

# Comprehensive and Consistent PropBank Light Verb Annotation

Claire Bonial,<sup>1</sup> Martha Palmer<sup>2</sup>

<sup>1</sup>U.S. Army Research Laboratory, <sup>2</sup>University of Colorado, Boulder

<sup>1</sup>2800 Powder Mill Rd., Adelphi, MD, USA

E-mail: Claire.N.Bonial.civ@mail.mil, Martha.Palmer@colorado.edu

## Abstract

Recent efforts have focused on expanding the annotation coverage of PropBank from verb relations to adjective and noun relations, as well as light verb constructions (e.g., *make an offer*, *take a bath*). While each new relation type has presented unique annotation challenges, ensuring consistent and comprehensive annotation of light verb constructions has proved particularly challenging, given that light verb constructions are semi-productive, difficult to define, and there are often borderline cases. This research describes the iterative process of developing PropBank annotation guidelines for light verb constructions, the current guidelines, and a comparison to related resources.

**Keywords:** light verb constructions, semantic role labeling, NLP resources

## 1. Introduction

The goal of PropBank (Palmer et al., 2005) is to supply consistent, general-purpose labeling of semantic roles across different syntactic realizations. With over two million words from diverse genres, the benchmark annotated corpus supports the training of automatic semantic role labelers, which in turn support other Natural Language Processing (NLP) areas, such as machine translation. PropBank annotation consists of two tasks: sense and role annotation. Previously, PropBank annotation was focused on verbs (a separate, but related project, NomBank (Meyers et al., 2004), focused on noun relations), but recent efforts have shifted to expanding the annotation coverage of PropBank from verb relations to adjective and noun relations, as well as light verb constructions (LVCs; e.g., *have a nap*, *do an investigation*, *give a kiss*) (Bonial et al., 2014). This shift has allowed PropBank to capture more comprehensively all mentions of events of states, which can be realized alternately as verbs (*fear*), nouns, (*fear*), adjectives (*afraid*) or LVCs (*have fear*). Although each new relation type has presented unique annotation challenges, ensuring consistent and comprehensive annotation of LVCs has proved particularly challenging because LVCs are semi-productive, difficult to define objectively, and there are often borderline cases. This research describes the iterative process of developing PropBank annotation guidelines for LVCs (for a description of the early development stages of LVC annotation, see Hwang et al., 2010), the current guidelines,<sup>1</sup> and a comparison to the annotation practices of two related resources: the TectoGrammatical Treebank (Cinková et al., 2004) and Tu and Roth’s (2011) LVC dataset that was developed specifically for enabling automatic LVC detection.

## 2. PropBank Background

PropBank annotation consists of two tasks: sense annotation and role annotation. The PropBank lexicon provides a listing of the coarse-grained senses of a verb, noun, or adjective relation, and the set of roles associated with each sense (thus, called a “roleset”). The roles are listed as argument numbers (Arg0 – Arg6) and correspond to verb-specific roles. For example:

**Offer-01** (*transaction, proposal*):

Arg0: *entity offering*

Arg1: *commodity*

Arg2: *price*

Arg3: *benefactive, or entity offered to*

Previous versions of PropBank maintained separate rolesets for the verb *offer*, and the related nouns *offer/offering*, but these have been combined (Bonial et al., 2014).<sup>2</sup> Thus, a roleset now clearly maps and provides parallel annotations of related usages:

1. [NE Electric]<sub>ARG0</sub> **offered** [\$2 billion]<sub>ARG2</sub> [to acquire PS of New Hampshire]<sub>ARG1</sub>
2. [NE Electric]<sub>ARG0</sub> **made an offer** [of \$2 billion]<sub>ARG2</sub> [to acquire PS of New Hampshire]<sub>ARG1</sub>

PropBank 1.0 had no special guidelines for LVCs. Therefore, example (2) would have been annotated according to the semantic roles laid out in the verb roleset for *make-01*, which denotes creation events and is, at best, metaphorically related. This practice created equivalent semantic representations for both *make an offer* and *make a cake*, which ultimately have very different semantic features and entailments. The special semantic and syntactic structure of LVCs requires a unique annotation procedure, without which their semantics may not be recognized in NLP systems.

<sup>1</sup> <http://proppbank.github.io/>

<sup>2</sup> <https://github.com/proppbank/proppbank-frames>

### 3. LVC Background

Definitions of LVCs cross-linguistically and within a language remain nebulous, with multiple labels applied to similar phenomena and arguably distinct phenomena labeled as ‘light’ verbs (Jespersen, 1942). Nonetheless, LVCs in English are largely thought to consist of a semantically general, highly polysemous verb and a noun denoting an event or state (e.g., Butt, 2003). More detailed aspects of LVCs remain debatable; thus, different LVC resources encompass different scopes of constructions (see Section 5).

LVCs are semi-productive, meaning that some novel combinations of light verbs (LVs) and eventive/stative nouns arise (Stevenson et al., 2004), but other combinations are unacceptable. The seemingly idiosyncratic nature of LVC productivity contributes to the degree of difficulty in automatically detecting LVCs. Furthermore, surface-identical forms can be LVCs or heavy usages of the same verbs:

3. He **took a drink** of the soda. (LVC)
4. He **took a drink** off the bar. (non-LVC)

Such usages are syntactically indistinguishable (Butt & Geuder, 2001); thus, neither automatic systems nor annotators can rely on syntactic criteria for detection. Manual annotation must use semantic criteria instead, outlined in the guidelines and described in the next section.

### 4. PropBank LVC Annotation Guidelines

Like other resources that seek to distinguish LVCs (e.g., Tu & Roth, 2011), initial guidelines relied on the basic heuristic that the LVC candidate can be rephrased using a related lexical verb without any significant loss in meaning, as seen in the swap of *offer* and *make an offer* in examples (1) and (2). Although intuitively appealing, reliance on this heuristic provides an overly narrowly view of LVCs, since there are two types of LVCs that are difficult to rephrase. First, some LVCs in English have no etymologically related verbal counterpart (e.g., *commit suicide*). Secondly, some LVC counterparts of verbs affect valency in a manner similar to the passive voice. For example,

5. We **got a call** from the hospital.<sup>3</sup>

This can only be rephrased with the related verb *call* in the passive:

6. We **were called** by the hospital.

The active rephrasing gives the wrong meaning:

7. We **called** from the hospital.

<sup>3</sup> The borderline status of this LVC is discussed in Section 4.3.

This makes the rephrasing test rather complex for annotators to use consistently to determine what is “similar enough” to pass. As a result, under the first version of LVC guidelines, PropBank was not recognizing LVCs that couldn’t be straightforwardly rephrased as verbs.

Such gaps in annotation became clear when pairs of similar LVCs with different LVs were compared and found to have very different representations in PropBank. For example, *take a bath* was annotated as an LVC with roles from *bathe-01*, while *get a bath* was annotated according to the transfer semantics of *get-01*, alongside usages like *get a package*. To amend these gaps and inconsistencies, the guidelines were expanded and improved, as described in the next section.

### 4.1 Roleset Criteria

In order to ensure that ongoing and future annotation is as comprehensive and consistent as possible, the guidelines were updated to focus on the semantic nature of arguments. This also helped to ground the guidelines in theoretical research on English LVCs, which generally assumes that the semantic content of the construction stems from the noun, while the verb provides the syntactic scaffolding for the noun to act as the main predicate (Butt, 2003; Grimshaw & Mester, 1988). Accordingly, the arguments, including the syntactic subject of the verb, will carry the semantic roles of the noun. The annotators are instructed to compare the fit of the semantic roles listed for the verb’s roleset to that of the noun’s roleset. Figure 1 gives an overview of the heuristics annotators use to decide if a *Verb + Noun* candidate is an LVC.

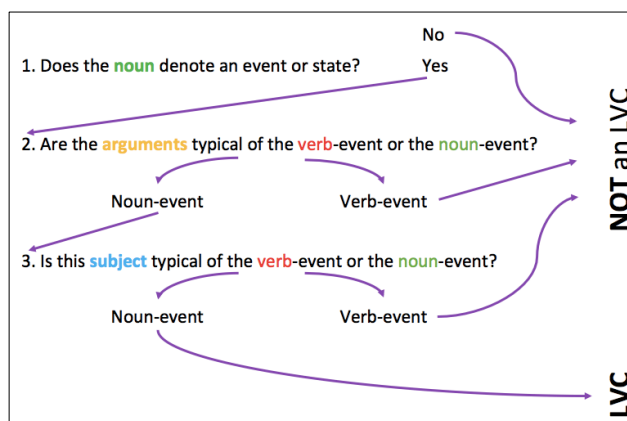


Figure 1: Flow chart for determining LVC status.

Consider the earlier sentence from examples (1) and (2) with the LVC *make\_offer* (repeated here for the reader’s convenience):

8. New England Electric **made an offer** of \$2 billion to acquire PS of New Hampshire.

Annotators would consider the roles laid out in *make-01* or *offer-01*, and decide which were more fitting of the

usage:

**Make-01** (*create*):

Arg0: *creator*

Arg1: *creation*

Arg2: *created-from, thing changed*

Arg3: *benefactive*

Annotation using the Make-01 roles:

9. [New England Electric]<sub>ARG0</sub> **made** [an offer of \$2 billion]<sub>ARG1</sub> [to acquire PS of New Hampshire]<sub>ARG-???</sub>

**Offer-01** (*transaction, proposal*):

Arg0: *entity offering*

Arg1: *commodity*

Arg2: *price*

Arg3: *benefactive, or entity offered to*

Annotation using the Offer-01 roles:

10. [New England Electric]<sub>ARG0</sub> **made an offer** [of \$2 billion]<sub>ARG2</sub> [to acquire PS of New Hampshire]<sub>ARG1</sub>

The verb annotation in (9) does not capture any of the semantics of the *offer* event, instead focusing on *make* as if it were a heavy, semantically rich verb. As a result, the argument “to acquire PS of New Hampshire” lacks an appropriate tag in the roleset. Furthermore, given the syntactic span of annotation that PropBank annotators attend to for verb relations, the *price* argument is lumped into the object noun phrase: “an offer of \$2 billion.” On the other hand, the noun roleset used in (10) provides the appropriate semantic roles to label each argument. Additionally, the LVC span of annotation labels syntactic arguments of the noun relation in addition to the verb relation, so the Arg2 *price* can be captured separately as a syntactic argument of the noun relation. Quite clearly in this case, the semantic roles stem from the noun relation, and the usage should be annotated as an LVC.

The resulting annotation as an LVC reflects the semantic contribution of the noun by using the noun’s roleset, but also captures the syntactic role of the verb, as syntactic arguments of both the light verb and noun are annotated. Although the light verb is initially marked as “LVB” (Light VerB) in annotation, post-processing completes a final step wherein the light verb and noun are combined into a single, complex relation (“REL”) *make + offer*:

11. [New England Electric]<sub>ARG0</sub> **made-REL an offer-REL** [of \$2 billion]<sub>ARG2</sub> [to acquire PS of New Hampshire]<sub>ARG1</sub>

## 4.2 Challenging Cases

The question of whether or not the subject argument specifically carries a role assigned by the noun relation may seem redundant, but it is of special importance. In some borderline cases, non-subject arguments carry the

semantic roles of the noun relation, but the subject argument introduces a role that is not intrinsic to the nominal event, instead often introducing an outside cause in the case of *give*, or a peripherally affected entity in the case of *have*. For example:

12. We’ve **had assassinations** before this.

‘We’ does not denote the person assassinated or the killer (noun roles), rather a group peripherally affected by the assassinations.

13. It **gives** you a **sense** of the terror she felt.

‘It’ does not denote the sensor or thing sensed (noun roles), rather an outside cause of sensation.

The above are not prototypical heavy usages of the verbs, but they are also not light verbs in the strictest sense, since the verb assigns a role to the subject argument, evidencing its semantic weight.

Perhaps the most difficult cases are what can be called Semi-Light Verb Constructions (semi-LVCs), truly borderline cases where the semantics of the verb and noun overlap so much that it is impossible to determine which is projecting the semantic roles. *Give* is often semi light, as is *get*. Consider the subcategorization frame of the transfer sense of *give*:

14. I gave her a book - NP1<sub>[AGENT]</sub> V NP2<sub>[RECIPIENT]</sub> NP3<sub>[THEME]</sub><sup>4</sup>

If *give* is an LV, then the semantic roles assigned to arguments stem instead from the noun relation:

15. I gave them my blessing - NP1<sub>[Agent]</sub> LV NP2<sub>[PATIENT OF BLESS]</sub> NP3<sub>[AGENT OF BLESS, NOUN PREDICATE]</sub>

In (15), *them* is more accurately interpreted as the *thing blessed*, as opposed to the Recipient of an abstract item, *blessing*. This is not to say that the two interpretations are not related, or that the latter interpretation is “wrong,” but for the purposes of Natural Language Understanding, recognizing a *blessing* event provides much more semantic specificity and allows for more appropriate inferences than interpreting this as a *transfer* event.

If *give* is used in a semi-LVC, both *give* and the predicating noun share the same semantic roles; thus, it is unclear where the semantics stem from:

16. I gave them a hint. - NP1<sub>[AGENT]</sub> LV NP2<sub>[RECIPIENT]</sub> NP3<sub>[NOUN PREDICATE/ THEME]</sub>

In this case, *them* could be understood as the Recipient of either *give* or *hint* (i.e. *give to them* or *hint to them*). Similarly, the Agent subject could be assigned by either

<sup>4</sup> NP=Noun Phrase argument, V=Verb, LV=Light Verb

relation. Unlike the true LVC in (15), the indirect object cannot be understood as the Patient or Theme of *hint* — *them* is not the *thing hinted*. Similarly, in example 5, *We got a call from the hospital*, the roles of *get* and *call* overlap, as both essentially involve an Agent, Theme and Recipient. Initially, semi-LVCs were not marked as LVCs in PropBank, but this again led to troubling inconsistencies, such as *give a speech* being marked as a transfer event, and *make a speech* being annotated as an LVC speech event. Thus, a greater set of potential LVCs, including all instances of *give* and *get*, were pulled for re-annotation under improved guidelines. These improvements clarified PropBank’s position that such borderline, semi-LVC cases are best annotated as LVCs, thereby providing the most consistent representation of related LVCs that differ only in the light verb.

### 4.3 Evaluation

The success of these guidelines has been demonstrated in the high agreement rates between annotators. On a task composed solely of the most likely LVs (*give, have, take, make, do*), agreement rates between annotators was 93.8%.<sup>5</sup> As proof of the quality of PropBank’s annotations, an LVC detection system has been trained on the OntoNotes 4.99 corpus (Weischedel et al., 2011), of which PropBank is one layer. This system achieves an F-Score of 89% (Chen et al., 2015) when tested on the same British National Corpus (BNC) data used by Tu and Roth (2011), who achieve an F-Score of 86.3%. When tested on the more realistic and challenging OntoNotes corpus containing 1,768 LVC instances, the system achieves an F-Score of 80.7%.

## 5. Comparison to Other Resources

Training data annotated according to different definitions of LVCs will result in distinct successes and failures of automatic identification. Thus, it is important to consider and compare how LVCs are defined across other notable resources with LVC annotation: the TectoGrammatical Treebank (TTree) (Cinková et al., 2004) and the British National Corpus (BNC) data set developed by Tu and Roth (2011).

The TTree annotation guidelines require a specific treatment for “support verb constructions,” which are described as “multi-word predicates consisting of a semantically empty verb which expresses the grammatical meanings in the sentence and a noun (frequently denoting an event or a state), which carries the main lexical meaning of the entire predicate” (Cinková et al., 2004: 91). This is similar to how PropBank defines LVCs. The guidelines proceed with a description of the rephrasing test, which indicates that most support verb constructions can be rephrased with a one-word predicate. Under the TTree guidelines, it is clear that the shared arguments of support verb constructions can be licensed by either the verb or the

noun predicate. This is quite distinct from PropBank LVC guidelines, which define LVCs more narrowly as constructions wherein the verb does not assign semantic roles to the arguments of the construction, but merely syntactically licenses arguments, including the subject argument.

This difference in views results in a very different data set of TTree Support Verb Constructions (SVCs), encompassing a much wider variety of support verbs beyond LVCs. A total of 694 support verb cases from TTree were examined for this research, and refined into five categories of support verb subtypes, including LVCs as defined by PropBank. A sixth category can perhaps be more properly thought of as idiomatic expressions. The findings of this comparison are summarized Table 1.

SVC Type	TTree Example	Semantic Contribution of Verb
LVC	<b>commit</b> crime, <b>have</b> right	None
Semi-LVC	<b>give</b> hint	Undeterminable
Aspectual	<b>take</b> flight, <b>hold</b> <b>out</b> hope	Initiation, continuation, termination
Copular	<b>be</b> right/wrong	None
Outside Cause	<b>grant</b> them the right to...	Causation
Idiomatic	<b>give</b> the cold shoulder	Non-compositional meaning

Table 1: Categories of SVCs in TTree annotations.

Overall, the TTree support verb annotations include a much wider variety of support verbs, with varying levels of semantics that are attributable to the verb. Nonetheless, TTree benefits from a very consistent treatment of all cases where verbs syntactically support potential arguments of a predicating noun (or adjective). Although PropBank already treats copular verbs with predicate adjectives in a distinct, but parallel fashion to LVC annotation, aspectual and causal support verbs are assigned heavy senses denoting their aspectual or causal semantics. Although this representation isn’t inaccurate, it should be explored as to whether a modified treatment could more clearly relate all of these cases as support verb constructions.

Tu and Roth’s (2011) dataset has been used to establish a state-of-the-art LVC detection system. Tu and Roth manually construct a dataset of 2,162 English sentences with positive and negative (semantically full or heavy usages of common LVs) LVC instances from the BNC. The short guidelines for identifying LVCs rely solely on the rephrasing test.<sup>6</sup> The authors restrict their data set to LVCs involving the six most frequent LVs: *do, get, give, have, make, take*. In PropBank, any verb can be marked as light, resulting in identification of LVCs like *bring charges against* and *conduct repairs*. Tu and Roth also

<sup>5</sup> <http://verbs.colorado.edu/propbank/ita/webtext-p25-SEL-LightVerb.html>

<sup>6</sup> <http://cogcomp.cs.illinois.edu/~ytu/test/LVCmain.html>

filter their data set by including only LVCs with nouns that are zero-derived nominals (e.g., *offer*), or derivationally related to a verb (e.g. *destruction*). PropBank includes LVCs with nouns that have no related verb, such as *take a trip*. The broader range of LVCs found in PropBank and OntoNotes does make detection more challenging, as evidenced in the lower Chen et al. (2015) classifier score (80.7%) on this test set as compared to the BNC test set (89%). Nonetheless, PropBank provides a more complete view and representation of the full inventory of LVCs in English.

## 6. Conclusion & Future Work

A lack of consensus regarding the definition of LVCs in the linguistic and NLP community has hindered progress in the detection and interpretation of LVCs. This research presents the PropBank guidelines for LVC annotation that is grounded in linguistic theory and practical for NLP. Compared to other resources, PropBank espouses a narrower view of LVCs than TTree, but a more inclusive view than Tu and Roth's dataset. Raising awareness of the differences between these resources and their unique definitions of LVCs gives the community the opportunity to arrive at a consensus concerning the delimitation of these constructions. PropBank treatments of both support verbs and light verbs have been developed with careful consideration of what distinctions are maximally informative that can be made consistently by annotators. Once detection is enabled, a system must also interpret LVCs correctly. Thus, future work includes an exploration of using mappings from PropBank to VerbNet (Kipper et al., 2008) to access semantic representations of the verb counterparts of LVCs, and porting this over as a semantic representation of the LVC. This relies upon the assumption that the semantics of the LVC can be adequately captured by essentially replacing the LVC with a verb related to the noun predicate. (Notably, replacing the LVC with a related verb is precisely how the Abstract Meaning Representation project (Banarescu et al., 2012) represents LVCs.) Although this approach may be adequate for the interpretation of LVCs in English, it must be evaluated for LVCs in other languages, since there is a cross-linguistic tendency for LVCs to cover the semantic space that is not covered by lexical verbs. Certainly further exploration into using NLP resources for the correct interpretation of LVCs is needed. Here, we have taken the preliminary steps for consistently and comprehensively defining and annotating LVCs, allowing for a focus on interpretation in the future.

## 7. Acknowledgements

We gratefully acknowledge the support of the National Science Foundation Grants NSF: 0910992 IIS:RI: Large: Collaborative Research: Richer Representations for Machine Translation and IIS-1116782: A Bayesian Approach to Dynamic Lexical Resources for Flexible

Language Processing, and the support of DARPA BOLT - HR0011-11-C-0145 and DEFT - FA-8750-13-2-0045 via a subcontract from LDC. 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, DARPA or the US government.

## 8. Bibliographical References

- Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. and Schneider, N., (2012). Abstract meaning representation (AMR) 1.0 specification. In *Parsing on Freebase from Question-Answer Pairs.* In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle: ACL (pp. 1533-1544).
- Bonial, Claire, et al. (2014). PropBank: semantics of new predicate types. *The 9th edition of the Language Resources and Evaluation Conference*. Reykjavik, Iceland.
- Butt, Miriam. (2003). The Light Verb Jungle. In G. Aygen, C. Bowern & C. Quinn (Eds.) *Papers from the GSAS/Dudley House Workshop on Light Verbs*. Cambridge, Harvard Working Papers in Linguistics, p. 1-50.
- Butt, Miriam, and Wilhelm Geuder. (2001). On the (Semi)Lexical Status of Light Verbs. In *Semi-Lexical Categories: On the content of function words and the function of content words*, ed. Norbert Corver and Henk van Riemsdijk. 323-370. Berlin: Mouton De Gruyter.
- Chen, Wei-Te, Claire Bonial & Martha Palmer. (2015). English Light Verb Construction Identification Using Lexical Knowledge. In proc. of the *Twenty-Ninth Conference on Artificial Intelligence (AAAI-15)*.
- Cinková et al. (2004). *Annotation of English on the tectogrammatical level*. Technical Report TR-2006-35, UFAL/CKL, Prague, Czech Republic.
- Grimshaw, J., and A. Mester. (1988). Light verbs and  $\theta$ -marking. *Linguistic Inquiry*, 19(2):205-232.
- Hwang, J.D., Bhatia, A., Bonial, C., Mansouri, A., Vaidya, A., Xue, N. and Palmer, M. (2010). Propbank annotation of multilingual light verb constructions. In *Proceedings of the Fourth Linguistic Annotation Workshop*(pp. 82-90). Association for Computational Linguistics.
- Jespersen, Otto. (1942). A Modern English Grammar. Part VI, Morphology. *Ejnar Munksgaard, Copenhagen*.
- Kipper, Karin, et al. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation* 42.1: 21-40.
- Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., & Grishman, R. (2004). Annotating Noun Argument Structure for NomBank. In *LREC* (Vol. 4, pp. 803-806).
- Palmer, Martha, Daniel Gildea, and Paul Kingsbury. (2005). The Proposition Bank: An annotated corpus of

semantic roles. *Computational linguistics* 31.1: 71-106.

Stevenson, Suzanne, Afsaneh Fazly, and Ryan North. (2004). Statistical measures of the semi-productivity of light verb constructions. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*. Association for Computational Linguistics.

Tu, Yuancheng, and Dan Roth. (2011). Learning English Light Verb Constructions: Contextual or Statistical. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*, pp. 31-39.

Weischedel, R., E. Hovy, M. Palmer, M. Marcus, R. Belvin, S. Pradhan, L. Ramshaw and N. Xue. (2011). Ontonotes: A large training corpus for enhanced processing. In Olive, J.; Christianson, C.; and McCary, J., eds., *Handbook of Natural Language Processing and Machine Translation*. Springer.