Universal Decompositional Semantics on Universal Dependencies

Aaron Steven White Drew Reisinger Keisuke Sakaguchi Tim Vieira
Sheng Zhang Rachel Rudinger Kyle Rawlins Benjamin Van Durme
Johns Hopkins University

Abstract

We present a framework for augmenting data sets from the Universal Dependencies project with *Universal Decompositional Semantics*. Where the Universal Dependencies project aims to provide a syntactic annotation standard that can be used consistently across many languages as well as a collection of corpora that use that standard, our extension has similar aims for semantic annotation. We describe results from annotating the English Universal Dependencies treebank, dealing with word senses, semantic roles, and event properties.

1 Introduction

This paper describes the *Universal Decompositional Semantics* (Decomp) project, which aims to augments Universal Dependencies (UD) data sets with robust, scalable semantic annotations based in linguistic theory. The UD project[1] aims to provide (i) a syntactic dependency annotation standard that can be used consistently across many languages and (ii) a collection of corpora that use that standard (De Marneffe et al., 2014; Nivre et al., 2015). Decomp provides complementary semantic annotations that scale across different types of semantic information and different languages and can integrate seamlessly with any UD-annotated corpus.

Decomp has two mutually supportive tenets—*semantic decomposition* and *simplicity*. As we discuss further in the next section, these tenets allow us to collect annotations from everyday speakers of a language that are rooted in basic, commonsensical aspects of meaning and that can be straightforwardly explained and generally agreed upon in context.

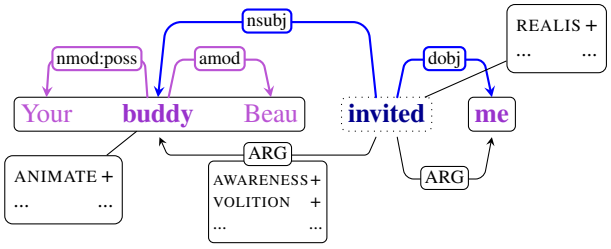

Figure 1: Decompositional semantics atop syntax.

In this paper, we describe Decomp protocols for three domains—semantic role decomposition, event decomposition, and word sense decomposition—and we present annotation results on top of the English UD v1.2 (EUD1.2) treebank.[2] We begin in §2 by connecting Decomp with previous work on decomposition in linguistic theory. In §3, we present *PredPatt*, which is a software package for preprocessing UD annotated corpora for input into Decomp protocols. In §4, we present a major revision to Reisinger et al.'s (2015) semantic role decomposition protocol (SPR1). Our revision, SPR2, brings SPR1 into full alignment with Decomp while adding various new features. In §5, we present Decomp-aligned annotation of event properties, focusing specifically on event realis. Finally, in §6, we describe a Decomp-aligned word sense decomposition protocol and associated set of annotations.

2 Universal Decompositional Semantics

A range of perspectives suggest that the proper representation for word meanings is *decompositional*.

[1] `http://universaldependencies.org`

[2] All datasets are available at `http://decomp.net`.

For example, Dowty (1979) followed by a substantial amount of research (e.g. Jackendoff 1990; Rappaport-Hovav and Levin 1998; Levin and Rappaport Hovav 2005) suggests that word meanings can be factored into (i) idiosyncratic, item-specific components and (ii) general components, such as CAUSATION, that are used across the lexicon. In the domain of thematic roles, Dowty (1991) argues that notions such as AGENT should be decomposed into simpler, more primitive properties such as volitional participation in an event. Pustejovsky (1991) decomposes word meanings into *qualia structures* that again incorporate more primitive properties of events and individuals. In spite of this wealth of theory, existing annotation protocols rarely take the idea into account, operating at the level that the above approaches decompose with very few exceptions (Greene and Resnik, 2009; Hartshorne et al., 2013; Reisinger et al., 2015). Decomp's premise is that a decompositional approach to large-scale annotation has benefits for both the annotation process and downstream uses of annotated data (cf. He et al. 2015; Bos et al. 2017 for recent non-decompositional approaches).

To capture these benefits, Decomp incorporates *semantic decomposition* directly into its protocols by mapping from decompositional theories, such as Dowty's, into straightforward questions on binary properties that are easily answered. This method of constructing annotation protocols gives rise to *simplicity* in the protocol, since the resulting questions are much more commonsensical and easily explained than the concepts they are decomposing. For instance, Dowty (1991) decomposes the relatively unintuitive notion (for ordinary speakers) of AGENT into much simpler properties, such as VOLITION and MOVEMENT. Instead of asking about whether a verbal argument is an AGENT (with the concomitant complex training process for annotators), a Decomp protocol might then ask about whether the referent of the argument had volition in or moved as a result of the event. Simplicity in the protocol in turn allows a Decomp protocol to gather annotations from untrained native (and naïve) speakers of a language. In large part, the current paper is focused on developing questions that everyday speakers can agree on and that are key for lexical representations.

An added benefit of questions on binary properties is they allow the use of ordinal prompts (Likert scales), allowing annotators to record subjective uncertainty of a property in a given context, which can then be aggregated across multiple responses with less severe impact to inter-annotator agreement.

3 Predicative patterns

In this section, we introduce PredPatt, which is a lightweight tool for identifying the structure of predicates and arguments from Universal Dependencies. We use the PredPatt's output as input to the Decomp-aligned annotation protocols we describe in §4 – §6. To ensure that this output is accurate, we evaluate on multiple UD-annotated corpora: automatically generated English parses and gold treebanks in Chinese, English, Hebrew, Hindi, and Spanish.[3]

PredPatt employs deterministic, unlexicalized rules over UD parses. We provide a high-level overview of the process here.[4]

Input UD Parse with Universal POS tags
1. Predicate and argument root identification
2. Argument resolution
3. Predicate and argument phrase extraction
4. Optional Post-processing

Output collection of predicate-argument structures

UD Parse A universal dependency (UD) parse, is a set of labeled pairs. Each pair has the form RELATION(DEPENDENT, GOVERNOR). The UD parse also includes a sequence of Universal POS tags. An example of a UD parse is in Figure 1.

Predicate and argument root identification Predicate and argument roots (i.e., dependency tree nodes) are identified by local configurations—specifically, edges in the UD parse. The simplest example is NSUBJ(s, v) and DOBJ(o, v), which indicate that v is a predicate root, and that s and o are argument roots. Similarly, roots of clausal subjects and clausal complements are also predicate roots. Nominal modifiers inside adverbial modifiers

[3] While we are not aware of a similar tool for Universal Dependencies, PredPatt is similar to ClausIE (Del Corro and Gemulla, 2013) and ArgOE (Gamallo et al., 2012), which supports Spanish, Portuguese, Galician and English.

[4] A detailed description of PredPatt is available at `https://github.com/hltcoe/PredPatt`. PredPatt derives from the system described by Rudinger and Van Durme 2014.

are arguments to the verb being modified, e.g., *Investors turned away from [the stock market]*. PredPatt also extracts relations from appositives, possessives, copula, and adverbial modifiers.

Argument resolution PredPatt includes argument resolution rules to handle missing arguments of many syntactic constructions, including predicate coordination, relative clauses, and embedded clauses. Argument resolution is crucial in languages that mark arguments using morphology, such as Spanish and Portuguese, because there are more cases of covert subjects. Other common cases for argument resolution are when predicates appear in a conjunction, e.g., *Chris likes to sing and dance*, has no arc from *dance* to its subject *Chris*. In relative clauses, the arguments of an embedded clause appear outside the subtree, e.g., *borrowed* in *The books John borrowed from the library are overdue.* has *books* as an argument and so does *are-overdue*.

Predicate extraction PredPatt extracts a descriptive name for *complex predicates*. For example, ***[PredPatt] finds [structure] in [text]*** has a 3-place predicate named (`?a` **finds** `?b` **in** `?c`). The primary logic here is (a) to lift mark and case tokens (e.g., *in*) out of the argument subtree, (b) to add adverbial modifiers, auxiliaries, and negation (e.g., ***[Chris] did not sleep quietly***). PredPatt uses the text order of tokens and arguments to derive a name for the predicate; no effort is made to further canonicalize this name, nor align it to a verb ontology.

Argument phrase extraction Argument extraction filters tokens from the dependency subtree below the argument root. These filters primarily simplify the subtree, e.g., removing relative clauses and appositives inside an argument. The default set of filters were chosen to preserve meaning, since it is not generally the case that all modifiers can safely be dropped (more aggressive argument simplification settings are available as options).

Post-processing PredPatt implements a number of optional post-processing routines, such as conjunction expansion, argument simplification (which filters out non-core arguments, leaving only subjects and objects), and language-specific hooks.[5]

Lang	#Sent	#Output	Precision
Chinese	98	375	69.1% ±4.7%
English	79	210	86.2% ±4.7%
Hebrew	12	30	66.7% ±17.9%
Hindi	22	50	52.0% ±14.3%
Spanish	27	55	70.9% ±12.4%

Table 1: Results of manual evaluation of PredPatt on UD

Gold treebanks in multiple languages We evaluated PredPatt manually on several randomly sampled sentences taken from the UD banks of Chinese, English, Hebrew, Hindi and Spanish. This evaluation runs PredPatt with the gold standard UD parse. We report the number of sentences evaluated along with the number of extractions from those sentences (a proxy for recall) and precision (95% confidence interval) for each language in Table 1.

4 Semantic role decomposition

A decompositional strategy has been successfully used by Reisinger et al. (2015) to annotate thematic role information under their Semantic Proto-Role labeling protocol (SPR1), which is based on Dowty's (1991) seminal thematic proto-role theory.[6]

In this section, we present a major revision to SPR1 aimed at strengthening and generalizing the protocol beyond Reisinger et al.'s dataset. We present three pilots aimed at validating our new protocol as well as a bulk annotation of a large subset of core arguments in EUD1.2. Finally, we describe, deploy, and validate methods for extending SPR2's reach beyond this subset, resulting in SPR2.1.

4.1 SPR1 protocol

In the SPR1 protocol, each core argument of a verb is annotated for the likelihood that particular properties hold of that argument's referent as a participant in the event denoted by the verb.

Property questions The properties selected for this purpose, given in Table 2, are based on those invoked by Dowty (1991) in his prototype-theoretic

[5]UD itself allows for language-specific exceptions to the "universal" standard, and we therefore allow that practice here.

[6]See Kako 2006; Greene and Resnik 2009; Madnani et al. 2010; Hartshorne et al. 2013 for work using similar protocols.

Role property	How likely or unlikely is it that...
instigation	ARG caused the PRED to happen?
volition	ARG chose to be involved in the PRED?
awareness	ARG was/were aware of being involved in the PRED?
sentient	ARG was/were sentient?
change of location	ARG changed location during the PRED?
-exists as physical	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?
change of possession	ARG changed possession during the PRED?
change of state	ARG was/were altered or somehow changed during or by the end of the PRED?
-stationary	ARG was/were stationary during the PRED?
-location of event	ARG described the location of the PRED?
-physical contact	ARG made physical contact with someone or something else involved in the PRED?
was used	ARG was/were used in carrying out the PRED?
-pred changed arg	The PRED caused a change in ARG?
+was for benefit	PRED happened for the benefit of ARG?
+partitive	*Only* a part or portion of ARG was involved in the PRED?
+change of state continuous	The change in ARG happened throughout the PRED?

Table 2: Questions posed to annotators. + indicates questions new to SPR2; - indicates SPR1 questions dropped in SPR2.

reconstruction of linking theory. Reisinger et al.'s (2015) SPR1 dataset, produced under this protocol, provides annotations of the Wall Street Journal portions of the Penn Treebank (PTB; Marcus et al. 1993) that are also annotated for core argument PropBank (Palmer et al., 2005) roles.

Filtering and data collection In Reisinger et al. 2015, verbs were excluded that occur in certain syntactic environments that interfere with property judgments. In particular, participles and imperatives were excluded, as well as verbs in embedded clauses, in questions, and under negation or auxiliaries. We carry these filters forward to our own bulk annotation by implementing them over PredPatt output. We show in §4.6 how these filters can be lifted.

Annotators To ensure internal consistency of the judgments, Reisinger et al.'s data was based on a single Amazon Mechanical Turk annotator.

4.2 SPR2 protocol

First we update both the set of questions and the method for presenting these questions in order to streamline the annotation process and simplify Reisinger et al.'s protocol. Second, to deal with potentially ungrammatical sentences, as well as to add an extra layer of quality control to the generation of property questions, we add an acceptability judgment question to the protocol. Finally, we collect annotations from multiple trusted annotators with two-way redundancy, allowing us to normalize the data in a way that is impossible with SPR1.

Property questions While Reisinger et al.'s properties were mainly motivated by linguistic theory, in the process of developing SPR2 we identified several redundancies as well as potential sources of error; these changes are summarized in Table 2. Redundancies include `stationary` being essentially the negation of `change of location`, and `predicate changed argument` being almost identical to `change of state`. The property `exists as physical` was dropped because it is a purely referential (non-relational) property of the argument; thus, it is redundant with our more elaborated decompositional word sense protocol. The `location of event` and `physical contact` properties were removed because of lower interannotator agreement and high within-annotator response variance in SPR1.

In addition to this streamlining, we added three new properties that target new types of arguments: benefactives, partitives, and incremental themes. Benefactive arguments and partitive arguments often have special morphosyntactic properties in many languages. In English, for example, benefactives can appear in double-object constructions with verbs like *buy*, and in many languages they correlate with special morphology. Partitives involve partial affectedness and similarly are often marked with morphological case (Kiparsky, 1998). The third new property, `change of state continuous`, is a plain-language version of Dowty's (1991) *incremental theme* proto-role property, which Reisinger et al. (2015) did not include. An argument is an INCREMENTAL THEME with respect to an event if the temporal progress of the event can be measured in terms of, or put into correspondence with, the part-whole structure of that argument which undergoes some gradual change (Tenny, 1987; Krifka, 1989, a.o.). For example, in an event of *mowing the lawn*, *the lawn* is an incremental theme because the progress of mowing is directly related to the portion of the lawn that has been mowed. Though incremental theme is quite abstract in comparison to other proto-role properties, it is widely agreed that something like this property is involved in linking thematic roles to syntactic position.

Dynamic reveal The question corresponding to the `change of state continuous` property

presupposes that the argument under consideration did, in fact, undergo some change of state. This means that if an annotator has previously determined that the property `change of state` does not apply, then asking about `change of state continuous` is at best inefficient, since we can deterministically predict that the answer should be *NA*, and at worst confusing to the annotator, since the question triggers a presupposition failure.

To avoid such presupposition failures in SPR2, which we suspect led to additional noise and annotation time as part of SPR1, we modified the annotation interface so that certain questions are revealed dynamically based on the answers to other questions. The set of questions is now organized hierarchically instead of as a flat list. In this hierarchical structure, `change of state` is a parent of `change of state continuous`, which means that the latter question only appears if the annotator gives a high ordinal value to the former. Questions that remain hidden are assumed to have *NA* as their answer. For SPR2, this pair of properties is the only one affected by the dynamic reveal feature, though this aspect of the protocol will be extended in later versions.

Acceptability judgments Two kinds of grammatical acceptability judgments were collected. The first kind, collected on a five-point scale, asked about the acceptability of the sentence containing the argument in question. The second kind, collected as a binary judgment, asked whether the question was hard to answer because of grammatical errors. This second was triggered only when annotators gave a response on the bottom three values of the ordinal scale for the relevant property question. We do not analyze these judgments here for reasons of space, but they are available as part of the released dataset.

Multiple annotators with redundancy Reisinger et al.'s (2015) reason for not using redundant annotations was that a single annotator would provide internally consistent judgments, but this consistency comes at the cost of potential bias in the judgments.[7] In order to evaluate bias, we move to

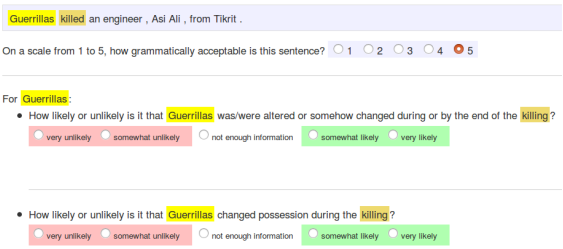

Figure 2: Example of semantic role decomposition task.

two-way redundancy (and later versions of the protocol are compatible with greater redundancy).

Heterogeneous data SPR2 extends the coverage of Semantic Proto-Role Labeling to heterogeneous genres. The SPR1 dataset contains only annotations of newswire text. This is not ideal for either practical or scientific purposes, since newswire tends to be biased toward otherwise rare word senses—often pertaining to financial markets—but low coverage of otherwise common word sense.

To remedy these coverage issues, we extend SPR1 to the English Universal Dependencies (version 1.2) treebank (EUD1.2). EUD1.2 is based on the Linguistic Data Consortium's English Web Treebank (Bies et al., 2012) and contains a much wider set of genres than the Penn Treebank—including weblogs, newsgroup discussions, emails from the EnronSent Corpus, reviews from English Google reviews, and answers from Yahoo! Answers. EUD1.2 has the added benefit of being natively annotated with gold-standard Universal Dependencies (UD) parses (Nivre et al., 2015).

4.3 Pilot experiments

In this section, we present three pilot experiments conducted on a subset of EUD1.2 and aimed at validating the updated protocol in preparation for deployment of the full task. In the first, we use the SPR1 protocol to obtain judgments on a small sample of EUD1.2 sentences from the same trusted annotator that produced SPR1. In the second, we open the same task to multiple annotators. And in the third, we deploy our updated SPR2 protocol on the

[7] For example, in analyzing the SPR1 dataset that Reisinger et al. make available, we noted that their annotator has a somewhat idiosyncratic way of answering the `was used` question, which aims at identifying instruments: the annotator marks *him* in (i) as likely to have been used in carrying out the advising.

(i) Sen. Bill Bradley of New Jersey *advised* **him** that the Dow Jones Industrial Average had declined by 190 points.

This is a general pattern for this question for this annotator.

same subset of EUD1.2—again, open to multiple annotators. We use these three pilots to evaluate interannotator agreement within and across protocols (where possible) and to construct a pool of trusted annotators to work on the full annotation task.

Item selection For each pilot, the same set of sentences were used. These sentences were selected based on properties of both the predicate and its corresponding arguments in each sentence. The pilot experiments were limited to the same 10 verbs (*want*, *put*, *think*, *see*, *know*, *look*, *say*, *take*, *tell*, *give*) that were considered in Reisinger et al.'s pilot.

As in Reisinger et al. 2015, tokens were excluded with verbs that occur in certain syntactic environments that interfere with property judgments. We used the same filters described by those authors, modified for UD. Additionally, verbs occurring as the second item in a conjunction were removed, as EUD1.2 does not have sufficient annotation to identify all arguments of such verbs from the syntax.

Verbal arguments were defined as the subtrees governed by a verb via a core grammatical relation (`nsubj`, `nsubpass`, `dobj`, and `iobj`). In addition, occurrences of the pronoun *it* in subject position were excluded because of inconsistencies in the annotation of expletive subjects in EUD1.2.

Pilots 1 & 2: SPR1 protocol Pilot 1, designed to compare SPR1 directly to SPR2, used the same protocol described in Reisinger et al. 2015 and was deployed on 99 argument tokens selected based on the method above.[8] To ensure that the only difference between SPR1 and this pilot was which sentences were annotated, we obtained the AMT identifier for the SPR1 annotator from Reisinger et al. Thus, the only annotator in this pilot was the same one that produced all the annotations for the SPR1 dataset.

The data from this pilot cannot be compared to the SPR1 dataset on a token level, since the items do not come from the same dataset. But these data can be compared to the SPR1 dataset on a type level by averaging responses to particular questions asked about particular argument positions (e.g., `nsubj`, `dobj`, etc.) for a particular predicate (e.g., *want*, *put*, etc.) and then comparing the correlation between these averages. The average type-level correlation between the average by-predicate, by-argument relation ratings in the SPR1 dataset and those in the current pilot was high for all verb-argument pairs (Spearman ρ=0.82).

Pilot 2 uses the same materials and protocol as Pilot 1. The only difference between the two is that this pilot was open to multiple annotators. A total of 33 annotators participated, one of whom was the same annotator that produced all the annotations for the SPR1 dataset and participated in Pilot 1.

For each argument token, we collected five judgments per property question. Interannotator agreement was calculated by argument token for the likelihood responses using pairwise Spearman rank correlations. The mean ρ across all annotator pairs and argument tokens was 0.562 (95% CI=[0.549, 0.574]) and, due to heavy left skew, the median was 0.618 (95% CI=[0.603, 0.631]). This agreement is relatively high, suggesting that different annotators tend to agree on the relative likelihood of a property applying to an argument.

Since the SPR1 and Pilot 1 annotator was among this group, we can also assess the extent to which the Pilot 1 annotator is consistent with other annotators. Comparing this annotator to every other annotator that annotated the same argument token, the mean ρ was 0.499 (95% CI=[0.451, 0.546]), and the median was 0.565 (95% CI=[0.504, 0.637]). This suggests that, on average, the other annotators are even more consistent with each other than they are with the original SPR1 annotator, vindicating the use of multiple annotators.

Pilot 3: SPR2 protocol Pilot 3 uses the same materials as Pilots 1 and 2 but introduces the SPR2 protocol laid out above. A total of 57 annotators participated in this pilot. For each argument token, we again collected five judgments per property.

Interannotator agreement was calculated by argument token for the likelihood responses using pairwise Spearman rank correlations. The mean ρ across all annotator pairs and argument tokens was 0.622 (95% CI=[0.610, 0.634]) and, again due to heavy left skew, the median was 0.677 (95% CI=[0.662, 0.690]). This higher agreement compared to Pilot 2 likely arises due to the fact that we have fewer questions in the SPR2 protocol and suggests that we suc-

[8] For each verb, 10 arguments were selected, with the exception of *see*, which only had 9 due to an off-by-one error.

ceeded in removing noisy questions without adding questions that were similarly noisy.

Since we use the same materials as Pilots 1 and 2, we can also compare the SPR1 and SPR2 protocols on the subset of questions they share. We find similar mean agreement, at 0.593 (95% CI=[0.580, 0.607]), and median agreement, at 0.665 (95% CI=[0.652, 0.672]), to that we found within the Pilots 2 and 3 results. This suggests that the addition and subtraction of questions does not substantially alter annotators' judgments on the questions that both protocols share.

4.4 Trusted annotator pool

To ensure annotation consistency in our bulk annotation, we constructed a pool of trusted annotators from those annotators that participated in Pilots 2 and 3. We used two metrics to construct this pool: rating agreement and applicability agreement. Both of these metrics control for various factors that might raise or lower agreement independent of the annotator—e.g., the particular question, the particular sentence, the particular argument type, etc.—using generalized linear mixed effects models. This pool contains a total of 86 trusted annotators.

4.5 Bulk task

For our bulk task, we used the SPR2 protocol to annotate a total of 3,806 argument tokens spanning 2,759 unique predicate lemmas. These argument tokens were part of a filtered set constructed using Reisinger et al.'s filtering scheme described above.

We collected two judgments per property, per argument token. Interannotator agreement was calculated in the same way as for the pilots. The mean ρ was 0.617 (95% CI=[0.611, 0.623]), and the median was 0.679, (95% CI=[0.673, 0.686]). This agreement is very close to that found in the pilots, suggesting that rating consistency extends beyond the constrained set of predicates used in the pilots.

4.6 Beyond filters

One issue with SPR1 that remains unaddressed in SPR2 is the use of filters. This significantly reduces the potential coverage of the protocol and relies on extremely rich syntactic annotation. This second is not problematic when we have gold standard treebanks like EUD1.2, but it becomes an issue when moving beyond such treebanks.

To alleviate this filter issue, we propose a further revision of SPR2. In this version (SPR2.1), we alter the SPR2 instructions to take into account cases where the property questions may be difficult to answer. These fall into at least three categories: eventualities that haven't happened (*irrealis eventualities*), generics, and habituals. In SPR2.1, annotators are instructed about each case and to answer as if a specific event of that kind did actually happen.

We annotated predicates that occurred in a sentence from the previous bulk task but were filtered from that task based on Reisinger et al.'s (2015) filters. We have so far annotated all such predicates with less than 100 instances in all of EUD1.2 and plan to continue annotation to get full coverage of these sentences.

A total of 26 annotators from our trusted pool participated in this annotation. As in the previous bulk task, we collected two judgments per property, per argument token. The interannotator agreement was calculated in the same way as for the previous bulk task and pilots and was reasonably high with a mean ρ of 0.528 (95% CI=[0.522, 0.535]) and median ρ of 0.571 (95% CI=[0.563, 0.580]). This somewhat lower agreement is to be expected, since these predicates were selected to be harder than those in the previous task.

4.7 Discussion

We presented a major revision to Reisinger et al.'s (2015) decompositional Semantic Proto-Role Labeling protocol (SPR1) and deployed this revised protocol (SPR2) in three validation pilots and a bulk task. We then described two extensions to this protocol aimed at expanding the annotable arguments.

One issue that arises with SPR2.1 is that it substantially complicates the instructions, clashing with Decomp *simplicity* tenet. In the next section, we describe a task aimed at allowing us to better target predicates that need these more elaborated instructions, allowing us to use the simpler SPR2 protocol where possible.

5 Event decomposition

As discussed in §4, SPR1 and SPR2 employ filters that run on top of dependency parses to ensure that

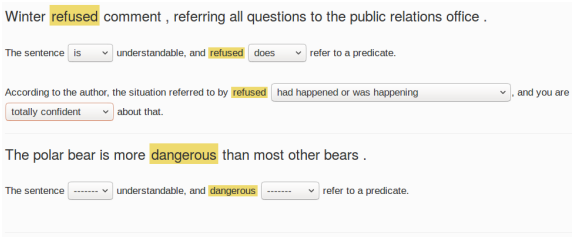

Figure 3: Example of the event decomposition task

proto-role property questions about particular arguments are answerable. We showed that these filters can be bypassed by altering the instructions given to annotators. This approach substantially increases the length of the tasks instructions, however, and so ideally, these lengthened instructions should be used only when absolutely necessary. One place it seems likely to be necessary is when the event in a sentence did not in fact occur.

In this section, we present a protocol, inspired by the one developed by de Marneffe et al. (2012), for targeting these sorts of sentence with special instructions in future versions of SPR (see also Saurí and Pustejovsky 2012). A major benefit of this protocol is that it produces a foundation for future decompositional event annotations.

Protocol The protocol has four major components: questions about (i) whether or not a particular word refers to an eventuality (event or state); (ii) whether the sentence is understandable; (iii) whether or not, according to the author, the event has already happened is currently happening; and (iv) how confident the annotator is about their answer to (iii).

The first two components were included to filter out items that are either incorrectly labeled as predicates or that the annotator could not annotate for components (iii) and (iv), and if an annotator answered *no* to either for a particular predicate candidate, (iii) and (iv) did not appear. Thus, like SPR2.x, this protocol incorporates a hierarchy of questions that can be elaborated in future versions.

Data collection We applied this protocol to every predicate candidate found in an EUD1.2 sentence annotated under SPR2 and SPR2.1. This yields annotations for a superset of the predicates annotated under SPR2.x, and thus components (i) and (ii) of these annotations can be used as a *post hoc* filter on the SPR2.x annotations or to decide on whether to include a predicate for future SPR2.x tasks.

A total of 6,930 predicate candidates were annotated in batches of 10 by 24 unique annotators recruited from the trusted annotator pool built for SPR2.x. Each predicate candidate was judged by two distinct annotators.

Data validation For each of the four components interannotator agreement was computed by each group of 10 predicates. For the categorical responses, we would ideally use Cohen's κ, but there were so many cases of perfect agreement for the categorical responses that Cohen's κ is ill-defined in many cases. As such, we report raw agreement here.

The mean raw agreement for whether each predicate candidate was a predicate was 0.955 (95% CI=[0.950, 0.960]). The mean raw agreement for whether the sentence was understandable was [0.976, (95% CI=[0.971, 0.980]); and the mean raw agreement for whether the eventuality happened or was happening was 0.820 (95% CI=[0.811, 0.829]).

Discussion We presented the first version of a new event decomposition protocol. This protocol integrates with and is in the same spirit as the SPR2.x protocols produced in the previous section.

In the next section, we describe a complementary protocol for decomposing word sense, focusing specifically on noun senses. This last protocol completes a picture wherein we decompose predicate argument semantics into three parts: the properties of a predicate independent of its arguments, the properties of a predicate's arguments in relation to the event the predicate denotes, and the properties of an argument independent of the predicate.

6 Word sense decomposition

In §4 and §5, we focused on semantic questions that deal with eventualities. In this section, we describe a Decomp protocol for decomposing word sense. Our goal is similar to that in previous sections: elicit responses from everyday speakers of the language regarding basic properties, in relation to the context of a natural language sentence.

Protocol If directly following the strategy explored thus far, we would create an interface that enumerated many dozens (or hundreds) of semantic properties one might ask about a word in context,

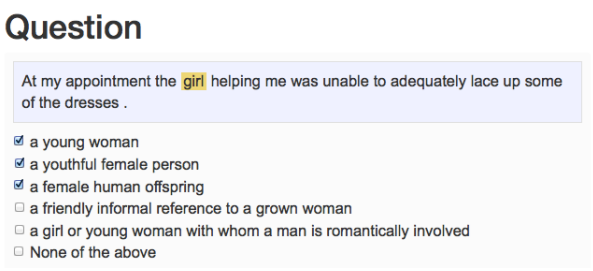

Figure 4: Example of the word sense task.

and in further developments of Decomp there may be specific properties that are deemed essential for direct querying of annotators. However, here we rely on the rich pre-existing taxonomy of lexical knowledge captured in the WordNet hierarchy (Miller, 1995) in order to more efficiently gather implicit property responses. Everyday speakers can perform basic word sense disambiguation (Snow et al., 2008): this falls under the simplicity tenant of Decomp. Once a word is disambiguated in context we then can infer automatically whether an instance is, e.g., a `physical object`.

While WordNet is a valuable resource, the selection of a specific categorical sense under an enumerated set of prespecified options is troubling in a similar way as Dowty was concerned with thematic roles (see Kilgarriff 1997). Therefore we follow a path similar to Sussna (1993) in asking annotators for zero or more senses that are appropriate.[9]

Candidate senses are extracted from WordNet *synsets*. We have grounded argument tokens in WordNet in order to make efficient use of existing lexical semantic resourses, but this protocol could in principle be used with any other lexical semantic resource. We believe these annotations will be useful already in the context of the other annotations, but in addition, future work will use these sense groundings to derive commonsense properties beyond those directly encoded in the WordNet hierarchy.

Data collection A total of 18,054 word tokens (arguments) in 10,833 total sentences extracted from EUD1.2 were annotated for sense by at least three annotators recruited from Amazon Mechanical Turk. Each token had an average of 5.63 candidates senses for annotators to choose from (Figure 4). In total, 1,065 unique annotators participated.

Data validation Inter-annotator agreement was computed by lemma by taking the Jaccard index for each pair of annotators that judged the senses for that lemma: $\frac{\text{\# of senses checked by both annotators}}{\text{\# of senses checked by either annotator}}$. The overall inter-annotator agreement using this measure was 0.592: this is reasonably high considering the extremely low chance-level.

In total, 9,317 token-sense pairs were agreed upon by all annotators. We refer to these token-sense pairs as *gold word sense(s)* for the token. If we relax the agreement threshold for a token-sense pair to be gold to 0.5—i.e. half or more annotators agreed on that pair—the number of *gold word sense(s)* goes up to 27,326. Out of 18,054 individual arguments, 8,553 of them have a single *gold word sense* and 370 have two or more *gold word senses* as in the example in Figure 4. Similarly, if we relax the agreement threshold down to 0.5, 9,656 arguments have a single *gold word sense* and 17,281 arguments have two or more *gold word senses*.

7 Conclusion

We have described the *Universal Decompositional Semantics* (Decomp) project, which aims to construct and deploy a set of cross-linguistically robust semantic annotation protocols that are based in linguistic theory and that integrate seamlessly with the Universal Dependencies project. We then proposed Decomp-aligned protocols for three domains—semantic role decomposition, event decomposition, and word sense decomposition—and presented annotations, all freely available, that use these protocols and are constructed on top of the English UD v1.2 treebank. In future work, we intend to further revise and extend these protocols as well as produce novel protocols aligned with Decomp.

Acknowledgments

This research was supported by the Human Language Technology Center of Excellence (HLTCOE), DARPA DEFT, DARPA LORELEI, and NSF INSPIRE BCS-1344269 (*Gradient symbolic computation*). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not

[9] Not only does this weaken the commitment to a single categorical meaning, but it also reduces concerns of annotators being confused by overly fine-grain definitions (Navigli, 2006).

be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

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

Johan Bos, Valerio Basile, Kilian Evang, Noortje Venhuizen, and Johannes Bjerva. The Groningen Meaning Bank. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*. Springer, Berlin, 2017.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333, 2012.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of LREC*, volume 14, pages 4585–4592, 2014.

Luciano Del Corro and Rainer Gemulla. Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*, pages 355–366, 2013.

David Dowty. *Word Meaning and Montague Grammar*. D. Reidel Publishing Company, 1979.

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

Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. Dependency-based open information extraction. In *Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, pages 10–18, 2012.

Stephan Greene and Philip Resnik. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511. Association for Computational Linguistics, 2009. ISBN 1-932432-41-8.

Joshua K. Hartshorne, Claire Bonial, and Martha Palmer. The VerbCorner Project: Toward an Empirically-Based Semantic Decomposition of Verbs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1438–1442, 2013.

Luheng He, Mike Lewis, and Luke Zettlemoyer. Question-Answer Driven Semantic Role Labeling: Using Natural Language to Annotate Natural Language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, 2015.

Ray S. Jackendoff. *Semantic Structures*. MIT Press, 1990.

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

Adam Kilgarriff. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113, 1997.

Paul Kiparsky. Partitive case and aspect. *The Projection of Arguments: Lexical and compositional factors*, 265:307, 1998.

Manfred Krifka. Nominal reference, temporal constitution, and quantification in event semantics. In R. Bartsch, J. van Benthem, and P. von Emde Boas, editors, *Semantics and Contextual Expression*. Foris Publications, 1989.

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

Nitin Madnani, Jordan Boyd-Graber, and Philip Resnik. Measuring transitivity using untrained annotators. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 188–194, 2010.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2): 313–330, June 1993. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=972470.972475`.

George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995. ISSN 0001-0782. doi: 10.1145/219717.219748. URL `http://doi.acm.org/10.1145/219717.219748`.

Roberto Navigli. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 105–112, Sydney, Australia, July 2006. Association for Computational Linguistics. doi: 10.3115/1220175.1220189. URL http://www.aclweb.org/anthology/P06-1014.

Joakim Nivre, Željko Agić, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Cristina Bosco, Sam Bowman, Giuseppe G. A. Celano, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Daniel Galbraith, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Berta Gonzales, Bruno Guillaume, Jan Hajič, Dag Haug, Radu Ion, Elena Irimia, Anders Johannsen, Hiroshi Kanayama, Jenna Kanerva, Simon Krek, Veronika Laippala, Alessandro Lenci, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Shunsuke Mori, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Prokopis Prokopidis, Sampo Pyysalo, Loganathan Ramasamy, Rudolf Rosa, Shadi Saleh, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Kiril Simov, Aaron Smith, Jan Štěpánek, Alane Suhr, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Sumire Uematsu, Larraitz Uria, Viktor Varga, Veronika Vincze, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. Universal Dependencies 1.2. *http://universaldependencies.github.io/docs/*, November 2015. URL https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1548.

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

James Pustejovsky. The generative lexicon. *Computational Linguistics*, 17(4):409–441, 1991.

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

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

Rachel Rudinger and Benjamin Van Durme. Is the stanford dependency representation semantic? In *ACL Workshop: EVENTS*, 2014.

Roser Saurí and James Pustejovsky. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299, 2012.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D08-1027.

Michael Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of the Second International Conference on Information and Knowledge Management*, CIKM '93, pages 67–74, New York, NY, USA, 1993.

Carol Lee Tenny. *Grammaticalizing aspect and affectedness*. Thesis, Massachusetts Institute of Technology, 1987. URL http://dspace.mit.edu/handle/1721.1/14704.