data, stopping when this quantity begins to decrease.

To select $|\mathcal{T}_{\text{event}}|$, $|\mathcal{T}_{\text{ent}}|$, $|\mathcal{R}_{\text{role}}|$, and $|\mathcal{R}_{\text{rel}}|$, we use a simplified version of Algorithm 1 that removes all of the implied factor nodes—effectively, fitting separate mixture models for each ontology. For each ontology, we then compute the evidence that the simplified model assigns to the development data given some number of types, choosing the smallest number such that there is no reliable increase in the evidence for any larger number. To determine reliability, we compute 95% confidence intervals using nonparametric bootstraps.

Results Based on the procedure described above, $|\mathcal{T}_{\text{event}}| = 4$, $|\mathcal{T}_{\text{ent}}| = 8$, $|\mathcal{R}_{\text{role}}| = 2$, and $|\mathcal{R}_{\text{rel}}| = 5$. To interpret these classes, we inspect the property means $\boldsymbol{\mu}_t$ associated with each type t and give examples from UD-EWT for which the posterior probability of that type is high.

Event Types While our goal was not necessarily to reconstruct any particular classification laid out in the theoretical literature, the four event types align reasonably well with those proposed by Vendler (1957): statives (16), activities (17), achievements (18), and accomplishments (19).

(16) I have finally found a mechanic I **trust**!!
(17) his agency is still **reviewing** the decision.
(18) A suit against [. . .] Kristof was **dismissed**.
(19) a consortium [. . .] **established** in 1997

This alignment is surprising given that Vendler's classification was not developed with actual language use in mind and thus abstracted away from complexities that arise when dealing with, e.g., non-factual or generic events. Nonetheless, there do arise cases where a particular predicate has a wider distribution across types than we might expect based on prior work. For instance, *know* is prototypically classed as a stative; and while it does get classed that way by our model, it also gets classed as an accomplishment or achievement (though rarely an activity)—e.g. when it is used to talk about coming to know something, as in (20).

(20) Please let me **know** how[. . .]to proceed.

Entity Types Our entity types are: person/group (21), concrete artifact (22), contentful artifact (23), particular state/event (24), generic state/event (25), time (26), kind of concrete objects (27), and particular concrete objects (28).

(21) Have a real **mechanic** check[...]
(22) I have a [. . .] cockatiel, and there are 2 **eggs** in the bottom of the cage[. . .]
(23) Please find attached a credit **worksheet**[. . .]
(24) He didn't take a **dislike** to the kids[...]
(25) They require **a lot of attention** [. . .]
(26) Every move Google makes brings this particular **future** closer.
(27) And what is their big / main **meal** of the day.
(28) Find him before he finds the dog **food**.

Role Types The optimality of two role types is consistent with Dowty's (1991) proposal that there are only two abstract role prototypes—*proto-agent* and *proto-patient*—into which individual thematic roles—i.e. those specific to particular predicates—cluster. Further, the means for the two role types we find very closely track those predicted by Dowty, with clear proto-agents (29) and proto-patients (30) (see also White et al. 2017).

(29) **they** don't ***press*** their sandwiches.
(30) you don't ever feel like you **ate** too ***much***.

Relation Types The relation types we obtain track closely with approaches that use ontologies of underspecified temporal relations (Cassidy et al., 2014; O'Gorman et al., 2016; Zhou et al., 2019, 2020; Wang et al., 2020): e_1 starts before e_2 (31), e_2 starts before e_1 (32), e_2 ends after e_1 (33), e_1 contains e_2 (34), and $e_1 = e_2$ (35).

(31) [. . .]the Spanish, Thai and other contingents are already **committed** to **leaving** [. . .]
(32) And I have to **wonder**: Did he **forget** that he already has a memoir[...]
(33) no, i am not **kidding** and no i don't want it b/c of the taco bell dog. i want it b/c it is really **small** and cute.
(34) they **offer** cheap air tickets to their country [...] you may get excellent discount airfare, which may even **surprise** you.
(35) the food is good, however the tables are so **close together** that it feels very **cramped**.

7 Comparison to Existing Ontologies

To explore the relationship between our induced ontologies and existing event and role ontologies, we ask how well our event, role, and entity types map onto those found in PropBank and VerbNet. Importantly, the goal here is not perfect alignment between our types and PropBank and VerbNet types but rather to give a sense for the similarities and differences between our empirically derived ontology and existing ontologies.

	Role	Precision	Recall	F1
argnum	A0	0.59	0.65	0.61
	A1	0.73	0.78	0.75
functag	pag	0.57	0.64	0.61
	ppt	0.66	0.75	0.70
verbnet	agent	0.72	0.46	0.56
	patient	0.21	0.16	0.18
	theme	0.59	0.59	0.59

Table 2: Test set results for all role types that are labeled on at least 5% of the development data.

To carry out these comparisons, we use the the parameters of the posterior distribution over event types for each predicate, the posterior distribution over role types for each argument of each predicate, and the posterior distribution over entity types for each argument as features in an SVM with radial basis function kernel predicting the event/role types found in PropBank and VerbNet. All experiments use the standard training, development, and testing splits given in UD-EWT.

Role Type Comparison We first obtain a mapping from UDS predicates and arguments to the PropBank predicates and arguments annotated in EWT. Each such argument in PropBank is annotated with an argument number (A0-A4) as well as a function tag (PAG = *agent*, PPT = *patient*, etc.).[3] We then compose this mapping with the mapping given in the PropBank frame files from PropBank rolesets to sets of VerbNet classes and from PropBank roles to sets of VerbNet roles (AGENT, PATIENT, THEME, etc.) to obtain a mapping from UDS arguments to sets of VerbNet roles. Because a particular argument maps to a set of VerbNet roles, we treat predicting VerbNet roles as a multi-label problem, fitting one SVM per role.

Table 2 gives the test set results for all role types labeled on at least 5% of the development data. Our roles tend to align well with agentive roles—PAG, AGENT, and A0—and some non-agentive roles—PPT, THEME, and A1—but they align less well with other non-agentive roles—PATIENT. This result suggests that our two-role ontology aligns fairly closely with the agentivity distinctions in PropBank and VerbNet, as we would expect if our roles in fact captured something like Dowty's coarse distinction among prototypical agents and patients.

[3]PropBank argument numbers are roleset specific, but the annotation guidelines explicitly draw a link between agentivity and A0 and non-agentivity and the remaining roles.

Predicate	Precision	Recall	F1
cause	0.50	0.98	0.66
do	0.34	0.27	0.30
has_possession	0.31	0.26	0.28
has_location	0.19	0.21	0.20
motion	0.14	0.14	0.14

Table 3: Test set results for all VerbNet predicates that are labeled on five most frequent predicates.

Event Type Comparison The PropBank roleset and VerbNet class ontologies are extremely fine-grained, with PropBank rolesets capturing a single sense of a predicate and VerbNet classes capturing very fine-grained syntactic behavior of a generally small set of predicates. Since our event ontology is intended to be more general than either, we do not compare it directly to PropBank rolesets or VerbNet classes. Instead, we compare to the VerbNet semantics layer, which provides a neo-Davidsonian event decomposition.

We specifically use the generative lexicon-inspired (DEM) variant of this semantics layer (Brown et al., 2018). An example of this layer for `give-13.1` is has_possession(e1, Ag, Th) & transfer(e2, Ag, Th, Rec) & cause(e2, e3) & has_possession(e3, Rec, Th). We predict only the predicates in this decomposition—e.g. transfer or cause—treating the problem as multi-label and fitting one SVM per predicate.

Table 3 gives the test set results for the five most frequent predicates in the corpus. Our ontology generally agrees with VerbNet in terms of its classification of events that involve causation (CAUSE) and, to a lesser extent, eventivity (DO) and possession. The ontologies' agreement in terms of whether an event involves (change of) physical location or motion is lower, though non-negligible.

8 Conclusion

We have presented an event structure ontology derived from inferential properties annotated on sentence- and document-level semantic graphs. We induced this ontology jointly with semantic role, entity type, and event-event relation ontologies using a document-level generative model. Our model identifies types—notably, four event types like Vendler's, and two proto-agent and proto-patient role types like Dowty's—consistent with theoretical predictions.

References

Omri Abend and Ari Rappoport. 2013. Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Alan Agresti. 2014. *Categorical Data Analysis*. John Wiley & Sons.

James Allen, Hannah An, Ritwik Bose, Will de Beaumont, and Choh Man Teng. 2020. A broad-coverage deep semantic lexicon for verbs. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3243–3251, Marseille, France. European Language Resources Association.

Emmon Bach. 1986. The Algebra of Events. *Linguistics and Philosophy*, 9(1):5–16.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley Framenet Project. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90. Association for 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, Sofia, Bulgaria. Association for Computational Linguistics.

Michael Ruisdael Bennett and Barbara Hall Partee. 1978. *Towards the logic of tense and aspect in English*. Indiana University Linguistics Club, Bloomington, IN.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English web treebank. *Linguistic Data Consortium, Philadelphia, PA*.

Susan Windisch Brown, James Pustejovsky, Annie Zaenen, and Martha Palmer. 2018. Integrating Generative Lexicon event structures into VerbNet. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics.

Lucas Champollion. 2010. *Parts of a whole: Distributivity as a bridge between aspect and measurement*. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 33–40.

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

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.

Bonnie Jean Dorr. 1993. *Machine translation: a view from the Lexicon*. MIT press.

David Dowty. 1979. *Word Meaning and Montague Grammar: The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*, volume 7. Springer Science & Business Media.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

George Ferguson and James F. Allen. 1998. TRIPS: An integrated intelligent problem-solving assistant. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98, page 567–572, USA. American Association for Artificial Intelligence.

Francis Ferraro and Benjamin Van Durme. 2016. A unified bayesian model of scripts, frames and language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

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

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

Annemarie Friedrich and Alexis Palmer. 2014b. Situation entity annotation. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 149–158, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

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, Berlin, Germany. Association for Computational Linguistics.

Venkata Govindarajan, Benjamin Van Durme, and Aaron Steven White. 2019. Decomposing generalization: Models of generic, habitual, and episodic statements. *Transactions of the Association for Computational Linguistics*, 7:501–517.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity. In *Workshop on events: Definition, detection, coreference, and representation*, pages 21–28.

Nancy Ide, Collin Baker, Christiane Fellbaum, Charles Fillmore, and Rebecca Jane Passonneau. 2008. MASC: The manually annotated sub-corpus of American English. In *6th International Conference on Language Resources and Evaluation, LREC 2008*, pages 2455–2460. European Language Resources Association (ELRA).

Ray Jackendoff. 1990. *Semantic Structures*, volume 18. MIT press.

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

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Manfred Krifka. 1989. Nominal reference, temporal constitution and quantification in event semantics. In Renate Bartsch, Johan van Benthem, and Peter van Emde Boas, editors, *Semantics and Contextual Expressions*, pages 75–115. Foris, Dordrecht.

Manfred Krifka. 1992. Thematic relations as links between nominal reference and temporal constitution. *Lexical matters*, 2953.

Manfred Krifka. 1998. The Origins of Telicity. In Susan Rothstein, editor, *Events and Grammar*, Studies in Linguistics and Philosophy, pages 197–235. Springer Netherlands, Dordrecht.

K. Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Content Analysis: An Introduction to Its Methodology. Sage.

George Lakoff. 1965. *On the nature of syntactic irregularity*. PhD Thesis, Massachusetts Institute of Technology.

Beth Levin. 1993. *English Verb Classes and Alternations: A preliminary investigation*. University of Chicago Press, Chicago.

Beth Levin and Malka Rappaport Hovav. 1991. Wiping the slate clean: A lexical semantic exploration. *Cognition*, 41(1-3):123–151.

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

Alexander P. D. Mourelatos. 1978. Events, Processes, and States. *Linguistics and Philosophy*, 2(3):415–434. Publisher: Springer.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

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, Austin, Texas. Association for Computational Linguistics.

Mari Broman Olsen. 1997. *A semantic and pragmatic model of lexical and grammatical aspect*. Outstanding Dissertations in Linguistics. Garland.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Christopher Piñón. 1995. *An Ontology for Event Semantics*. Ph.D. thesis, Stanford University, Palo Alto.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

James Pustejovsky. 2013. Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 1–10, Pisa, Italy. Association for Computational Linguistics.

James Pustejovsky, Marc Verhagen, Roser Saurí, Jessica Littman, Robert Gaizauskas, Graham Katz, Inderjeet Mani, Robert Knippen, and Andrea Setzer. 2006. TimeBank 1.2. *Linguistic Data Consortium*, 40.

Malka Rappaport Hovav and Beth Levin. 1998. Building verb meanings. *The projection of arguments: Lexical and compositional factors*, pages 97–134.

Malka Rappaport Hovav and Beth Levin. 2001. An event structure account of english resultatives. *Language*, 77(4):766–797.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic proto-roles. *Transactions of the Association for Computational Linguistics*, 3:475–488.

Rachel Rudinger, Adam Teichert, Ryan Culkin, Sheng Zhang, and Benjamin Van Durme. 2018. Neural-Davidsonian semantic proto-role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 944–955, Brussels, Belgium. Association for Computational Linguistics.

Carlota S Smith. 2003. *Modes of discourse: The local structure of texts*, volume 103. Cambridge University Press.

Barry Taylor. 1977. Tense and continuity. *Linguistics and Philosophy*, 1(2):199–220.

Carol Lee Tenny. 1987. *Grammaticalizing aspect and affectedness*. Thesis, Massachusetts Institute of Technology.

Robert Truswell, editor. 2019. *The Oxford Handbook of Event Structure*. Oxford University Press, Oxford. Publication Title: The Oxford Handbook of Event Structure.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy. Association for Computational Linguistics.

Zeno Vendler. 1957. Verbs and Times. *Philosophical Review*, 66(2):143–160.

Henk J. Verkuyl. 1972. *On The Compositional Nature Of The Aspects*, volume 15 of *Foundations of Language*. D. Reidel Publishing Company, Dordrecht.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706, Online. Association for Computational Linguistics.

Aaron Steven White, Kyle Rawlins, and Benjamin Van Durme. 2017. The semantic proto-role linking model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 92–98, Valencia, Spain. Association for Computational Linguistics.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, Texas. Association for Computational Linguistics.

Aaron Steven White, Elias Stengel-Eskin, Siddharth Vashishtha, Venkata Subrahmanyan Govindarajan, Dee Ann Reisinger, Tim Vieira, Keisuke Sakaguchi, Sheng Zhang, Francis Ferraro, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2020. The universal decompositional semantics dataset and decomp toolkit. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5698–5707, Marseille, France. European Language Resources Association.

Sheng Zhang, Rachel Rudinger, and Benjamin Van Durme. 2017. An Evaluation of PredPatt and Open IE via Stage 1 Semantic Role Labeling. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *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 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589, Online. Association for Computational Linguistics.