Expanding VerbNet with Sketch Engine

Claire Bonial, Orin Hargraves & Martha Palmer
Department of Linguistics,
University of Colorado at Boulder
Hellems 290, 295 UCB
Boulder, CO 80309-0295
{Claire.Bonial, Orin.Hargraves, Martha.Palmer}@colorado.edu

Abstract

This research describes efforts to expand the lexical resource VerbNet with additional class members and completely new verb classes. Several approaches to this in the past have involved automatic methods for expansion, but this research focuses on the addition of frequent, yet particularly challenging verbs that require manual additions after a survey of each verb's syntactic behaviors and semantic features. Sketch Engine has been an invaluable tool in this process, allowing for a comprehensive, yet detailed view of the behavior of a given verb, along with efficient comparisons to the behaviors of other verbs that might be included in VerbNet already. The incorporation of light verbs into VerbNet has presented particular challenges to this process, these are described along with a proposed resource to supplement VerbNet with information on light verbs.

1 Introduction

VerbNet (VN) (Kipper et al., 2008) is a classification of English verbs, expanded from Levin's (1993) classification. VN serves as a valuable lexical resource, facilitating a variety of Natural Language Processing (NLP) tasks such as semantic role labeling (Swier and Stevenson, 2004), inferencing (Zaenen et al., 2008), and automatic verb classification (Joanis et al., 2008). VN currently contains entries for about 6300 verbs, with continuous efforts to expand VN's coverage. VN is one resource included in SemLink (Palmer, 2009; Loper et al., 2007), which is both a mapping resource, unifying a variety of complementary lexical resources, and an annotated corpus. Through its unification of resources, SemLink provides an efficient way in which to compare resources and understand their strengths in weaknesses, including deficiencies in coverage. In an investigation of the coverage of VN for verbs found in the SemLink corpus, which consists of 112,917 instances of the Wall Street Journal, approximately 20 verbs were discovered with relatively high frequencies that were not accounted for in VN. These instances make up 14,878, or 78%, of the 19,070 SemLink instances missing VerbNet classes. These verbs include, for example, *account, be, benefit, cite, do, finance, let, market, tend, trigger,* and *violate*. Thus, while past efforts to expand VN have used automatic methods (Korhonen and Briscoe, 2004) primarily grouping verbs by syntactic patterns, these efforts take these highly frequent verbs as a starting point, as their addition to VN would greatly expand its coverage and completeness. The drawback of this approach is that many of these verbs were not already included in VN precisely because they are quite unique in their syntax and semantics, thus making them difficult candidates for incorporation into VN's class structure, which is described in more detail in the sections to follow.

Sketch Engine's (Kilgarriff et al., 2004) Word Sketch and Thesaurus functions were found to be extremely helpful in the process of considering these verbs for addition, because these resources give a detailed snapshot of syntactic and collocational tendencies. Particularly difficult cases for addition are those where common, polysemous verbs are used in their 'light' sense while combining with another predicating element; for example, *Jessica **made an offer** to buy the house*. These cases are especially problematic for VN to account for because the structure of the lexicon assumes that the verb is the primary predicating element. The steps and challenges of these additions are discussed in turn. The overall successes of this expansion demonstrate the value of utilizing both the complementary lexical resources included in SemLink, as well as Sketch Engine.

2 Background

VN and Sketch Engine are two lexical resources that provide a wealth of information on the syntactic behaviors of certain lexical items. In the case of VN, these behaviors are expressed primarily through syntactic frames and alternations common to verb class members, listed in each class. The syntactic information of VN draws heavily from Levin's (1993) work, which documented the syntactic behavior of verbs as reflected in a survey of primarily literary sources. In the case of Sketch Engine, syntactic and collocational information is drawn algorithmically from very large corpora. Thus, the two resources are quite complementary because VN makes theoretically-grounded useful generalizations about the behaviors of classes of verbs, while Sketch Engine provides empirically-based statistical information about the behavior of verbs. SemLink is also instrumental in this process because the annotated corpus can reveal which verbs should be prioritized for addition to VN. Each of these resources is discussed in more detail in the next sections.

2.1 VerbNet Background

Class membership in VN is based on a verb's compatibility with certain syntactic frames and alternations. For example, all of the verbs in the Spray class, which includes the verb *load*, have the ability to alternate the Theme or Destination as a noun phrase (NP) object or as a prepositional phrase (PP): *Jessica loaded the boxes into the wagon*, or *Jessica loaded the wagon with boxes*. VN's structure is somewhat hierarchical, comprised of superordinate and subordinate levels within each verb class. In the top level of each class, syntactic frames that are compatible with all verbs in the class are listed. In the lower levels, or 'sub-classes,' additional syntactic frames may be listed that are restricted to a limited number of members. In each class and sub-class, an effort is made to list all syntactic frames in which the verbs of that class can be grammatically realized. Each syntactic frame is detailed with the expected syntactic phrase type of each argument, thematic roles of arguments, and a semantic representation; for example:

Frame NP V NP PP.destination
Example Jessica loaded boxes into the wagon.
Syntax Agent V Theme Destination
Semantics Motion(during(E), Theme)
Not(Prep-*into*(start(E), Theme, Destination))
Prep-*into*(end(E), Theme, Destination)
Cause(Agent, E)

The class numbers in VN also reflect larger groups of what can be thought of as meta-classes. Thus, for example, all classes beginning with the number 9 (9.1-9.10) are verbs of placement. Although this classification is primarily based on shared syntactic behaviors, there is clear semantic cohesion to each of the classes. As Levin hypothesizes, this is a result of the fact that verb behavior is determined by verb meaning.

The syntactic information of VN is intended to be comprehensive in the sense that it includes all grammatical realizations of core, or frequent arguments, including some that can be optional. As a result, it can be quite difficult to add class members and classes to VN. To add a member, the verb must firstly be compatible with the primary diathesis alternation characterizing that class, and it must be compatible with all other syntactic frames listed in its class (or subclass). To add a class, two or more verbs that share a diathesis alternation and other syntactic behaviors must be discovered. In many cases, finding existing classes that are compatible with a candidate for addition is not possible, and determining what verbs warrant a new class is also a difficult question. Sketch Engine is in many ways an ideal supplement to this process because its Word Sketch function provides detailed information on the syntactic behaviors of a verb, and the Thesaurus tool can offer verbs that are used very similarly that may be candidates for new classes.

2.2 Sketch Engine Background

Sketch Engine is a corpus query and processing system for the automatic extraction of lexical information (Kilgarriff et al., 2004). Used in conjunction with a large corpus, it can generate data that efficiently summarizes the behavior of any word representing a major part of speech (noun, verb, adjective, adverb). Sketch Engine was developed for the use of lexicographers compiling dictionaries but has found widespread use in NLP because of its sophisticated and varied corpus query tools.

The two Sketch Engine tools most pertinent to our inquiry are the Word Sketch and the The-

saurus function. A Word Sketch is an HTML-formatted listing of a keyword's functional distribution and collocation in a corpus. This information includes syntactic information, such as which parts of speech and lexical items frequently act as complements of verbs. This is very useful for considering VN class membership, as membership is based on compatibility with certain syntactic frames. The Thesaurus function in Sketch Engine provides a list of words with the same part of speech for a given word that are assigned a score above a certain threshold. The score is based on the number of triples that two words share across a corpus. The higher the score, the more similar the behavior of the two words, and thus the more likely they are to be synonyms for computational purposes. This function is also useful when considering VN membership, because similar words will often share classes.

2.3 SemLink Background

SemLink (Palmer, 2009; Loper et al., 2007) is both a mapping resource and an annotated corpus. It provides mappings between complementary lexical resources: PropBank (Palmer et al., 2005), VN, FrameNet (Fillmore et al., 2002), and the recently added OntoNotes sense groupings (Pradhan et al., 2007). Each of these lexical resources varies in the level and nature of semantic detail represented, since each was created independently with somewhat differing goals. Nonetheless, all of these resources can be used to associate semantic information with the propositions of natural language. SemLink serves as a platform to unify these resources and therefore combine the fine-granularity and rich semantics of FrameNet, the syntactically-based generalizations of VN, and the relatively coarse-grained semantics of PropBank, which has been shown to be effective training data for supervised Machine Learning techniques. The recent addition of the OntoNotes sense groupings, which can be thought of as a more semantically general, or coarse-grained, view of WordNet (Fellbaum, 1998), provides even broader coverage for the resource.

The SemLink annotated corpus consists of approximately 112,000 instances of the Wall Street Journal, wherein ideally each verb is annotated with its VN class, PropBank 'roleset,' (i.e. coarse-grained sense), FrameNet frame, and OntoNotes sense number. Each argument of the verb is labelled with its VN theta role, PropBank argument number and FrameNet frame element label. The current version of SemLink includes about 78,000 instances with complete annotation; yet there are about 19,000 instances with PropBank annotations but no VN annotations because the verb is simply not present in VN. PropBank is the most comprehensive resource because, unlike FrameNet and VN, the primary goal in developing PropBank was not lexical resource creation, but the development of an annotated corpus to be used as training data for supervised machine learning systems. PropBank, like FrameNet, also includes relations other than verb relations, with annotations for noun, adjective, and complex light verb construction predicates (see http://verbs.colorado.edu/propbank/EPB-Annotation-Guidelines.pdf for full annotation guidelines, see Hwang et al., 2010-a for a description of the annotation of light verbs). As mentioned previously, verbs that are present in PropBank, and therefore SemLink, but not present in VN are prime candidates for addition.

3 Challenges of Adding VerbNet Members

The motivation for the expansion of VN is to make it a more robust tool for use in NLP by increasing its coverage. Pursuant to that, we work from a list of verbs that are relatively frequent in SemLink. In some cases, intuitive or lexicographic examination of a verb is sufficient for locating its destination in VN. When a verb has the same syntactic behavior as its super-type, for example, and the super-type is already in VN, it's possible that a new verb can simply be added to the same class its super-type is in. The relatively infrequent verb *abominate* is a synonym/subtype of *hate* and instantiates syntactic patterns similar to those of *hate*. It can be added to VN in the Admire class, where *hate* is already present.

A more complicated scenario arises when a verb shares some but not all syntactic or semantic properties of a synonym or super-type verb already in VN. In these cases, it is helpful to consult the Thesaurus function of Sketch Engine to see what verbs share the greatest number of patterns with a candidate for addition to VN. If any of these verbs is already in VN, its class can be examined for suitability with regard to the new verb. As a case in point: the transitive verb *authenticate* is not cur-

rently in VN. A thesaurus query in Sketch Engine shows *authenticate* to be syntactically and semantically similar to (in descending order) *substantiate, verify, validate, falsify,* and *corroborate*. Of these verbs, two are in the VN Indicate class (*verify* and *corroborate*), and in fact, *authenticate* fits well in the Indicate class as well.

Sketch Engine is less successful at predicting the appropriate target class of a candidate for addition to VN in three general cases:

1. when there is a sparsity of data for the candidate verb in the corpus
2. when the candidate verb's behavior does not closely match any class existing in VN
3. when the candidate verb has strong semantic ties or syntactic ties with verbs in more than one VN class but doesn't exactly fit in any of them.

In the first case there is little to be gained from examining Sketch Engine data. In cases of data sparsity, Sketch Engine may show words that are not even the same part of speech as the queried word. A query on the verb *dissimulate*, for example, returns only the adjective *glum* in Sketch Engine, with an extremely low similarity score. In the latter two cases above, examination of Sketch Engine data is still useful because it may point out possible weaknesses in VN: it may indicate a need to subdivide or reanalyze a current class, or to create a new class.

3.1 Case Studies: Successful Additions

The highly frequent verb *discuss* has recently been added to VN, in the Chit_Chat class. Information from the Thesaurus function in Sketch Engine was instrumental in helping us to arrive at the correct placement for *discuss*, which involved a minor reanalysis of the Chit_Chat class.

Most of the verbs sharing significant patterns with *discuss* as reported in Sketch Engine are already located in either of two broad classes in VN. There is *explain, mention, suggest,* and *note* (all located in the Chit_Chat class), and *consider, describe, accept,* and *believe* (all located in the larger group of classes beginning with 29, including the Characterize, Consider and Conjecture classes). Examination of the subclasses in these two broad classes did not turn up an exact match for *discuss* that allowed for instantiation of all frames, but we found that the formerly undivided Chit_Chat class could easily be split into two sibling classes that would enable us to find a perfect fit for *discuss* (which is now in 37.6-2, a subclass of Chit_Chat). It also resulted in a more rational organization for the class overall, with verbs in each of the two sibling classes fully functional in all the frames listed.

Here it is interesting to note that Levin's work, largely theoretical, insight-based, and undertaken before the availability of examining verb behavior in large corpora, is largely supported by empirical data, based entirely on the behavior of words computed statistically. Verbs that Levin had classified as near neighbors and that occupy adjacent classes in VN are demonstrably similar in their behavior as shown by the distributional analysis delivered by Sketch Engine.

3.2 Case Studies: Difficult Additions

A case where Sketch Engine fails to deliver information that facilitates the placement of a verb in VN can be illustrated with the rather complex and frequent verb *cite*. In PropBank, *cite* is represented by two senses or numbered 'rolesets': the far more frequent cite.01, which covers uses such as 'cite an example/source/case/reason' and 'Weed control is cited as the single most important challenge in organic farming,' and the less frequent cite.02, which has only the single pattern 'cite (a person) for (a violation).'

The statistical analysis delivered by Sketch Engine for *cite* draws far more from cite.01 than from cite.02 and offers verbs with high similarity scores that in VN are located mainly in classes beginning with the number 37, which are verbs of reporting: *mention, note, acknowledge, discuss, claim, explain,* and *state*. Despite these many similarities, there is not a class or subclass of 37.* that accounts well for the behavior of *cite*, mainly because it has more selectional restrictions than many verbs in those classes. *Cite*, for example, is not typically followed by a relative clause, which is characteristic of reporting verbs.

The 17th verb in terms of similarity scores provided by the Sketch Engine thesaurus for *cite* is *criticize*. This verb is in VN's Judgment class, and this seems to be a recognition of the less frequent use of *cite*, that is, cite.02 from PropBank, 'cite the witness for contempt.' The statistical algorithm for generating word similarities is surely the explanation for this much lower similarity score, because of the relative infrequency of this meaning of *cite*. Nonetheless, *cite* has been added to the

Judgment class, after reorganization and the addition of a subclass that allowed for the class to not only accommodate this verb, but also more precisely the capture the behaviors of all verbs in the class.

4 Adding Classes

When Sketch Engine shows no easily interpretable pattern for the placement of a verb in VN, and the verb is frequent, with many reportable patterns, it provides an occasion to examine whether VN is deficient in having no established class that captures the syntax and semantics of such a verb. A case we recently examined is the verb *benefit*.

4.1 Benefit Class

Benefit is reported in Sketch Engine to share significant patterns with several verbs: *gain, encourage, enable, help, attract,* and *suffer*, for example. We used the Word Sketch function in Sketch Engine for an analysis of the patterns exemplified by *benefit*, and it indicates that *benefit* has an important diathesis alternation that is not possible for any of these verbs. We can say, for example,

4. The program benefits minorities.
5. Minorities benefit from the program.
6. Minorities benefit.

and get approximately the same meaning. Like ergative English verbs, there is a strong overlap between the most frequent subjects and objects of the verb *benefit*. In one corpus we examined, for example, the five nouns *people, community, patient, child,* and *student* were the most frequent as both the subjects and the objects of *benefit*. None of *benefit's* pattern-similar verbs show this, and as a result, none of the verbs noted above that were already in VN could accept *benefit* as a new member in their class. On the basis of this analysis, we created a new class for *benefit* (Benefit-72.1), which instantiates the patterns noted above. The verb *profit* has also been added to this class since it can also instantiate these patterns.

5 Adding Light Verbs to VerbNet

Comparisons of VN and PropBank reveal another important difference in coverage: PropBank provides annotations recognizing the unique semantics of English light verb constructions (LVCs). LVCs include expressions like *do an investigation, give a groan, have a drink, make an offer,* and *take a bath*. These constructions therefore consist of a highly polysemous, semantically 'light' verb (Jespersen, 1942) as well as a noun predicate, denoting an event or state, found either in a noun phrase or prepositional phrase complement (e.g. *take into consideration*). In Goldberg's terms (2006: 109), the verbs found in these constructions have relatively low 'cue validity,' indicating that they are not a good predictor of overall sentence meaning. Rather, it is the noun that carries most of the event semantics. The verb does, however, modulate the event semantics in different manners and extents, depending on the LVC. For example, we can clearly see the contribution of the verb when comparing two LVCs with the same eventive noun: *give a bath* versus *take a bath*. Namely, the *give* LVC licenses an additional argument.

It should be noted that both the delimitation and labeling of the constructions outlined here remain nebulous and in debate. What is termed 'LVC' here has also fallen under the labels 'support verb construction,' and 'complex predicate' among others. Furthermore, since Jespersen's (1942, Volume VI:117) application of the term 'light verb' to English V + NP constructions, the term has been extended to constructions with Japanese *suru* 'do' (Grimshaw and Mester, 1988), Romance causatives (Rosen, 1989), Hindi N + V constructions (Mohanan, 1994), Urdu V + V constructions (Butt, 1994), as well as a Chinese variant on control/raising constructions involving *ba* and *de* (Huang, 1992).

It is extremely important for NLP resources to recognize the distinct semantics of LVCs. To support automatic semantic role labeling and inferencing, it is necessary to know, for example, that *Sarah took a bath* does not mean that Sarah grasped a bathtub and went dragging it around somewhere. Instead, this should be recognized as a bathing event. While VN has good coverage of most of the common English light verbs (*do, give, have, make, take*), it does not currently recognize the potential for these verbs to be used within LVCs, and would therefore inevitably misrepresent the semantics of such constructions.

Unfortunately, LVCs can be extremely difficult to detect. LVCs arguably exist on a continuum from purely compositional language that can be interpreted compositionally (e.g. *She made a dress*) to fixed idiomatic expressions with meanings that go far beyond that of the individual lex-

ical items (e.g. *She kicked the bucket*) (Nunberg, Sag and Wasow, 1994). LVCs share some properties of each of these extremes of language because their interpretation is somewhat idiomatic in that the listener must be able to recognize firstly that the verb shouldn't be interpreted in its normal, literal ('heavy') sense, and secondly that the overall meaning stems primarily from the noun. However, they are not completely idiomatic because the noun can usually be interpreted literally, and they certainly cannot be classed with fixed idiomatic expressions because there is quite a bit of syntactic flexibility and, to some extent, substitutability of terms, reflecting LVC's semi-productivity (Nickel, 1978).

LVCs are semi-productive in the sense that novel LVCs are theoretically possible in the pattern of *light verb + eventive/stative noun*, but there are constraints on this productivity. This results in what appear to be semantically similar families of LVCs (e.g. *make a statement, make a speech, make a declaration*), yet other arguably similar LVC combinations are not acceptable to most speakers (e.g. *?make a yell, *make advice*). Additionally, LVCs tend to be syntactically indistinguishable from compositional, heavy usages of the same verb, and in some cases their semantics can be interpreted as either heavy or light: *She made a backup*, which can be thought of as either *She created a backup* (reflecting the heavy sense) or *She backed up...* (reflecting the light sense). For these reasons, while novel LVCs can continuously enter the language, they can be very difficult for both humans and computers to detect and delimit.

Such semi-productive constructions are generally very problematic for lexical resources such as VN, but also FrameNet and WordNet (Fellbaum, 1998), because all of these resources are somewhat static in nature, such that they are currently unable to reflect the possibility for speakers to use verbs in novel contexts that shift and extend their meanings. LVCs, like caused-motion constructions (e.g. *She blinked the snow off of her eyelashes*), are productive enough to be extremely problematic for coverage by a lexical resource (Hwang et al., 2010-b; Bonial et al., 2011). Fixed idiomatic expressions, which are not productive and undergo only morphosyntactic variation, can be stored as a single entry or lexical item, following a words with spaces approach (more flexible idiomatic constructions require a more general treatment). In contrast, the productivity and flexibility of LVCs (both syntactic flexibility and flexibility of adding elements such as determiners and modifiers) make this somewhat impractical. There are promising approaches for the automatic identification of non-frozen, variable idiomatic expressions (e.g. *blow one's own trumpet* and *toot one's own horn*) using measures of both syntactic and lexical fixedness (Fazly, Cook and Stevenson, 2009). Although these methods may also be effective for identifying even low frequency LVCs, they have not yet been applied to this problem. Thus, ideally the constraints on productivity and family resemblances of well-attested LVCs could be leveraged to make predictions about likely LVCs, without the need to exhaustively list each unique LVC.

With such information, VN could be augmented with probabilities that verbs will participate in certain types of constructions, regardless of whether this is an LVC or a coercive construction. Therefore, current work on VN includes efforts to use Hierarchical Bayesian Modeling (HBM) to capture patterns of verb behavior, and therefore statistical likelihoods that a given verb will participate in a given construction, including LVCs (Bonial et al., 2011). As additional corpora are modeled, the HBM, and in turn VN, can continue to evolve to capture the flexible, dynamic nature of language including semi-productive expressions like LVCs.

However, in the case of LVCs, understanding the likelihood for a verb to participate in this construction only addresses half of the problem. Although there are 'families' of semantically similar nouns that pair with a given light verb, there are seemingly idiosyncratic constraints concerning which light verbs pair with which eventive or stative noun, but statistical patterns could also be of assistance in making this prediction. Some of this information is conveniently provided by Sketch Engine.

5.1 Assistance from Sketch Engine

An examination of the Word Sketches of the common light verbs *do, give, have, make* and *take* firstly underscores the importance of including light usages in lexical resources, because they are very common. For example, in the English TenTen corpus of approximately 3.2 billion tokens, the second most frequent object of *do* is the eventive noun *job*, the top four most frequent objects

of *give* are *rise, birth, notice, advice*, the second most frequent object of *have* is *effect*, the most frequent object of *make* is *decision*, and finally, the second most frequent object of *take* is *care*. The tendency for these verbs to pair with predicating nouns to form LVCs is quite clear from Sketch Engine, demonstrating the importance for such usages to be treated appropriately by lexical resources. While Sketch Engine can provide a wealth of information on what nouns are most likely to combine with a particular light verb to form an LVC, it cannot provide information on the semantic classes of nouns that often combine with a given light verb, and therefore can provide little assistance when it comes to detecting less frequent or novel constructions. Unfortunately, it is precisely such generalizations that could be most usefully incorporated into VN, therefore circumventing the need to simply list all attested LVC combinations.

With the aid of collocational tendencies from Sketch Engine, FrameNet can be used as a resource to predict other infrequent or even perhaps novel LVCs, by working under the assumption that if a frequent, attested LVC has a noun that falls into a particular frame, then it is likely that all noun members of that frame could potentially combine with the same light verb. For example, PropBank LVC annotations indicate that many eventive and stative noun collocates with *have* are nouns of mental activities and perception, e.g. *have knowledge, have a thought, have an understanding*. Sketch Engine also reflects this tendency with *have knowledge* as one of the most frequent collocates of *have* in the English TenTen corpus. The nouns of these LVCs are all found in FrameNet's Awareness frame. One could allow an automatic system to assume that any member of the Awareness frame could grammatically combine with *have* to form an acceptable LVC. To investigate the validity of this assumption, the Corpus of Contemporary American English (COCA) (Davies, 2008) was searched for each member of the Awareness frame in combination with *have* within a three word window. The results of this investigation are summarized in Table 1.

Similar searches of the Cogitation frame members and Purpose frame members, which also include nouns of frequent, attested LVCs like *have a thought* and *have the intention*, demonstrate that all of the noun members of these frames are also attested in COCA within light usages. These cursory findings demonstrate that each of the members of these frames have the potential to combine with *have* to create an attested LVC. However, these initial findings also include many false positives due to inevitable overlap with heavy senses and intervening material. For example *application* is a noun found in FrameNet's Purpose frame; *application* in its concrete sense frequently combines with *have* in its heavy, ownership sense: *I had the application on my desk*. Additionally, if the light verb *have* were replaced with a semantically similar verb, such as *possess* in these usages, it is likely that these too would work as LVCs; however, this requires further investigation.

It should be noted that not all of the potential combinations found in these frames would sound grammatical to all, or perhaps even most speakers. Thus, this process does not necessarily predict what would be acceptable LVCs. It simply would allow for computational systems to have a resource that essentially lists potential LVCs, and if and when these are actually used in a corpus, their semantics would be interpreted as likely LVCs instead of heavy usages of the verb. The problem of overlap with heavy senses of the same nouns should also be addressed through continued research using manual PropBank annotations of LVCs and HBM.

5.2 Incorporating Light Verb Resources

The challenge remains of how exactly to incorporate information on light verbs into VN's class structure. This is particularly difficult since VN's existing class membership assumes that event semantics stem primarily from verbs. Thus, it seems most appropriate for this information to exist in a supplementary resource to VN. When a verb is relatively frequently realized as a light verb, then this would be added as a sense when one searched for this verb in VN. Instead of this search taking one to a sense located in a VN class, however, selecting the light sense would provide information on the most common eventive and stative noun collocates of that light verb, along with links to the associated FrameNet frames. This information can currently be drawn from the manual PropBank annotations, and ideally in the future could be expanded through the aforementioned research using HBM. The collocational tendencies found in the smaller PropBank corpus can also be verified

Awareness Frame Members	Number of Instances	Example Usage
awareness	687	She **had** a fascinating **awareness** of the space around her.
comprehension	107	This suggests that students in our sample **had**, on average, higher **comprehension** in Spanish than in English.
conception	300	They **had** a different **conception** of what was going to happen.
consciousness	504	That night she **had** a new **consciousness** of the country, felt almost a new relation to it.
hunch	319	I **had** a **hunch** you were more than just a pretty face.
ignorance	103	But we, the art-beholders, **have** no such **ignorance**.
knowledge	4729	This study found that pet owners **had** a basic **knowledge** of rabies and the quarantine.
presumption	73	Americans have always **had** a **presumption** that you will not do your job.
suspicion	934	In my mind, I truly **had suspicion** that she had tried to take her own life on that cliff.
thought	4831	Acknowledge he **had thoughts** of leaving her.
understanding	2449	That much planning implies he **had** a clear **understanding** of his actions and he understood the consequences...

Table 1: COCA instances of *have* LVCs from FrameNet Awareness frame.

against those of Sketch Engine's larger corpora, and additional LVC combinations could also be discovered through a manual inspection of common light verb's collocations in Sketch Engine.

6 Conclusions and Future Work

This research generally demonstrates the efficacy of considering complementary lexical resources together, for frequently the information provided by such comparisons is greater than what can be gleaned by an individual resource alone. Specifically, Sketch Engine's Word Sketch and Thesaurus function can be extremely informative and useful in expanding and adding VN classes. The two resources are quite complementary in that VN makes important theoretical assumptions about syntax underlying semantics, and Sketch Engine simply reports syntactic and collocational information from large corpora, yet this information often leads to fruitful expansion of classes. In some senses, the very fact that Sketch Engine can be used so successfully to expand VN underscores the validity of Levin's hypothesis that syntax is a reflection of semantics.

Sketch Engine can also be useful in discovering common LVCs, and FrameNet can be leveraged to discover other likely LVC combinations based on the frequent existing LVCs. Although this has yet to be fully investigated, the Thesaurus function could also be used to find semantically similar nouns to those in attested LVCs, which could lead to the discovery of additional families of LVCs, in the same way that FrameNet frames could be used. The Thesaurus function could then potentially lead to suggestions for new candidates to add to FrameNet frames as well. Future work will also include an investigation of whether or not there is a systematic selection of nouns or noun classes that are compatible with specific light verbs. Of special interest on this topic would be whether or not more semantically general super-type nouns are more often compatible within LVCs as compared to more semantically specified subtypes (e.g. *take a walk* vs. **take a limp*).

With the help of these resources, VN can be expanded to include the relatively frequent, but difficult cases discussed here. Future work expanding LVC annotations in PropBank and discovering LVCs automatically using HBM will allow for VN to flexibly account for constructions like LVCs and coercive constructions, in which verbs can be used in novel and semantically distinct ways. Such flexibility will greatly improve currently static resources like VN and allow lexicons to more closely reflect what one might imagine a speaker's lexicon to be: dynamically updating through experience with novel lexical items in novel contexts.

Acknowledgments

We gratefully acknowledge the support of the National Science Foundation Grant NSF-IIS-1116782, A Bayesian Approach to Dynamic Lexical Resources for Flexible Language Processing. 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.

References

Claire Bonial, Susan Windisch Brown, Jena D Hwang, Christopher Parisien, Martha Palmer and Suzanne Stevenson. 2011. Incorporating Coercive Constructions into a Verb Lexicon. *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics held in conjunction with ACL-2010*. Portland, Oregon.

Miriam Butt. 1995. *The Structure of Complex Predicates in Urdu*. Dissertation in Linguistics. CSLI Publications, Stanford.

M Davies. 2008-. *The Corpus of Contemporary American English (COCA): 425 million words*. Available online at http://www.americancorpus.org.

Afsaneh Fazly, Paul Cook and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*. 35(1): 61–103 Uppsala, Sweden.

Christiane Fellbaum (Ed.) 1998. *Wordnet: An Electronic Lexical Database*. MIT press, Cambridge.

Charles J Fillmore, Christoper R Johnson, and Miriam R L Petruck. 2002. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Adele Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, New York.

J. Grimshaw and A. Mester. 1988. Light verbs and Theta-marking. *Linguistic Inquiry*, 19(2):205-232.

Charles Huang. 1992. Complex Predicates in Control. *Control and Grammar*, ed. Richard Larson, Sabine Iatridou, and Utpal Lahiri. 109-147. Kluwer Academic Publishers.

Jena D. Hwang, Archna Bhatia, Claire Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue and Martha Palmer. 2010-a. PropBank Annotation of Multilingual Light Verb Constructions *Proceedings of the Linguistic Annotation Workshop held in conjunction with ACL-2010.*. Uppsala, Sweden.

Jena D. Hwang, Rodney D. Nielsen and Martha Palmer. 2010-b. Towards a Domain Independent Semantics: Enhancing Semantic Representation with Construction Grammar. *Proceedings of Extracting and Using Constructions in Computational Linguistic Workshop held in conjunction with NAACL HLT 2010*. Los Angeles, CA.

Otto Jespersen. 1942. *A Modern English Grammar on Historical Principles*. Part VI: Morphology. Ejnar Munksgaard, Copenhagen.

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

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. *Proceedings of EURALEX*. Lorient, France.

Karin Kipper, Anna Korhonen, Neville Ryant and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation Journal*, 42:21–40

Anna Korhonen and T. Briscoe. 2004. Extended lexical-semantic classification of english verbs. *Proceedings of HLT/NAACL Workshop on Computational Lexical Semanticsl* Boston, Massachusetts.

Beth Levin. 1983. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Edward Loper, S. Yi, and Martha Palmer. 2007. Combining lexical resources: Mapping between PropBank and VN. *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS-7)* Tilburg.

Tara Mohanan. 1994. *Argument Structure in Hindi*. CSLI Publications, Stanford.

Gerhard Nickel. 1978. Complex Verbal Structures in English. *Studies in Descriptive English Grammar* 63–83.

G. Nunberg, Ivan A Sag, and T. Wasow. 1994. Idioms. *Language* 70: 491–538.

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

Martha Palmer. 2009. Semlink: Linking PropBank, VN and FrameNet. *Computational Linguistics* 31(1):71–106.

Sameer Pradhan, E. Hovy, M. Marcus, M. Palmer, L. Ramshaw and R. Weischedel. 2007. OntoNotes: A unified relational semantic representation. *Proceedings of the First IEEE International Conference on Semantic Computing(ICSC-07)*. Irvine, CA.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

Sara Rosen. 1989. Argument Structure and Complex Predicates. Doctoral dissertation, Brandeis University.

R. Swier and Suzanne Stevenson. 2004. Unsupervised semantic role labeling. *Proceedings of the Generative Lexicon Conference, GenLex-09*. Pisa, Italy.

Annie Zaenen, C. Condoravdi, and D G. Bobrow. 2008. The encoding of lexical implications in VN. *Proceedings of LREC 2008*. Morocco