Semantic Proto-Roles

Drew Reisinger Rachel Rudinger Francis Ferraro Craig Harman
Kyle Rawlins* Benjamin Van Durme*
{rawlins@cogsci, vandurme@cs}.jhu.edu
Johns Hopkins University

Abstract

We present the first large-scale, corpus based verification of Dowty's seminal theory of proto-roles. Our results demonstrate both the need for and the feasibility of a property-based annotation scheme of semantic relationships, as opposed to the currently dominant notion of categorical roles.

1 Introduction

For decades researchers have debated the number and character of *thematic roles* required for a theory of the syntax/semantics interface. AGENT and PATIENT are canonical examples, but questions emerge such as: should we have a distinct role for BENEFICIARY? What about RECIPIENT? What are the boundaries between these roles? And so on.

Dowty (1991), in a seminal article, responded to this debate by constructing the notion of a *Proto*-Agent and *Proto*-Patient, based on entailments that can be mapped to questions, such as: *"Did the argument change state?"*, or *"Did the argument have volitional involvement in the event?"*. Dowty argued that these properties group together in the lexicon non-categorically, in a way that aligns with classic Agent/Patient intuitions. For instance, a Proto-Patient often both changes state (but might not), and often is causally affected by another participant.

Various resources have been developed for computational linguists working on 'Semantic Role Labeling' (SRL), largely under the classical, categorical notion of role. Here we revisit Dowty's research as computational linguists desiring data for a new task, Semantic Proto-Role Labeling (SPRL), in which existing coarse-grained categorical roles are replaced by scalar judgements of Dowty-inspired properties. As the availability of supporting data is a critical component of such a task, much of our efforts here are focused on showing that everyday English speakers (untrained annotators) are able to answer basic questions about semantic relationships.

In this work we consider the following questions: (i) can crowdsourcing methods be used to empirically validate the formal linguistic theory of Dowty, following prior work in psycholinguistics (Kako, 2006b)? (ii) How might existing semantic annotation efforts be used in such a pursuit? (iii) Can the pursuit of Dowty's semantic properties be turned into a practical and scalable annotation task? (iv) Do the results of such an annotation task (at various scales, including over very large corpora) continue to confirm Dowty's proto- role hypothesis? And finally, (v) how do the resulting configurations of fine-grained role properties compare to coarser annotated roles in resources such as VerbNet?[1]

We first derive a set of basic semantic questions pertaining to Dowty-inspired properties. These questions are used in two Mechanical Turk HITs that address the above issuess. In the first HIT, we build on psycholinguistic work (Kako, 2006b) to directly access 'type-level' intuitions about a lexical item, by asking subjects property-questions using made-up ("nonce") words in argument positions. Our results

*Corresponding authors.

[1]To be clear, Dowty himself does not make direct predictions about the distribution of proto-role properties within a corpus, except insofar as a corpus is representative of the lexicon.

Transactions of the Association for Computational Linguistics, vol. 3, pp. 475–488, 2015. Action Editor: Diana McCarthy.
Submission batch: 3/2015; Revision batch 6/2015; Published 8/2015.

replicate these previous experiments, and demonstrate that what can be done in this domain in a controlled lab experiment can be done via crowdsourcing. We extend this to a large-scale MTurk annotation task using corpus data. This task presents an annotator with a particular ('token-level') sentence from PropBank (Palmer et al., 2005) and a highlighted argument, and asks them for a likelihood judgment about a property; for example, "*How likely or unlikely is it that* ARG *is sentient?*". By looking across many token-level instances of a verb, we can then infer type-level information about the verb.

We discuss results from this task over 11 role properties annotated by a single (trusted) annotator on approximately 5000 verb tokens. Our results represent the first large-scale corpus study explicitly aimed at confirming Dowty's proto-role hypothesis: Proto-Agent properties predict the mapping of semantic arguments to subject and object. We show that this allows us to both capture and discover fine-grained details of semantic roles that coarser annotation schemes such as VerbNet do not: empirically, this data set shows a great degree of *role fragmentation*, much greater than any existing annotation scheme allows. The results of this task represent a new large-scale annotated resource, involving close to 345 hours of human effort.[2]

2 Background

2.1 Roles in linguistics

Thematic roles have been a key analytical component in modern linguistic theory.[3] Despite the vast literature, there is surprisingly little consensus over what a thematic role is, or how to identify or precisely characterize them. A 'textbook' approach, influential in linguistics and computer science, is that there is a (short) list of core *Generalized Thematic Roles*, such as AGENT, PATIENT, EXPERIENCER, etc. that verbs assign to arguments. However, it has been known for some time that this view is problematic (see Levin and Rappaport Hovav (2005) for an overview). Perhaps the best known arguments emerge from the work of David Dowty.

Proto-roles Dowty (1991), in an exhaustive survey of research on thematic roles up to that point, identifies a number of problems with generalized thematic roles. First and foremost, if the inventory of role types is small, then it proves impossible to clearly delineate the boundaries between role types. This situation pushes researchers who want clean role boundaries towards a very large inventory of specialized, fine-grained thematic roles – what Dowty termed *role fragmentation*. A large, fragmented set of role-types may be useful for many purposes, but not for expressing generalizations that should be stated in terms of thematic roles. Dowty (1991) focuses on generalizations related to *the mapping problem*: how are syntactic arguments mapped to semantic arguments? The mapping problem is not just a linguistic puzzle, but a central problem for tasks such as SRL, semantic parsing, etc.

Dowty offers a solution to the mapping problem couched not in terms of fine-grained fragmented thematic roles, but in terms of what Dowty analogizes to 'prototype' concepts constructed over fine-grained role properties. In particular, the role-properties are features such as whether the participant in question causes the event to happen, or whether the participant changes state. Dowty groups properties into two classes: Proto-Agent properties, and Proto-Patient properties. A semantic argument is more AGENT-like the more Proto-Agent properties it has, and more PATIENT-like the more Proto-Patient properties it has. These two sets of properties are in competition, and an argument can have some of each, or even none of the properties. Dowty's role properties (slightly modified) are shown in Table 1; we use these as a starting point for our own choice of fine-grained features in §3.[4]

Classic role types fall out from what we will

[2] Available through the JHU Decompositional Semantics Initiative (Decomp): `http://decomp.net`.

[3] A full accounting of the history of thematic roles is beyond the scope available here (Blake, 1930; Gruber, 1965; Fillmore, 1966; 1976; 1982; Castañeda, 1967; Jackendoff, 1972; 1987; Cruse, 1973; Talmy, 1978; Chomsky, 1981; Carlson, 1984; Carlson and Tanenhaus, 1988; Rappaport and Levin, 1988; Rappaport Hovav and Levin, 1998; Levin and Rappaport Hovav, 2005; Dowty, 1989; 1991; Parsons, 1990; Croft, 1991; Davis and Koenig, 2000, among others).

[4] Dowty's Argument Selection Principle: *"In predicates with grammatical subject and object, the argument for which the predicate entails the greatest number of Proto-Agent properties will be lexicalized as the subject of the predicate; the argument having the greatest number of Proto-Patient entailments will be lexicalized as the direct object."* (Dowty 1991:31)

Proto-Agent properties	Proto-Patient properties
a. volitional involvement	f. changes state
b. sentience (/perception)	g. incremental theme
c. causes change of state	h. causally affected
d. movement (relative)	i. stationary (relative)
e. independent existence	j. no indep. existence

Table 1: Proto-role properties (Dowty 1991:27–28).

term *configurations* of these properties. A 'core' AGENT, for example, would have all of the Proto-Agent properties. An EXPERIENCER would have Proto-Agent properties (b) and (e), and Proto-Patient property (h), and so would be less AGENT-like than a core AGENT. This idea is further developed by Grimm (2005; 2011), who points out that when combinations of proto-role properties are looked at as a lattice structure, generalized thematic roles can be identified with particular parts of the lattice. If Dowty's proposal is right, the lexicon will instantiate a very large number of property configurations, rather than a small and constrained set.

A key result of this theory is explanation of the contrast between what Dowty terms *stable* and *unstable* predicates. A stable predicate is one like *kill* whose mapping behavior is similar across languages – the KILLER is mapped to subject, and the VICTIM to object. An unstable predicate is one where this is not so. Instability can also manifest within a language, in the form of lexical doublets such as *buy* and *sell*. The Proto-Patient argument for these verbs is stable, but the subject alternates: for *buy* it is the GOAL argument that appears as subject while for *sell* it is the SOURCE. Dowty's explanation is that for transaction events, SOURCE and GOAL are very similar in their Proto-Agent properties, and so compete equally for subject position.

Dowty's linguistic proposal, if correct, has substantial implications for human language technology (see also discussion in Palmer et al. (2005)). It suggests an approach to semantic annotation, semantic parsing, and related tasks that focuses on this fine-grained level of proto-role properties, with any more generalized thematic roles as emergent property configurations. If lexical argument structure is organized around proto-roles, then we predict that we will find this organization reflected in corpora, and that token-level annotations of verb meanings would benefit from observing this organization. In particular, an annotation strategy that takes the proto-role hypothesis seriously would annotate verbs for properties such those shown in Table 1.

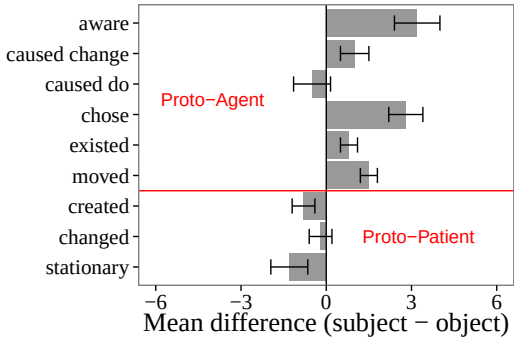

Figure 1: Proto-role properties in Kako 2006 exp. 1 (reproduction of Kako's Fig. 1). Error bars in all figures show 95% t-test CIs.

Experimental work Can the proto-role hypothesis be operationalized? A starting point is experimental work by Kako (2006a,b), who took the proto-role hypothesis into the lab. Kako developed several experimental versions of the hypothesis, whereby participants were asked simplified question-based versions of Dowty's proto-role properties about sentences of English. Kako did not use actual or attested sentences of English, but rather focused on 'nonce'-based tasks. That is, he constructed stimuli by taking constructed sentences of English containing the target verbs, and replacing noun positions with nonce words like *dax*. Subjects were then presented with these nonce sentences and asked questions such as, *"How likely is it that the dax moved?"*.

The nonce-method is designed to access 'type-level' judgments about verbs across frames. Across all experiments, Kako confirms a version of the proto-role hypothesis: subject arguments across the verbs he examines have significantly more Proto-Agent than Proto-Patient properties, and vice versa for objects. Fine-grained results for individual proto-role properties from one of his experiments are shown in Figure 1: this presents an aggregate measure of the success of the proto-role hypothesis, showing the mean difference between property ratings for subject vs. object arguments. Dowty's mapping hypothesis predicts that subjects should skew towards Proto-Agent properties, and objects towards

Proto-Patient properties, exactly Kako's finding.

Kako's work help lead Alishahi and Stevenson (2010) to annotate a small collection of child directed speech with Dowty-inspired properties, used to evaluate a Bayesian model for inducing what they termed *semantic profiles*.[5]

2.2 Roles in computational linguistics

PropBank[6] PropBank (Palmer et al., 2005) layers predicate/argument annotations on the English portion of the Penn Treebank (PTB) (Marcus et al., 1993), treating semantic role annotation as a sort of slot-filling exercise: a *frameset* defines a set of semantic roles that a particular type of predicate may use. Every verb is assigned a frameset (roughly, a verb sense), and arguments of the verb (potentially a non-contiguous span) are labeled with a particular role. Coarse categorical labels, such as ARG_0 and ARG_1, allow PropBank to both capture some of Levin (1993)'s syntactic variations, and imbue this syntactic information with shallow semantics. Annotations do not cross sentence boundaries.

As every verb in the PTB was annotated, PropBank has good coverage: 4,500 framesets cover around 3,300 verb types. Additional resources have adopted and extended PropBank, e.g. (Weischedel et al., 2013, etc.), and there have been multiple shared tasks centered around PropBank-style SRL (Carreras and Màrquez, 2005). However, at three days (Palmer et al., 2005), the training time for an annotator is significantly higher than the crowdsourcing solution we pursue here.

VerbNet and SemLink VerbNet (Schuler, 2005) provides a class-based view of verbs. It applies Levin's verb classes (Levin, 1993) to more than five thousand (English) verbs, categorizing them according to their syntactic behaviors. Beyond this grouping, which includes a shallow semantic parse frame, VerbNet provides its own semantic role labels, and a neo-Davidsonian-inspired logical form. All information within VerbNet is class-specific; the frames and roles apply equally to all verbs within a class.[7] Further, VerbNet's lexical entries allow for assigning selectional restrictions on thematic roles, e.g. requiring a participant be CONCRETE, or ANIMATE. While these restrictions take the form of properties, the thematic roles themselves are left categorical.

Bonial et al. (2011) united VerbNet's semantic roles with those of LIRICS[8], a standardization effort to facilitate multilingual NLP. Motivated in part by the properties of Dowty, they constructed a hierarchy of 35 roles interrelated through their property requirements, implicit in the organization of the hierarchy paired with natural language role definitions. The properties bundled into these roles are then taken taken to be type-level, hard constraints: they cannot reflect semantic nuances within individual sentences, and are strictly boolean (a property cannot hold to a degree, or with some uncertainty).

The SemLink project (Loper et al., 2007) provides a mapping between VerbNet, PropBank, FrameNet (see below) and WordNet (Fellbaum, 1998). Crucially for our work (see §6), SemLink provides a mapping from the role hierarchy of Bonial et al. (2011) to the argument annotations of PropBank.

VerbCorner VerbCorner (Hartstone et al., 2013; Hartshorne et al., 2014) is an on-going effort to validate VerbNet's semantic annotations, focusing at a finer-grained level of role information. For a particular verb and semantic features, annotators are provided context through a small, made-up story. Annotators then read example sentences pulled from VerbNet and determine whether those sentences violate the contextual expectations. As with the present work, VerbCorner crowd-sources the anno-

[5] Probability distributions over observed configurations that capture a generalized notion of semantic (proto-)role.

[6] This section is not a fully exhaustive list of resources, and we omit discussion of several important ones that are complementary to our efforts. For example, resources such as the Pattern Dictionary of English Verbs (Hanks, 2013), currently in progress, could be supplemented by our SPRL annotations. (The PDEV will contain valency patterns for thousands of verbs along with restrictions on the semantic types of their arguments based on (Pustejovsky et al., 2004)'s ontology.) Also important is early connectionist work, which proposed "semantic microfeatures" to model semantic role generalizations; see e.g. Hinton (1981; 1986) and McClelland and Kawamoto (1986).

[7] For instance, the lemmas *break* and *shatter* are both members of the same class (BREAK-45.1), capturing the causative alternation. Both senses can be used transitively (*"John broke/shattered the mirror"*) or intransitively (*"The mirror broke/shattered."*), while semantic roles assign *John* to AGENT and *the mirror* to PATIENT in both syntactic frames, capturing the logical entailment of a resulting degraded physical form.

[8] Linguistic InfRastructure for Interoperable ResourCes and Systems (LIRICS): `http://lirics.loria.fr/`

tation, though there are key differences: Hartstone et al. (2013) are focused on logical entailments (what *must* be true) whereas we are focused on *strongly suggested* implications (what is *likely* to be true).

FrameNet The Berkeley FrameNet Project (Baker et al., 1998) is an instantiation of Fillmore's frame semantic theory (Fillmore, 1982). FrameNet describes events via a frame, consisting of lexical triggers and semantic roles that are expected to be filled. This is similar to PropBank's take on predicate/argument structure, though there are significant differences: (1) FrameNet triggers may be multiword, verbal or nominal expressions; (2) unlike PropBank, FrameNet defines interframe relations; (3) FrameNet is extremely fine-grained (embraces role-fragmentation), opting for semantic completeness rather than annotator ease. FrameNet has inspired semantic role labeling (Gildea and Jurafsky, 2002; Litkowski, 2004), in addition to frame semantic parsing (Baker et al., 2007; Das et al., 2010).

3 Experimental Setup

The literature review makes clear that understanding and annotating fine-grained role properties is valuable in both linguistic theory and in computational linguistics: under many sets of assumptions, such properties ground out the theory of coarse-grained roles. We follow Hartstone et al. (2013) in directly addressing fine-grained properties, here in the context of the proto-role theory. The proto-role approach gives us a set of testable questions to assess on a corpus. We focus on two main issues: (i) whether the proto-role solution to the mapping problem scales up to very large sets of data, and (ii) the prediction that there will be a very large set of property configurations attested as roles in a large data set. If the predictions from the proto-role theory are true, then we conclude that a large data set annotated with fine-grained role properties may be valuable in tasks related to semantic roles and event detection.

To assess these predictions, we broadly follow Kako (2006b) in operationalizing proto-roles using likelihood questions targeting specific role properties in sentences of English. This paper presents two experiments that implement this strategy. In the remainder of this section we describe the general setup of the experiments. In particular, we describe a process for arriving at the specific fine-grained property questions we ask, the creation of the data set that we ask the questions about, the task that Mechanical Turkers are presented with, and the manner in which we analyze and display the results.

Role property	Q: How likely or unlikely is it that...
instigated	`Arg` caused the `Pred` to happen?
volitional	`Arg` chose to be involved in the `Pred`?
awareness	`Arg` was/were aware of being involved in the `Pred`?
sentient	`Arg` was sentient?
moved	`Arg` changes location during the `Pred`?
phys_existed	`Arg` existed as a physical object?
existed_before	`Arg` existed before the `Pred` began?
existed_during	`Arg` existed during the `Pred`?
existed_after	`Arg` existed after the `Pred` stopped?
changed_poss	`Arg` changed possession during the `Pred`?
changed_state	The `Arg` was/were altered or somehow changed during or by the end of the `Pred`?
stationary	`Arg` was stationary during the `Pred`?

Table 2: Questions posed to annotators.

We first inspected the role hierarchy of Bonial et al. (2011) along with the associated textual definitions: these were manually decomposed into a set of explict binary properties. For example, we define the SemLink ACTOR role as a participant that has the binary property of INSTIGATION. From these properties we subselected those that were most similar to the original questions proposed by Dowty (see Table 1). For each such property we then generated a question in natural language to be posed to annotators given an example sentence (see Table 2). The set we report on here represents a subset of the questions we have tested; in ongoing work we are evaluating whether we can expand Dowty's set of questions, e.g. to capture roles such as INSTRUMENT.

Methods Because we are interested in the potential impact of Dowty's proto-roles theory on human language technologies, we perform a number of related crowdsourcing experiments, with the dual aim of validating the existing (psycho-)linguistic literature on proto-roles as well as piloting this highly scalable framework for future decompositional semantic annotation efforts.

All of the crowdsourcing experiments in this paper are run using Amazon Mechanical Turk, and (ex-

cept for the kappa scores reported for experiment 2) all workers were recruited from the MTurk worker pool. The basic setup of the experiments in Sections 4 and 5 is the same. The Mechanical Turk worker is presented with a single sentence with a highlighted verb and one highlighted argument of that verb. Then the worker answers all of the questions in Table 2 for that verb-argument pair using a Likert scale from 1 to 5, with the response labels: *very unlikely*, *somewhat unlikely*, *not enough information*, *somewhat likely*, and *very likely* (See Figure 2). Each Mechanical Turk HIT yields responses for all the questions in Table 2 applied to a single verb-argument pair. The Mechanical Turk experiments are run with two types of sentences: those with real verbs and nonsense ("nonce") arguments, and those with entirely real English sentences. Section 4 discusses the former "type-level" HIT with nonce arguments, while Section 5 discusses the latter "token-level" annotation task with real arguments.

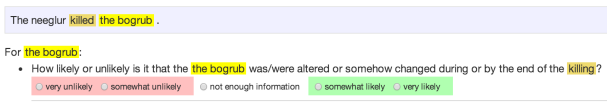

Figure 2: Example HIT question with nonce arguments.

Data To obtain verb-argument pairs for the task described here, we drew sentences from the subset of PropBank that SemLink annotates for VerbNet roles. From these, we removed verbs annotated as participles, verbs with trace arguments, verbs under negation or modal auxiliaries, and verbs in embedded clauses to ensure that annotators only saw verbs in *veridical* contexts – contexts where logical operations such as negation do not interfere with direct judgments about the verbs. For example, in *John didn't die*, negation reverses the change-of-state judgment for the whole sentence, despite that being part of the meaning of the verb *die*. We also removed clausal arguments, as most of the questions in Table 2 do not make sense when applied to clauses; in ongoing work we are considering how to extend this approach to such arguments. A total of 7,045 verb tokens with 11,913 argument spans from 6,808 sentences remained after applying these filters.

Analysis To evaluate whether the results of the following experiments accord with Dowty's proposal, we follow Kako (2006b) in taking the mean difference between the property ratings of the subject and object across sentences; see §2.1. We present these differences in the same format as in Figure 1. Here we stick with Kako's evaluation of the results, in order to demonstrate the convergence of the linguistic and psycholinguistic evidence with computational linguistic approaches; our immediate goal in the present work is not to advance the methodology, but to show that these techniques can be pursued through large-scale crowdsourcing.

We perform two Mechanical Turk experiments on verbs: one with nonce arguments, and one with real data in Section 5. Because nonce arguments have no meaning in their own right, we assume that the properties that annotators assign these arguments are a function of the verb and role, not the argument itself. Hence, we assume that these annotations are at the verb-role *type* level. Conversely, the experiment in Section 5 are at the *token* level, because all arguments have real English instantiations.

4 Experiment 1: Nonce-based

The first experiment we run with nonce arguments is an attempt to replicate the results of Kako (2006b). Recall that Kako (2006b) upholds the psychological validity of Dowty (1991)'s *Argument Selection Principle*, by demonstrating that human subjects assign Proto-Agent and Proto-Patient properties to grammatical subject and object arguments according to Dowty's prediction. (See Figure 1.)

In this experiment, we generate simple transitive sentences with a small set of real verbs and nonce arguments. The set of verbs are precisely those selected by Kako (2006b) in his first experiment: *add, deny, discover, finish, find, help, maintain, mention, pass, remove, show, write*. The questions we ask workers to answer come from a slightly expanded set of proto-role properties.[9] There were 16 partic-

[9] As pointed out by a reviewer, a verb in a nonce sentence is potentially ambiguous. Because we constructed the nonce sentences from actual frames in PropBank examples, an annotator will have at least coarse cues to the intended sense. In this respect we follow Kako, and established protocol in nonce experiments in general. We leave the effect of sense ambiguity on nonce property judgments for future work.

ipants in the experiment, recruited from the MTurk worker pool, each completing 7.5 HITs on average.

The results of this experiment, broadly, replicate Kako (2006b)'s earlier findings: human annotators on average indicate that, within the same sentence, the subject-position argument is more likely to have Proto-Agent properties than the object-position argument, and the object-position argument is more likely to have Proto-Patient properties than the subject-position argument. This finding is illustrated in Figure 3. In addition, the basic facts match Kako's original finding; compare Figures 3 and 1. (Our INSTIGATION property is equivalent to Kako's CAUSED CHANGE property, and we do not have an analogue of his CAUSED DO property.) Proto-Agent properties have a greater effect than Proto-Patient properties, and CAUSATION, VOLITION, and AWARENESS are all strong Proto-Agent properties. CREATION and STATIONARY are all weaker, but non-zero, Proto-Patient properties for these verbs. There are some differences that are apparent. First of all, where Kako did not (in this particular experiment) find an effect of CHANGE OF STATE, we did; this is broadly consistent with Kako's overall findings. We did not get an effect for MOVEMENT or for PHYSICAL EXISTENCE in this experiment, in contrast to Kako's results.

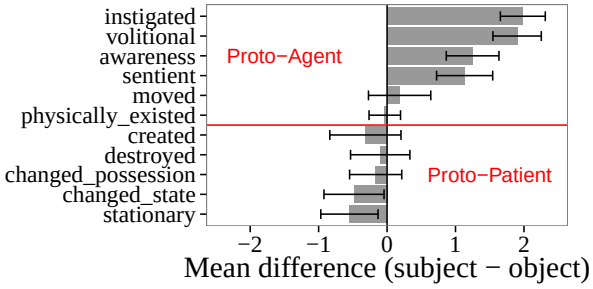

Figure 3: Mechanical Turk results for the nonce experiment. A positive value for a property indicates that, on average, subject-position arguments received a higher score for that property than object-position arguments.

Our ability to replicate Kako (2006b) is significant for two reason: (i) it lends further credence to the proto-role hypothesis, and (ii) it establishes that crowd-sourcing with non-experts in a less controlled situation than a formal experiment results in reasonable annotations for this task with minimal training.

5 Experiment 2: Corpus-based

Can this result extend to real corpus data? If so, the proto-role theory can lead to a valuable source of annotation information about thematic roles. To assess this, we moved from a synthetic nonce task to a much larger scale version of the task using data from PropBank (Palmer et al., 2005). Each item in this task presents the annotator with a PropBank sentence with the predicate and argument highlighted, and asks them the same questions about that actual sentence. The sentences were sampled from PropBank as described in §3.

Our primary goal in this collection effort was to obtain internally consistent, broad-coverage annotations. Thus we worked through a number of pilot annotation efforts to determine cross-annotator reliability between annotators and with our own judgements. From the final version of our pilot[10] we selected a single annotator with strong-pairwise agreement amongst the other most prolific annotators. Compared to the five other most prolific annotators in our final pilot, the pair-wise average Cohen's Kappa with squared metric on an ordinal interpretation of the Likert scale was 0.576.[11]

In our large-scale annotation task, we have collected property judgments on over 9,000 arguments of near 5,000 verb tokens, spanning 1,610 PropBank rolesets. This represents close to 350 hours of annotation effort. The results are shown in Figure 4. Because some arguments in PropBank are abstract, for which many of the questions in Table 2 do not make sense, we added an additional response field that asks *"Does this question make sense"* if the worker gives a response lower than 3 (Figure 6). Figure 5 shows the results with N/A responses removed. For presentation purposes, we convert the temporal existence properties to CREATION and DESTRUCTION.

Discussion The overall results substantially resemble both Kako's original results and our experi-

[10] Based on a set of 10 verbs selected based on frequency in the CHILDES corpus, filtering for verbs that had enough tokens in PropBank; *want.01, put.01, think.01, see.01, know.01, look.02, say.01, take.01, tell.01,* and *give.01*.

[11] One of those five annotators had less stable judgements than the rest, which we identified based on a pair-wise Kappa score of only 0.383 with our final annotator. If removing that annotator the average pair-wise score with the remaining four annotators then rose to 0.625.

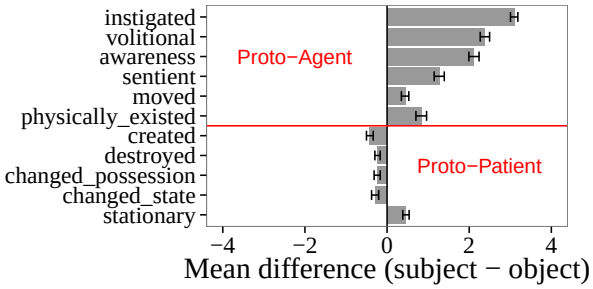

Figure 4: Mechanical Turk results for experiment 2.

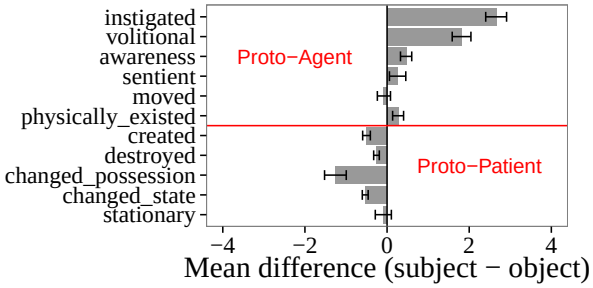

Figure 5: Experiment 2 with N/A removed.

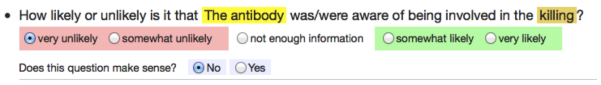

Figure 6: An additional response field appears for questions that might not be applicable. (Example sentence for this question: "The antibody then *kills* **the cell**.")

ment 1.As predicted, the Proto-Agent properties are predictors of whether an argument will be mapped to subject position, and the Proto-Patient properties similarly predict objecthood. Not all properties are equal, and on this much larger data set we can clearly see that INSTIGATION (causation) is the strongest property. Because we have many data points and a reliable annotator, the variance on this data is much smaller. This graph confirms the the proto-role hypothesis over a large corpus: fine-grained role properties predict the mapping of semantic roles to argument position. This data set puts us in a position to ask a wide range of followup questions about the nature of thematic roles, many of which we cannot adddress in the present paper. The central question we do address here is about property configurations: since each property configuration represents a coarse-grained role, we can ask what the distribution of property configurations is over this corpus. Dowty's prediction is that we should see some clustering around common configurations, but a long tail representing role fragmentation. The prediction of classical approaches is that we should see only the common configurations as clusters, with no long tail. We turn to this issue in the following sections, comparing our role annotations also to roles in VerbNet and FrameNet (using SemLink as the mapping among the three data sets).

Example	Rtg
(A) He *earned* **a master's degree in architecture** from Yale.	N/A
(B) The bridge normally *carries* **250,000 commuters a day**.	1
(C) Baskets of roses and potted palms *adorned* **his bench**.	5

Table 3: STATIONARY examples from experiment 2.

One key difference from Dowty's predictions is that STATIONARY appears to act as a Proto-Agent property. First, we are using a slightly different notion of stationary to Dowty, who proposed that it be relative to the event – in this we follow Kako. Second, our MOVEMENT property is really about change of location (see Table 2) and so is not the negation of STATIONARY. Third, our corpus is heavily biased to non-physical events and states, where the notion of motion does not apply, and so in this respect may not be fully representative of a more naturalistic corpus. Within the relatively small proportion of data that is left, we find that objects do not tend to be stationary, and so if this is correct, it may simply be wrong to classify the absolute version of STATIONARY as a Proto-Patient property. Three examples from the data set are shown Table 3, for each case – the result is that once N/A responses are excluded, examples such as (B) are still more the norm than examples such (C).

Annotation quality To assess annotation quality we began with a stratified sample based on each PropBank argument ID in the set $\{0, 1, 2, 3, 4, 5\}$.[12] Local researchers each then answered 209 questions over this sample. One of the authors participated,

[12]While argument IDs are meant to be meaningful when conditioned on a roleset, the values still correlate with the "coreness" of an argument even when independent of roleset (e.g., argument IDs 0 and 1 are most likely to be AGENT and PATIENT): our stratification aimed to survey across both "core" and "peripheral" argument role types.

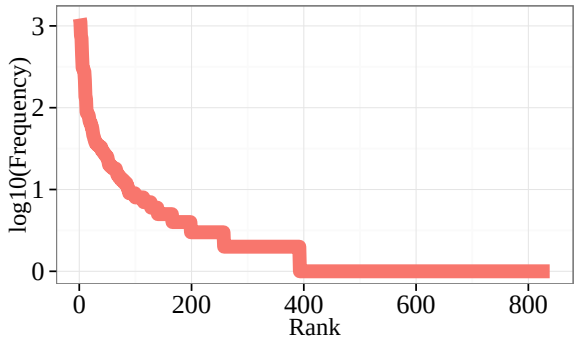

Figure 7: Distribution of property configurations in experiment 2. To obtain categorical roles for purposes of comparison, responses of 2/4 were mapped to 1/5, giving configurations on 11 properties over what we might coarsely consider: {False (1), Unknown (3), True (5)}.

achieving a Kappa score of 0.619 with the annotator. Two colleagues generally familiar with thematic roles but without prior experience with the protocol or our goals achieved scores of 0.594 and 0.642. Finally, a colleague who speaks English as a second language acheived a Kappa score of 0.479. These correlations, along with our initial selection criteria for the annotator, and then combined with those correlations observed in Table 6 (discussed below), suggests our process resulted in a useful resource which we will release to the community.

In section 6 we additionally provide a qualitative indicator of annotation quality, in the form of an alignment to VerbNet roles.

6 Comparison to Other Rolesets

A prediction emerging from the proto-role hypothesis is that, when a set of role-relevant properties such as those in Table 2 are tested on a large scale, we should not find clean role-clusters. We do expect to find certain common role-types appearing frequently, but we also expect there to be a long tail of combinations of properties. This is exactly what we find when examining our results. Figure 6 shows the frequency of property configurations in the data set. Around 800 configurations are attested, with nearly 75% of those making up the tail.

The proto-role hypothesis predicts that there are natural sentences in which an argument can be AGENT/PATIENT-like, yet be missing one or more Proto-agent/patient properties. This is what gives rise to the observed long tail of property configurations: cases that would otherwise be lumped together as, e.g., AGENT, are instead placed in a more diverse set of bins. While Dowty's theory is really about roles at the type-level, these bins are also useful for understanding role annotations at the token level, i.e. capturing exactly those properties that hold of the given argument in context.

Table 4 shows three real-world sentences taken from the Wall Street Journal involving the verb *kill*. Each sentence has what PropBank would call a KILL.01, ARG$_0$-PAG, or the first argument of the roleset KILL.01, a particular sense of the word *kill*.[13] Further, each of these arguments are labeled as a VerbNet AGENT and FrameNet KILLER/CAUSE through SemLink. These sentences were selected purely because they were the only instances of *kill* in our dataset with SemLink role annotations. Then, when examining our annotations for these arguments, we find that our motivations from §3 for this enterprise are justified. At the token level, there are robust inferences leading to different results on each example for key proto-role properties, but in each case the subject is still a better Proto-agent than the object. From this triplet, we learn that the subject of *kill* needn't be volitionally involved (as in the accidental death in A), needn't be aware of the killing, and even need not be sentient. The present annotation scheme, in contrast to the coarse label provided to these examples in VerbNet, captures this variation while still allowing inference to type-level properties of the verb *kill*. (These examples also clearly illustrate the degree to which noun semantics can influence thematic role-related judgments when carried out on natural data, something the fine-grained approach allows us to explore directly.) We can also clearly see from this triplet that INSTIGATION is constant across examples, as is PHYSICAL EXISTENCE. Interestingly, the example (B) shows that killing does not even preclude the continuation

[13] PAG is a recent addition to PropBank semantics, standing for Proto-Agent but interpreted as an unweighted disjunction of features: "it acts volitionally, is sentient, or perceives, causes a changes of state, or moves" (Kübler and Zinsmeister, 2015). Another addition, PPT, stands for Proto-Patient. While motivated by Dowty's terminology, these additions do not capture the individual property-based notion we advocate for here.

Sentences	Property	(A)	(B)	(C)
(A) **She** was untrained and, in one botched job *killed* a client.	instigated	5	5	5
(B) **The antibody** then *kills* the cell.	volitional	2	1	5
(C) **An assassin in Colombia** *killed* a federal judge on a Medellin street.	awareness	3	1	5
PropBank KILL.01, ARG_0-PAG: killer	sentient	5	1	5
VerbNet MURDER-42.1-1, AGENT: ACTOR in an event who initiates and carries out the event intentionally or consciously, and who exists independently of the event	moved	3	3	3
	phys_existed	5	5	5
	created	1	1	1
	destroyed	1	3	1
FrameNet KILLING, KILLER/CAUSE: (The person or sentient entity) / (An inanimate entity or process) that causes the death of the VICTIM.	changed_poss	1	1	1
	changed_state	3	3	3
	stationary	3	3	3

Table 4: Comparison of role annotations for *kill* across resources. Ratings: 1=very unlikely, 5=very likely.

Sentences	Property	(A)	(B)	(C)
(A) **The stock** *split* four-for-one on Oct. 10.	instigated	1	1	1
(B) "In 1979, the pair *split* **the company** in half, with Walter and his son, Sam, agreeing to operate under the Merksamer Jewelery name."	volitional	1	1	1
	awareness	1	5	1
(C) The company downplayed the loss of Mr. Lesk and *split* **his merchandising responsibilities** among a committee of four people.	sentient	1	1	1
	moved	1	1	1
PropBank SPLIT.01, ARG_1-PPT: thing being divided	phys_existed	1	1	1
VerbNet SPLIT-23.2, PATIENT: UNDERGOER in an event that experiences a change of state, location or condition, that is causally involved or directly affected by other participants, and exists independently of the event.	created	1	1	1
	destroyed	1	5	1
	changed_poss	1	5	5
FrameNet CAUSE_TO_FRAGMENT, WHOLE_PATIENT: The entity which is destroyed by the AGENT and that ends up broken into PIECES.	changed_state	5	5	4
	stationary	1	1	1

Table 5: Comparison of role annotations for *split* across resources.

of existence after the event, so the EXISTENCE property may not be fully independent.

Table 5 makes a similar point using the verb *split*. These three instances of *split*, labeled with the same role (and verb sense) in PropBank/VerbNet, show clear differences in terms of fine-grained role properties. (Note also that in (A), a PropBank ARG_1 appears in subject position.) While there is consensus on CHANGE OF STATE, there is variation in whether the argument is DESTROYED, CHANGES POSSESSION, and is AWARE of its involvement in the event.

Alignment with VerbNet In what follows we explore a non-exact mapping where we have taken sentences in SemLink annotated with VerbNet coarse-grain roles, and simply projected the mean 1-5 proto-role ratings (subtracting N/A) onto each role. This serves two purposes: (1) the quality of this mapping serves to verify the quality of the proto-role annotations, and (2) this alignment helps compare between coarse and fine-grained role annotations. This alignment is a proof-of-concept, and we leave a deeper exploration of ways of doing this sort of alignment for the future. Table 6 shows the full alignment. A value of 5 indicates that the role tends to determine the proto-property positively, i.e. AGENTS are extremely likely to be judged as instigators. A value close to 3 indicates that the role is neutral with respect to the proto-property, e.g. AGENTS may or may not move. A value close to 1 indicates that the arguments with that role are likely to have the negative version of the proto-property, e.g. AGENTS tend not to CHANGE POSSESSION. At a broad level the results are strong, though we will not be able to discuss every detail here.

In this alignment the judgments of N/A have been removed.[14] In the case of e.g. the INSTIGATION value for THEME, this supports interpreting the role as assigning no value to instigation at all; similarly for some of the other values for THEME. In some

[14]This is not the only way to treat N/A ratings, and we will leave a full exploration to future work.

Role	Freq	instigated	volitional	awareness	sentient	moved	existed	created	destroyed	chg poss	chg state	stationary
Agent	1546	4.9 (1355)	4.8 (1273)	4.9 (1275)	4.8 (810)	3.1 (897)	4.7 (947)	1.1 (1413)	1.1 (1508)	1.7 (432)	3.3 (1489)	2.8 (874)
Theme	1153	3.4 (214)	3.9 (215)	4.6 (226)	4.7 (147)	3.6 (335)	4.3 (412)	1.9 (986)	1.3 (1037)	3.3 (339)	3.9 (999)	2.5 (300)
Patient	312	3.1 (77)	3.2 (80)	4.5 (85)	4.4 (47)	3.3 (75)	4.6 (100)	1.1 (285)	1.6 (293)	3.4 (99)	4.5 (294)	2.7 (69)
Experiencer	210	4.4 (161)	4.3 (167)	4.9 (169)	4.8 (128)	3.1 (135)	4.7 (139)	1.0 (195)	1.1 (204)	1.4 (41)	3.6 (204)	2.8 (137)
Stimulus	129	4.3 (64)	4.0 (33)	4.1 (35)	4.2 (26)	2.9 (35)	3.7 (42)	1.7 (107)	1.1 (114)	1.8 (20)	3.1 (115)	2.9 (32)
Topic	114	4.0 (2)	2.3 (3)	2.5 (4)	3.5 (4)	3.0 (4)	3.0 (7)	2.0 (92)	1.1 (91)	2.6 (18)	3.4 (74)	3.0 (3)
Destination	91	1.6 (5)	2.9 (15)	4.5 (16)	4.8 (8)	2.3 (24)	4.9 (48)	1.5 (74)	1.2 (75)	2.2 (22)	4.2 (75)	4.1 (39)
Recipient	88	1.4 (37)	3.6 (58)	4.8 (60)	4.9 (35)	3.0 (46)	4.5 (52)	1.5 (84)	1.0 (85)	2.3 (30)	3.7 (82)	3.0 (40)
Extent	87	— (0)	— (0)	— (0)	— (0)	— (0)	— (0)	1.0 (1)	1.0 (2)	— (0)	3.0 (1)	— (0)
... 12 roles omitted ...												
Instrument	16	4.4 (9)	4.5 (8)	4.5 (8)	5.0 (4)	3.8 (5)	4.3 (7)	1.3 (15)	1.0 (15)	1.3 (6)	3.3 (13)	2.0 (4)
Initial Loc.	15	2.2 (5)	2.3 (7)	4.2 (8)	3.5 (2)	2.5 (4)	3.0 (4)	1.0 (14)	2.1 (14)	1.8 (6)	4.2 (13)	2.3 (3)
Beneficiary	13	3.6 (5)	5.0 (5)	5.0 (5)	5.0 (1)	3.0 (1)	3.5 (4)	2.2 (10)	1.0 (13)	3.0 (4)	3.7 (13)	3.0 (1)
Material	9	5.0 (1)	5.0 (1)	5.0 (2)	5.0 (1)	3.7 (3)	5.0 (3)	1.0 (7)	1.2 (8)	5.0 (1)	3.7 (7)	1.7 (3)
Predicate	8	— (0)	5.0 (1)	5.0 (1)	— (0)	— (0)	— (0)	1.0 (8)	2.2 (8)	— (0)	3.4 (5)	— (0)
Asset	7	— (0)	— (0)	— (0)	— (0)	3.0 (1)	3.0 (1)	1.0 (5)	1.0 (6)	5.0 (1)	4.3 (3)	— (0)

Table 6: High and low frequency VerbNet roles (via SemLink) aligned with mean property ratings when excluding N/A judgments. Freq provides the number of annotations that overlapped with a role. In parenthesis is the number of cases for that property which were judged applicable (not N/A). E.g. we annotated 1,546 arguments that SemLink calls AGENT, where 1,355 of those were deemed applicable for the instigation property, with a mean response of 4.9. 12 mid-frequency roles are omitted here for space reasons; the full alignment is provided with the dataset for this paper.

cases with large numbers of N/A responses, e.g. the awareness and sentient properties for THEME, the provided mean is high, suggesting the role may be more heterogeneous than would otherwise appear.

In lieu of an exhaustive discussion, we will motivate the alignment with several interesting examples of AWARENESS. AWARENESS tended to be rated highly. Table 7 gives a range of examples illustrating particular tokens of judgements relative to VerbNet roles. In (A-C) we give three straightforward examples where the bolded argument was judged to be aware of involvement in the event. Abstract entities such as companies were consistently annotated as having the potential to be aware. Consequently, in B for example, Ford is annotated as being aware that Mazda makes the Tracer for the company. In these cases, it is intuitively right at the token level that the participant is likely to be aware of their involvement in the event, but this does not mean that we can conclude anything about the role; for example, for BENEFICIARIES and INSTRUMENTS we have only a few examples of the AWARENESS property.

In C-E, we have given three examples of different ratings for AWARENESS focusing on the DESTINATION role. All three ratings are intuitively straightforward at the token level; in D the recipient of the cases (the court) may not yet be aware of the decision. In E the recipient of the *sprinkling* was a baby and therefore was quite unlikely to be aware of her potential future singing career (a fact about the context and argument more than the verb). F helps illustrate the quality of our annotations: personal communication with SemLink researchers verified that we discovered a rare bug via our process.[15]

	Rtg	(Role) Example
A	5	(AGENT) **They** worry about their careers, *drink* too much and suffer [...]
B	5	(BENEFICIARY) Mazda *makes* the Tracer for **Ford**.
C	5	(DESTINATION) Commercial fishermen and fish processors *filed* suit in **federal court** [...]
D	3	(DESTINATION) But the court [...] *sent* the cases back to **federal district court in Dallas**.
E	1	(DESTINATION) When the good fairy [...] hovered over the cradle of Edita [...], she *sprinkled* **her** with high E flats, [...]
F	5	(INSTRUMENT*) **Guests** *bring* movies on tape , and show their favorite three-to-five minute segments on the screen [...]

Table 7: Examples of AWARENESS: how likely is it that the bold argument is aware of being involved in the event?

7 Semantic Proto-Role Labeling

SRL systems are trained to predict either: (i) a predicate or frame specific notion of role (e.g., FrameNet), or (ii) a cross-predicate, shared notion of role (e.g., PropBank). (i) allows for fine-grain distictions specific to a single predicate, but risks data sparsity (needing many examples per predicate). (ii) allows for sharing statistics across predicates, but requires careful, manual cross-predicate

[15]The role via SemLink should be AGENT.

	instigated	volitional	awareness	sentient	moved
Null	57.4	59.0	58.4	74.0	64.1
Pos	79.0	74.9	75.5	74.1	64.1
Full	82.9	74.9	75.5	75.2	67.1
existed	created	destroyed	chg poss	chg state	stationary
57.8	69.4	80.4	72.4	45.9	65.0
59.7	69.6	80.7	72.4	46.5	65.1
64.8	72.0	82.3	72.3	58.0	69.0

Table 8: Test classification accuracies for each property.

analysis to ensure equivalent role-semantics (Loper et al., 2007), and as in seen Tables 4 and 5 it may not be feasible to ensure exact equivalence.

Our approach addresses this challenge by dropping the notion of categorical role entirely, replacing it with responses to proto-role questions that can be shared across all arguments[16] and predicates.[17] Further, as likelihood judgements may be interpreted as scalars, then this may provide a smoother representation for prediction and downstream use, akin to the recent push to replace categorical "1-hot" word representations with vector-space models.

As an example SPRL model, we trained separate log-linear classifiers with L_2 regularization on the judgments of each property in the results from Experiment 2. As in Fig. 6 we collapsed ratings to a categorical $\{1,3,5\}$, and included N/A, for a resultant 4-way classification problem.[18] The 9,778 arguments that appear in the dataset were divided into training (7,823), development (978), and test (977).

We trained three models: *Null*, with only an intercept feature[19]; *Pos*, which adds as a feature the linear offset of the argument relative to the verb (as a coarse proxy for syntactic structure); and Full, which added a vector embedding of the verb (Rastogi et al., 2015).[20] Even with this basic model we see evidence of learning property-specific distributions across verbal predicates, such as for CHANGED STATE.

[16]E.g., the notions of ACTOR, AGENT, and even PATIENT may overlap in their underlying properties.

[17]E.g., the proto-Agent of *build* will be related but not identical to that of *kill*: where commonalities exist, predictive models can benefit from the overlap.

[18]Future work on prediction may explore alternative formulations, such as a 2-step process of first predicting N/A, then performing regression on likelihood.

[19]The Null classifier predicts a rating of 1 for CREATED and DESTROYED and N/A for all the other properties.

[20]`http://cs.jhu.edu/~prastog3/mvlsa/`

8 Conclusions and Future Work

In this paper we have adopted from theoretical linguistics the idea that thematic roles should be decomposed into more fine-grained properties that have a prototype structure – the *Proto-role Hypothesis*. We developed an annotation task based on this idea, and tested it both in a smale scale nonce-based version and a very large scale version using real data from PropBank(/WSJ). One main result is that the proto-role hypothesis holds true at this very large scale. A second result is that, at this scale we gain evidence for a substantial amount of 'role fragmentation' in the lexicon of English: we find approximately 800 discrete property configurations. The proto-role approach allows us to cope with fragmentation by focusing on the fine-grained properties that make up roles. We showed this allows a greater degree of accuracy in role annotations, for example handling variability in fine-grained properties across tokens of a verb in a corpus that lead to coarse-grained categorization challenges. Finally, we have shown it practical to directly annotate a corpus with fine-grained properties and produced a large collection of such annotations, which we release to the community. We are currently expanding the annotation set beyond WSJ, and beyond English, as well as applying it to theoretical questions about verb class and argument structure (Davis and Koenig, 2000; Kako, 2006b), along with word sense. Finally, we are building on the baseline model in §7 to more broadly investigate how decompositional semantic annotations can guide linguistically motivated representation learning of meaning.

Acknowledgments Great thanks to Martha Palmer, Tim O'Gorman, Scott Grimm, and the reviewers for their feedback; Robert Busby for annotations; Julian Grove for work on a predecessor project with Kyle Rawlins. Thanks to Sanjeev Khudanpur, John J. Godrey, and Jan Hajič, as well as JHU and Charles University for coordinating the Fred Jelinek Memorial Workshop in 2014, supported by NSF PIRE (0530118). Support came from an NSF Graduate Research Fellowship, DARPA DEFT FA8750-13-2-001 (*Large Scale Paraphrasing for Natural Language Understanding*), the JHU HLTCOE, the Paul Allen Institute of Artificial Intelligence (*Acquisition and Use of Paraphrases in a Knowledge-Rich Setting*), and NSF BCS-1344269 (*Gradient Symbolic Computation*).

References

Afra Alishahi and Suzanne Stevenson. 2010. A computational model of learning semantic roles from child-directed language. *Language and Cognitive Processes*, 25(1):50–93.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the International Conference on Computational Linguistics*, pages 86–90.

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. Semeval'07 task 19: Frame semantic structure extraction. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 99–104. Association for Computational Linguistics.

Frank R. Blake. 1930. A semantic analysis of case. In James T. Hatfield, Werner Leopold, and A. J. Friedrich Ziglschmid, editors, *Curme volume of linguistic studies*, pages 34–49. Linguistic Society of America.

Claire Bonial, William Corvey, Martha Palmer, Volha V. Petukhova, and Harry Bunt. 2011. A hierarchical unification of LIRICS and VerbNet semantic roles. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 483–489, September.

Greg N. Carlson and Michael K. Tanenhaus. 1988. Thematic roles and language comprehension. In W. Wilkins, editor, *Syntax and Semantics: Thematic Relations*. Academic Press.

Greg N. Carlson. 1984. Thematic roles and their role in semantic interpretation. *Linguistics*, pages 259–279.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 152–164. Association for Computational Linguistics.

Hector-Neri Castañeda. 1967. Comments on Donald Davidson's 'The logical form of action sentences'. In Nicholas Rescher, editor, *The Logic of Decision and Action*. University of Pittsburgh Press.

Noam Chomsky. 1981. *Lectures on government and binding*. Foris Publications.

W. Croft. 1991. *Syntactic categories and grammatical relations*. University of Chicago Press.

D. Cruse. 1973. Some thoughts on agentivity. *Journal of Linguistics*, 9:1–204.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Proceedings of Human Language Technologies Conference of the North American Chapter of the Association for Computational Linguistics*, pages 948–956.

A. R. Davis and J.-P. Koenig. 2000. Linking as constraints on word classes in a hierarchical lexicon. *Language*, 76:56–91.

David Dowty. 1989. On the semantic content of the notion 'thematic role'. In Barbara Partee, Gennaro Chierchia, and Ray Turner, editors, *Properties, types and meanings, vol II*. Kluwer.

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

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Charles Fillmore. 1966. Towards a modern theory of case. The Ohio State University project on linguistic analysis report 13, The Ohio State University.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32.

Charles Fillmore. 1982. Frame semantics. *Linguistics in the morning calm*, pages 111–137.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Scott Grimm. 2005. *The lattice of case and agentivity*. MSc thesis, University of Amsterdam.

Scott Grimm. 2011. Semantics of case. *Morphology*, 21:515–544.

Jeffrey S. Gruber. 1965. *Studies in lexical relations*. Ph.D. dissertation, Massachusetts Institute of Technology.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. MIT Press.

Joshua K. Hartshorne, Claire Bonial, and Martha Palmer. 2014. The VerbCorner Project: Findings from Phase 1 of crowd-sourcing a semantic decomposition of verbs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 397–402. Association for Computational Linguistics.

Joshua Hartstone, Claire Bonial, and Martha Palmer. 2013. The VerbCorner project: Toward an empirically-based semantic decomposition of verbs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1438–1442.

Geoffrey E. Hinton. 1981. Implementing semantic networks in parallel hardware. In Geoffrey E. Hinton and John A. Anderson, editors, *Parallel Models of Associative Memory*. Erlbaum, Hillsdale, NJ.

Geoffrey E. Hinton. 1986. Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pages 1–12. Amherst, MA.

Ray S. Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. The MIT Press.

Ray S. Jackendoff. 1987. The status of thematic relations in linguistic theory. *Linguistic Inquiry*, 18:369–411.

Edward Kako. 2006a. The semantics of syntactic frames. *Language and Cognitive Processes*, 21(5):562–575.

Edward Kako. 2006b. Thematic role properties of subjects and objects. *Cognition*, 101:1–42.

Sandra Kübler and Heike Zinsmeister. 2015. *Corpus Linguistics and Linguistically Annotated Corpora*. Bloomsbury Publishing.

Beth Levin and Malka Rappaport Hovav. 2005. *Argument Realization*. Cambridge University Press.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Ken Litkowski. 2004. Senseval-3 task: Automatic labeling of semantic roles. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 141–146.

Edward Loper, Szu-Ting Yi, and Martha Palmer. 2007. Combining lexical resources: Mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics*.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

James L. McClelland and Alan H. Kawamoto. 1986. Mechanisms of sentence processing: Assigning roles to constituents. In David E. Rumelhart and James L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2*. MIT Press, Cambridge, MA.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31:71–106.

Terence Parsons. 1990. *Events in the Semantics of English*. MIT Press, Cambridge, MA.

James Pustejovsky, Patrick Hanks, and Anna Rumshisky. 2004. Automated induction of sense in context. In *Proceedings of the International Conference on Computational Linguistics*.

Malka Rappaport and Beth Levin. 1988. What to do with theta roles. In W. Wilkins, editor, *Syntax and semantics 21: Thematic Relations*, pages 7–36. Academic Press.

Malka Rappaport Hovav and Beth Levin. 1998. Building verb meanings. In M. Butts and W. Geuder, editors, *The Projection of Arguments: Lexical and Compositional Factors*, pages 97–134. CSLI Publications.

Pushpendre Rastogi, Benjamin Van Durme, and Raman Arora. 2015. Multiview LSA: Representation learning via generalized CCA. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics*.

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

Leonard Talmy. 1978. Figure and ground in complex sentences. In Joseph Greenberg, editor, *Universals of Human Language*, pages 625–649. Stanford University Press.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0 LDC2013T19. Linguistic Data Consortium, Philadelphia, PA.