The VerbCorner Project: Findings from Phase 1 of Crowd-Sourcing a Semantic Decomposition of Verbs

Joshua K. Hartshorne
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139, USA
jkhartshorne@gmail.com

Claire Bonial, Martha Palmer
Department of Linguistics
University of Colorado at Boulder
Hellems 290, 295 UCB
Boulder, CO 80309, USA
{CBonial, MPalmer}@colorado.edu

Abstract

Any given verb can appear in some syntactic frames (*Sally broke the vase*, *The vase broke*) but not others (**Sally broke at the vase*, **Sally broke the vase to John*). There is now considerable evidence that the syntactic behaviors of some verbs can be predicted by their meanings, and many current theories posit that this is true for most if not all verbs. If true, this fact would have striking implications for theories and models of language acquisition, as well as numerous applications in natural language processing. However, empirical investigations to date have focused on a small number of verbs. We report on early results from VerbCorner, a crowd-sourced project extending this work to a large, representative sample of English verbs.

1 Introduction

Verbs vary in terms of which syntactic frames they can appear in (Table 1). In principle, this could be an unpredictable fact about the verb that must be acquired, much like the phonological form of the verb.

However, most theorists posit that there is a systematic relationship between the semantics of a verb and the syntactic frames in which it can appear (Levin and Hovav, 2005). For instance, it is argued that verbs like *break*, which describe a caused change of state, can appear in both the NP V NP form (*Sally broke the vase*) and the NP V form (*The vase broke*). Verbs such as *hit* and *like* do not describe a change of state and so cannot appear in both forms.[1] Similarly, only verbs that describe propositional attitudes, such as *like*, can take a *that* complement (*John liked that Sally broke the vase*).

Frame	hit	like	break
NP V NP	x	x	x
NP V	-	-	x
NP that S	-	x	-
NP V at NP	x	-	-

Table 1: Some of the syntactic frames available for *hit*, *like*, and *break*.

1.1 The Semantic Consistency Hypothesis

This account has a natural consequence, which we dub the Semantic Consistency Hypothesis: There is some set of semantic features such that verbs that share the same syntactic behavior are identical along those semantic features.[2] Note that on certain accounts, this is a strong tendency rather than a strict necessity (e.g., Goldberg, 1995).

It is widely recognized that a principled relationship between syntax and semantics would have broad implications. It is frequently invoked in theories of language acquisition. For instance, Pinker (1984, 1989) has described how this correspondence could solve long-standing puzzles about how children learn syntax in the first place. Conversely, Gleitman (1990) has shown such a syntax-semantics relationship could solve significant problems in vocabulary acquisition. In fact, both researchers argue that a principled relationship between syntax and semantics is necessary for language to be learnable at all.

In computational linguistics and natural language processing, some form of the Semantic Consistency Hypothesis is often included in linguistic resources and utilized in applications. We describe in detail one such resource, VerbNet,

[1] Note that this is a simplification in that there are non-causal verbs that appear in both the NP V NP frame and the NP V frame. For details, see (Levin, 1993).

[2] There is a long tradition of partitioning semantics into those aspects of meaning which are "grammatically relevant" and those which are not. We refer the interested reader to Pinker (1989), Jackendoff (1990), and Levin & Rappaport Hovav (2005).

which is highly relevant to our investigation.

1.2 VerbNet

VerbNet (Kipper et al., 2008; based on Levin, 1993) lists over 6,000 verbs, categorized into 280 classes according to the syntactic frames they can appear in. That is, all verbs in the same class appear in the same set of syntactic frames. Importantly, in addition to characterizing the syntactic frames associated with each class, VerbNet also characterizes the semantics of each class.

For instance, class 9.7, which comprises a couple dozen verbs, allows 7 different syntactic frames. The entry for one frame is shown below:

Syntactic Frame NP V NP PP.DESTINATION
Example Jessica sprayed the wall.
Syntax AGENT V THEME {+LOC|+DEST_CONF} DESTINATION
Semantics MOTION(DURING(E), THEME)
NOT(PREP(START(E), THEME, DESTINATION))
PREP(END(E), THEME, DESTINATION)
CAUSE(AGENT, E)

Importantly, the semantics listed here is not just for the verb *spray* but applies to all verbs from the Spray Class whenever they appear in that syntactic frame – that is, VerbNet assumes the Semantic Consistency Hypothesis.

VerbNet and its semantic features have been used in a variety of NLP applications, such as semantic role labeling (Swier and Stevenson, 2004), inferencing (Zaenen et al., 2008), verb classification (Joanis et al., 2008), and information extraction (Maynard et al., 2009). It has also been employed in models of language acquisition (Parisien and Stevenson, 2011; Barak et al., 2012). In general, there has been interest in the NLP literature in using these syntactially-relevant semantic features for shallow semantic parsing (e.g., Giuglea and Moschitti, 2006).

2 Empirical Status of the Semantic Consistency Hypothesis

Given the prominence of the Semantic Consistency Hypothesis in both theory and practice, one might expect that it was on firm empirical footing. That is, ideally there would be some database of semantic judgments for a comprehensive set of verbs from each syntactic class. In principle, these judgments would come from naive annotators, since researchers' intuitions about subtle judgments may be unconsciously clouded by theoretical commitments (Gibson and Fedorenko, 2013). The Semantic Consistency Hypothesis would be supported if, within that database, predicates with the same syntactic properties were systematically related semantically.

No such database exists, whether consisting of the judgments of linguists or naive annotators. Most theoretical studies report researcher judgments for only a handful of examples; how many additional examples were considered by the researcher goes unreported. In any case, to our knowledge, of the 280 syntactic verb classes listed by VerbNet, only a handful have been studied in any detail.

The strongest evidence comes from experimental work on several so-called alternations (the passive, causative, locative, and dative alternations). Here, there does appear to be a systematic semantic distinction between the two syntactic frames in each alternation, at least most of the time. This has been tested with a reasonable sample of the relevant verbs and also in both children and adults (Ambridge et al., 2013; Pinker, 1989). However, the relevant verbs make up a tiny fraction of all English verbs, and even for these verbs, the syntactic frames in question represent only a fraction of the syntactic frames available to those verbs.

This is not an accidental oversight. The limiting factor is scale: with many thousands of verbs and over a hundred commonly-discussed semantic features and syntactic frames, it is not feasible for a single researcher, or even team of researchers, to check which verbs appear in which syntactic frames and carry which semantic entailments. Collecting data from naive subjects is even more laborious, particularly since the average Man on the Street is not necessarily equipped with metalinguistic concepts like *caused change of state* and *propositional attitude*. The VerbCorner Project is aimed at filling that empirical gap.

3 VerbCorner

The VerbCorner Project[3] is devoted to collecting semantic judgments for a comprehensive set of verbs along a comprehensive set of theoretically-relevant semantic dimension. These data can be used to test the Semantic Consistency Hypothesis. Independent of the validity of that hypothesis, the

[3] http://gameswithwords.org/VerbCorner/

semantic judgments themselves should prove useful for any study of linguistic meaning or related application.

We address the issue of scale through crowdsourcing: Recruiting large numbers of volunteers, each of whom may provide only a few annotations. Several previous projects have successfully crowd-sourced linguistic annotations, such as Phrase Detectives, where volunteers have contributed 2.5 million judgments on anaphoric relations (Poesio et al., 2012).

3.1 Integration with VerbNet

One significant challenge for any such project is first classifying verbs according to the syntactic frames they can appear in. Thus, at least initially, we are focusing on the 6,000+ verbs already cataloged in VerbNet. As such, the VerbCorner Project is also verifying and validating the semantics currently encoded in VerbNet. VerbNet will be edited as necessary based on the empirical results.

Integration with VerbNet has additional benefits, since VerbNet itself is integrated with a variety of linguistic resources, such as PropBank and Penn TreeBank. This amplifies the impact of any VerbCorner-inspired changes to VerbNet.

3.2 The Tasks

We selected semantic features of interest based on those most commonly cited in the linguistics literature, with a particular focus on those that – according to VerbNet – apply to many predicates.

Previous research has shown that humans find it easier to reason about real-world scenarios than make abstract judgments (Cosmides and Tooby, 1992). Thus, for each feature (e.g., MOVEMENT), we converted the metalinguistic judgment ("Does this verb entail movement on the part of some entity?") into a real-world problem.

For example, in "Simon Says Freeze," a task designed to elicit judgments about movement, the Galactic Overlord (Simon) decrees "Galactic Stay Where You Are Day," during which nobody is allowed to move from their current location. Participants read descriptions of events and decide whether anyone violated the rule.

In "Explode on Contact," designed to elicit judgments about physical contact, objects and people explode when they touch one another. The participant reads descriptions of events and decides whether anything has exploded.

Note that each task is designed to elicit judgments about entailments – things that must be true rather than are merely likely to be true. If John greeted Bill, they might have come into contact (e.g., by shaking hands), but perhaps they did not. Previous work suggests that it is the semantic *entailments* that matter, particularly for explaining the syntactic behavior of verbs (Levin, 1993).

3.3 The Items

The exact semantics associated with a verb may depend on its syntactic frame. Thus *Sally rolled the ball* entails that somebody applied force to the ball (namely: Sally), whereas *The ball rolled* does not. Thus, we investigate the semantics of each verb in each syntactic frame available to it (as described by VerbNet). Below, the term *item* is the unit of annotation: a verb in a frame.

In order to minimize unwanted effects of world knowledge, the verb's arguments are replaced with nonsense words or randomly chosen proper names (*Sally sprayed the dax onto the blicket*). The use of novel words is explained by the story for each task.

3.4 The Phases

Given the sheer scale of the project, data-collection is expected to take several years at least. Thus, data-collection has been broken up into a series of phases. Each phase focuses on a small number of classes and/or semantic entailments. This ensures that there are meaningful intermediate results that can be disseminated prior to the completion of the entire project. This manuscript reports the results of Phase 1.

4 Results

The full data and annotations will be released in the near future and may be available now by request. Below, we summarize the main findings thus far.

4.1 Description of Phase 1

In Phase 1 of the project, we focused on 11 verb classes (Table 3) comprising 641 verbs and seven different semantic entailments (Table 2). While six of these entailments were chosen from among those features widely believed to be relevant for syntax, one was not: A Good World, which investigated evaluation (*Is the event described by the verb positive or negative?*). Although evaluation

Task	Semantic Feature	Anns.	Anns./Item	Mode	Consistency
Entropy	PHYSICAL CHANGE	23,875	7	86%	95%
Equilibrium	APPLICATION OF FORCE	27,128	8	79%	95%
Explode on Contact	PHYSICAL CONTACT	23,590	7	93%	95%
Fickle Folk	CHANGE OF MENTAL STATE	16,466	5	81%	96%
Philosophical Zombie Hunter	MENTAL STATE	24,592	7	80%	89%
Simon Says Freeze	LOCATION CHANGE	24,245	7	83%	88%
A Good World	EVALUATION	22,668	7	72%	74%

Table 2: Respectively: Task, semantic feature tested, number of annotations, mean number of annotations per item, mean percentage of participants choosing the modal response, consistency within class.

of events is an important component of human psychology, to our knowledge no researcher has suggested that it is relevant for syntax. As such, this task provides a lower bound for how much semantic consistency one might expect within a syntactic verb class.

In all, we collected 162,564 judgments from 1,983 volunteers (Table 2).

4.2 Inter-annotator Agreement

Each task had been iteratively piloted and redesigned until inter-annotator reliability was acceptable, as described in a previous publication. However, these pilot studies involved a small number of items which were coded by all annotators. How good was the reliability in the crowdsourcing context?

Because we recruited large numbers of annotators, most of whom annotated only a few items, typical measures of inter-annotator agreement such as Cohen's *kappa* are not easily calculated. Instead, for each item, we calculated the most common (modal) response. We then considered what proportion of all annotations were accounted for by the modal response: a mean of 100% would indicate that there was no disagreement among annotators for any item.

As can be seen in Table 2, for every task, the modal response covered the bulk responses, ranging from a low of 72% for EVALUATION to a high of 93% for PHYSICAL CONTACT. Since there were typically 4 or more possible answers per item, inter-annotator agreement was well above chance. This represents good performance given that the annotators were entirely untrained.

In many cases, annotator disagreement seems to be driven by syntactic constructions that are only marginally grammatical. For instance, inter-annotator agreement was typically low for class 63. VerbNet suggests two syntactic frames for class 63, one of which (NP V THAT S) appears to be marginal (?*I control that Mary eats*). In fact, annotators frequently flagged these items as ungrammatical, which is a valuable result in itself for improving VerbNet.

Class	Examples	PChange	Force	Contact	MChange	Mental	LChange
12	yank, press	-	x	d	-	-	d
18.1	hit, squash	d	x	d	-	-	d
29.5	believe, conjecture	-	-	-	-	d	-
31.1	amuse, frighten	-	-	-	x	d	-
31.2	like, fear	-	-	-	-	x	-
45.1	break, crack	x	d	d	-	-	d
51.3.1	bounce, roll	-	d	d	-	-	d
51.3.2	run, slink	-	d	-	-	-	d
51.6	chase, follow	-	-	-	-	-	d
61	attempt, try	-	-	-	-	-	-
63	control, enforce	-	-	-	-	-	-

Table 3: VerbNet classes investigated in Phase 1, with presence of semantic entailments as indicated by data. *x* = feature present; - = feature absent; *d* = depends on syntactic frame.

4.3 Testing the Semantic Consistency Hypothesis

4.3.1 Calculating consistency

We next investigated whether our results support the Semantic Consistency Hypothesis. As noted above, the question is not whether all verbs in the same syntactic class share the same semantic entailments. Even a single verb may have different semantic entailments when placed in different syntactic frames. Thus, calculating consistency of a *class* must take differing *frames* into account.

There are many sophisticated rubrics for calculating consistency. However, for expository purposes here, we use one that is intuitive and easy to interpret. First, we determined the annotation for each item (i.e., each verb/frame combination) by majority vote. We then considered how many verbs in each class had the same annotation in any given syntactic frame.

For example, suppose a class had 10 verbs and 2 frames. In the first frame, 8 verbs received the same annotation and 2 received others. The consistency for this class/frame combination is 80%. In the second frame, 6 verbs received the same annotation and 4 verbs received others. The consistency for this class/frame combination is 60%. The consistency for the class as a whole is the average across frames: 70%.

4.3.2 Results

Mean consistency averaged across classes is shown for each task in Table 2. As expected, consistency was lowest for EVALUATION, which is not expected to necessarily correlate with syntax. Interestingly, consistency for EVALUATION was nonetheless well above floor. This is perhaps not surprising: two sentences that have the same values for PHYSICAL CHANGE, APPLICATION OF FORCE, PHYSICAL CONTACT, CHANGE OF MENTAL STATE, MENTAL STATE, and LOCATION CHANGE are, on average, also likely to be both good or both bad.

Consistency was much higher for the other tasks, and in fact was close to ceiling for most of them. It remains to be seen whether the items that deviate from the mode represent true differences in semantics or reflect merely noise. One way of addressing this question is to collect additional annotations for those items that deviate from the mode.

4.4 Verb semantics

For each syntactic frame in each class, we determined the most common annotation. This is summarized in Table 3. The semantic annotation depended on syntactic frame nearly 1/4 of the time.[4]

These frequently matched VerbNet's semantics, though not always. For instance, annotators judged that class 18.1 verbs in the NP V NP PP.INSTRUMENT entailed movement on the part of the instrument (*Sally hit the ball with the stick*) – something not reflected in VerbNet.

5 Conclusion and Future Work

Results of Phase 1 provide support for the Semantic Consistency Hypothesis, at least as a strong bias. More work will be needed to determine the strength of that bias. The findings are largely consistent with VerbNet's semantics, but changes are indicated in some cases.

We find that inter-annotator agreement is sufficiently high that annotation can be done effectively using the modal response with an average of 6-7 responses per item. We are currently investigating whether we can achieve better reliability with fewer responses per item by taking into account an individual annotator's history across items, as recent work suggests is possible (Passonneau and Carpenter, 2013; Rzhetsky et al., 2009; Whitehill et al., 2009).

Thus, crowd-sourcing VerbNet semantic entailments appears to be both feasible and productive. Data-collection continues. Phase 2, which added over 10 new verb classes, is complete. Phase 3, which includes both new classes and new entailments, has been launched.

Acknowledgments

We gratefully acknowledge the support of the National Science Foundation Grant NSF-IIS-1116782, DARPA Machine Reading FA8750-09-C-0179, and funding from the Ruth L. Kirschstein National Research Service Award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[4]Note that this table was calculated based on whether the semantic feature was present or not. In many cases, the data was significantly richer. For instance, for APPLICATION OF FORCE, annotators determined which participant in the event was applying the force.

References

Ben Ambridge, Julian Pine, Caroline Rowland, Franklin Chang, and Amy Bidgood. 2013. The retreat from overgeneralization in child language acquisition: word learning, morphology and verb argument structure. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(1):47–62.

Libby Barak, Afsaneh Fazly, and Suzanne Stevenson. 2012. Modeling the acquisition of mental state verbs. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–10. Association for Computational Linguistics.

Leda Cosmides and John Tooby. 1992. Cognitive adaptations for social exchange. *The Adapted Mind*, pages 163–228.

Edward Gibson and Evelina Fedorenko. 2013. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, 28(1-2):88–124.

Ana-Maria Giuglea and Alessandro Moschitti. 2006. Shallow semantic parsing based on framenet, verbnet and propbank. In *Proceedings of the 217th European Conference on Artificial Intelligence*, pages 563–567, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Lila Gleitman. 1990. The structural sources of verb meanings. *Language Acquisition*, 1(1):3–55.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar approach to argument structure*. University of Chicago Press.

Eric Joanis, Suzanne Stevenson, and David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3):337–367.

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.

Diana Maynard, Adam Funk, and Wim Peters. 2009. Using lexico-syntactic ontology design patterns for ontology creation and population. In *Proc. of the Workshop on Ontology Patterns*.

Christopher Parisien and Suzanne Stevenson. 2011. Generalizing between form and meaning using learned verb classes. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*. Citeseer.

Rebecca J Passonneau and Bob Carpenter. 2013. The benefits of a model of annotation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 187–195.

Steven Pinker. 1984. *Language Learnability and Language Development*. Harvard University Press.

Steven Pinker. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2012. The phrase detective multilingual corpus, release 0.1. In *Collaborative Resource Development and Delivery Workshop Programme*, page 34.

Andrey Rzhetsky, Hagit Shatkay, and W John Wilbur. 2009. How to get the most out of your curation effort. *PLoS Computational Biology*, 5(5):1?13.

Robert S Swier and Suzanne Stevenson. 2004. Unsupervised semantic role labeling. In *Proceedings of the Generative Lexicon Conference*, volume 95, page 102.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier R Movellan. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, volume 22, pages 2035–2043.

Annie Zaenen, Daniel G Bobrow, and Cleo Condoravdi. 2008. The encoding of lexical implications in verbnet: Predicates of change of locations. In *Language Resources Evaluation Conference*.