PropBank Annotation of Multilingual Light Verb Constructions

Jena D. Hwang[1], Archna Bhatia[3], Clare Bonial[1], Aous Mansouri[1], Ashwini Vaidya[1], Nianwen Xue[2], and Martha Palmer[1]

[1]Department of Linguistics, University of Colorado at Boulder, Boulder CO 80309
[2]Department of Computer Science, Brandeis University, Waltham MA 02453
[3]Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana IL 61801

{hwangd,claire.bonial,aous.mansouri,ashwini.vaidya,martha.palmer}
@colorado.edu, bhatia@illinois.edu, xuen@brandeis.edu

Abstract

In this paper, we have addressed the task of PropBank annotation of light verb constructions, which like multi-word expressions pose special problems. To arrive at a solution, we have evaluated 3 different possible methods of annotation. The final method involves three passes: (1) manual identification of a light verb construction, (2) annotation based on the light verb construction's Frame File, and (3) a deterministic merging of the first two passes. We also discuss how in various languages the light verb constructions are identified and can be distinguished from the non-light verb word groupings.

1 Introduction

One of the aims in natural language processing, specifically the task of semantic role labeling (SRL), is to correctly identify and extract the different semantic relationships between words in a given text. In such tasks, verbs are considered important, as they are responsible for assigning and controlling the semantic roles of the arguments and adjuncts around it. Thus, the goal of the SRL task is to identify the arguments of the predicate and label them according to their semantic relationship to the predicate (Gildea and Jurafsky, 2002; Pradhan et al., 2003).

To this end, PropBank (Palmer et. al., 2005) has developed semantic role labels and labeled large corpora for training and testing of supervised systems. PropBank identifies and labels the semantic arguments of the verb on a verb-by-verb basis, creating a separate Frame File that includes verb specific semantic roles to account for each subcategorization frame of the verb. It has been shown that training supervised systems with PropBank's semantic roles for shallow semantic analysis yield good results (see CoNLL 2005 and 2008).

However, semantic role labeling tasks are often complicated by multiword expressions (MWEs) such as idiomatic expressions (e.g., 'Stop *pulling my leg*!'), verb particle constructions (e.g., 'You must *get over* your shyness.'), light verb constructions (e.g., '*take a walk*', '*give a lecture*'), and other complex predicates (e.g., V+V predicates such as Hindi's निकल गया *nikal gayaa*, lit. 'exit went', means 'left' or 'departed'). MWEs that involve verbs are especially challenging because the subcategorization frame of the predicate is no longer solely dependent on the verb alone. Rather, in many of these cases the argument structure is assigned by the union of two predicating elements. Thus, it is important that the manual annotation of semantic roles, which will be used by automatic SRL systems, define and label these MWEs in a consistent and effective manner.

In this paper we focus on the PropBank annotation of light verb constructions (LVCs). We have developed a multilingual schema for annotating LVCs that takes into consideration the similarities and differences shared by the construction as it appears in English, Arabic, Chinese, and Hindi. We also discuss in some detail the practical challenges involved in the crosslinguistic analysis of LVCs, which we hope will bring us a step closer to a unified crosslinguistic analysis.

Since NomBank, as a companion to PropBank, provides corresponding semantic role

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 82–90,
Uppsala, Sweden, 15-16 July 2010.

labels for noun predicates (Meyers et al., 2004), we would like to take advantage of NomBank's existing nominalization Frame Files and annotations as much as possible. A question that we must therefore address is, "Are nominalization argument structures exactly the same whether or not they occur within an LVC?" as will be discussed in section 6.1.

2 Identifying Light Verb Constructions

Linguistically LVCs are considered a type of a complex predicate. Many studies from differing angles and frameworks have characterized complex predicates as a fusion of two or more predicative elements. For example, Rosen (1997) treats complex structures as complementation structures, where the argument structure of elements in a complex predicate are fused together. Goldberg (1993) takes a constructional approach to complex predicates and arrives at an analysis that is comparable to viewing complex predicates as a single lexical item. Similarly, Mohanan (1997) assumes different levels of linguistic representation for complex predicates in which the elements, such as the noun and the light verb, functionally combine to give a single clausal nucleus. Alsina (1997) and Butt (1997) suggest that complex predicates may be formed by syntactically independent elements whose argument structures are brought together by a predicate composition mechanism.

While there is no clear-cut definition of LVCs, let alone the whole range of complex predicates, for the purposes of this study, we have adapted our approach largely from Butt's (2004) criteria for defining LVCs. LVCs are characterized by a light verb and a predicating complement (henceforth, *true predicate*) that "combine to predicate as a single element." (Ibid.) In LVC, the verb is considered semantically bleached in such a way that the verb does not hold its full predicating power. Thus, the light verb plus its true predicate can often be paraphrased by a verbal form of the true predicate without loss of the core meaning of the expression. For example, the light verb 'gave' and the predicate 'lecture' in 'gave a lecture', together form a single predicating unit such that it can be paraphrased by 'lectured'.

True predicates in LVCs can be a noun (the object of the verb or the object of the preposition in a prepositional phrase), an adjective, or a verb. One light verb plus true predicate combination found commonly across all our PropBank languages (i.e., English, Arabic, Chinese, and Hindi) is the noun as the object of the verb as in 'Sara **took [a stroll]** along the beach'. In Hindi, true predicates can be adjectives or verbs, in addition to the nouns.

मुझे तुम [**अच्छे**] **लगे** (Adjective)
to-me you [**nice] seem**
lit. 'You seem nice to me'
'You (are) liked to me (=I like you).'

मैंने सब कुछ [**कर**] **लिया** (Verb)
I-ERG everything [**do] took**
lit. 'I took do everything'
'I have done everything.'

As for Arabic, the LVCs come in verb+noun pairings. However, they surface in two syntactic forms. It can either be the object of the verb just like in English:

القى جورج [**محاضرة**] عن لبنان
gave.he Georges [**lecture**] PREP Lebanon
lit.'Georges gave a lecture about Lebanon'
'Georges lectured about Lebanon'

or the complement can be the object of a preposition:

سأقوم [بزيارة] سيدنا إلياس
conduct.I [PREP-visit] our.saint Ilias
lit. 'I will conduct with visit Saint Ilias's'
'I will visit Saint Ilias's'

3 Standard PropBank Annotation Procedure

The PropBank annotation process can be broken down into two major steps: creation of the Frame Files for verbs occurring in the data and annotation of the data using the Frame Files. During the creation of the Frame Files, the usages of the verbs in the data are examined by linguists (henceforth, "framers"). Based on these observations, the framers create a Frame File for each verb containing one or more framesets, which correspond to coarse-grained senses of the predicate lemma. Each frameset specifies the PropBank labels (i.e., ARG0, ARG1,…ARG5) corresponding to the argument structure of the verb. Additionally, illustrative examples are included for each frameset, which will later be referenced by the annotators. These examples also include the use of the ARGM labels.

Thus, the framesets are based on the examination of the data, the framers' linguistic knowledge and native-speaker intuition. At

times, we also make use of the syntactic and semantic behavior of the verb as described by certain lexical resources. These resources include VerbNet (Kipper et. al., 2006) and FrameNet (Baker et. al., 1998) for English, a number of monolingual and bilingual dictionaries for Arabic, and Hindi WordNet and DS Parses (Palmer et. al., 2009) for Hindi. Additionally, if available, we consult existing framesets of words with similar meanings across different languages.

The data awaiting annotation are passed onto the annotators for a double-blind annotation process using the previously created framesets. The double annotated data is then adjudicated by a third annotator, during which time the differences of the two annotations are resolved to produce the Gold Standard.

Two major guiding considerations during the framing and annotating process are data consistency and annotator productivity. During the frameset creation process, verbs that share similar semantic and syntactic characteristics are framed similarly. During the annotation process, the data is organized by verbs so that each verb is tackled all at once. In doing so, we firstly ensure that the framesets of similar verbs, and in turn, the annotation of the verbs, will both be consistent across the data. Secondly, by tackling annotation on verb-by-verb basis, the annotators are able to concentrate on a single verb at a time, making the process easier and faster for the annotators.

4 Annotating LVC

A similar process must be followed when annotating light verb constructions The first step is to create consistent Frame Files for light verbs. Then in order to make the annotation process produce consistent data at a reasonable speed, we have decided to carry out the light verb annotation in three passes (Table 1): (1) annotate the light verb, (2) annotate the true predicate, and (3) merge the two annotations into one.

The first pass involves the identification of the light verb. The most important parts of this step are to identify a verb as having bleached meaning, thereafter assign a generic light verb frameset and identify the true predicating expression of the sentence, which would be marked with ARG-PRX (i.e., ARGument-PRedicating eXpression). For English, for example, annotators were instructed to use Butt's (2004) criteria as described in Section 2. These criteria required that annotators be able to recognize whether or not the complement of a potential light verb was itself a predicating element. To make this occasionally difficult judgment, annotators used a simple heuristic test of whether or not the complement was headed by an element that has a verbal counterpart. If so, the light verb frameset was selected.

The second pass involves the annotation of the sentence with the true predicate as the relation. During this pass, the true predicate is annotated with an appropriate frameset. In the third pass, the arguments and the modifiers of the two previous passes are reconciled and merged into a single annotation. In order to reduce the number of hand annotation, it is preferable for this last pass, the Pass 3, to be done automatically.

Since the nature of the light verb is different from that of other verbs as described in Section 2, the advantage of doing the annotation of the light verb and the true predicate on separate passes is that in the light verb pass the annotators will be able to quickly dispose of the verb as a light verb and in the second pass, they will be allowed to solely focus on the annotation of the light verb's true predicate.

The descriptions of how the arguments and modifiers of the light verbs and their true predicates are annotated are mentioned in Table 1, but notably, none of the examples in it currently include the annotation of arguments

	Pass 1: Light Verb Annotation	Pass 2: True Predicate Annotation	Pass 3: Merge of Pass1&2 Annotation
Relation	Light verb	True predicate	Light verb + true predicate
Arguments and Modifiers	- Predicating expression is annotated with ARG-PRX - Arguments and modifiers of the light verb are annotated	- Arguments and modifiers of the true predicate are annotated	- Arguments and modifiers found in the two passes are merged, preferably automatically.
Frameset	Light verb frameset	True predicate's frameset	LVC's frameset
Example	***"John took a brisk walk through the park."***		
	REL: took ARG-PRX: a brisk walk	ARG-MNR: brisk REL: walk	REL: took walk ARG-MNR: brisk

Table 1. Preliminary Annotation Scheme

and modifiers. This is intentional, as coming to an agreement concerning the details of what exactly each of the three passes looks like while meeting the needs of the four PropBank languages is quite challenging. Thus, for the rest of the paper we will discuss the strengths and weaknesses of the two trial methods of annotation we have considered and discarded in Section 5, as well as the final annotation scheme we chose in Section 6.

5 Trials

5.1 Method 1

As our first attempt, the annotation of argument and adjuncts was articulated in the following manner (Table 2).

Pass 1:	Pass 2:
Light verb	True predicate
- Predicating expression is labeled ARG-PRX *- Annotate the Subject argument of the light verb as the Arg0.* *- Annotate the rest of the arguments and modifiers of the light verb with ARGM labels.*	*- Annotate arguments and modifiers of the true predicate within its domain of locality.*
Generic light verb Frame File	True predicate's Frame File
"John took a brisk walk through the park."	
ARG0: John REL: took ARG-PRX: a brisk walk ARG-DIR: through the park	ARG-MNR: brisk REL: walk

Table 2. Method 1 for annotation for Passes 1 and 2. Revised information is in italics.

In Pass 1, in addition to annotating the predicating expression of the light verb with ARG-PRX, the subject argument was marked with an ARG0. The choice of ARG0, which corresponds to a proto-typical agent, was guided by the observation that English LVCs tend to lend a component of agentivity to the subject even in cases where the true predicate would not necessarily assign an agent as its subject. The rest of the arguments and modifiers were labeled with corresponding ARGM (i.e., modifier) labels. The assumption here is that the arguments of the light verb will also be the arguments of the true predicate.

In Pass 2, then, the annotation of the arguments of the true predicate was restricted to its domain of locality (i.e., the span of the ARG-PRX as marked in Pass1). That is, in the example 'John took a brisk walk through the park', the labeled spans for the true predicate would be limited to the NP 'a brisk walk' and neither 'John' nor through the park' would be annotated as the arguments of the true predicate 'walk'.

Frame Files: This method would require three Frame Files: a generic light verb Frame File, a true predicate Frame File, and an LVC Frame File. The Frame File for the light verb would not be specific to the form of the light verb (e.g., same frame for *take* and *make*). Rather, it would indicate a skeletal argument structure in order to reduce the amount of Frame Files made, including only Arg0 as its argument[1].

5.2 Weakness of Method 1

This method has one glaring problem: the assumption that the semantic roles of the arguments as assigned by the light verb uniformly coincide with those assigned by the true predicate does not always hold. Consider the following English sentence[2].

> whether Wu Shu-Chen would **make** another **[appearance]** in court was subject to observation

In this example, 'Wu Shu-Chen' is the agent argument (Arg0) of the light verb 'make' and is the theme or patient argument (Arg1) of a typical 'appearance' event. Also consider the following example from Hindi.

It is possible that in a light verb construction, the light verb actually modifies the standard underlying semantics of a nominalization like appearance. In any event, we cannot assume that the expected argument labels for the light verb and for the standard interpretation of the nominalization will always coincide. Thus, we could say that Pass 2's true predicate annotation is only partial and is not representative of the complete argument structure. In particular, we are left with a very difficult merging problem, because the argument labels of the two separate passes conflict as seen in the above examples.

5.3 Method 2

In order to remedy the problem of conflicting argument labels, we revised Method 1's Pass 2 annotation scheme. This is shown in Table 3. Pass 1 remains unchanged from Method 1.

In this method, both the light verb and the true predicate of the sentence receive complete sets of

[1] This is why the rest of the argument/modifiers would be annotated using ARGM modifier labels.

[2] The light verb is in boldface, the true predicate is in bold and square brackets, and the argument/adjunct under consideration is underlined.

Pass 2:
True predicate
- *Annotate the Subject argument of the light verb with the appropriate role of the true predicate* - Annotate arguments and modifiers of the true predicate *without limitation as to the domain of locality.*
True predicate's Frame File
"He made another appearance at the party"
ARG1: He ARG-ADV: another REL: appearance ARG-DIR: at court

Table 3. Method 2 for annotation for Pass 2. Pass 1 as presented in Table 2 remains unchanged. Revised information for Pass 2 is in italics

argument and modifier labels. In Pass 2, the limitation of annotating within the domain of locality is removed. That is, the arguments and modifiers inside and outside the true predicate's domain of control are annotated with respect to their *semantic relationship to the true predicate* (e.g., in the English example of Section 5.2, 'Wu Shu-Chen' would be considered ARG1 of 'appearance').

Frame Files: This method would also require three Frame Files. The major difference is that with this method the Frame File for the true predicate includes arguments that are sisters to the light verb.

5.4 Weaknesses of Method 2

If in Method 1 we have committed the error of semantic unfaithfulness due to omission, in Method 2 we are faced with the problem of including too much. In the following sentence, consider the role of the underlined adjunct:

A New York audience … **gave** it a big round of **applause** when the music started to play.

By the annotation in Method 2, the underlined temporal adjunct 'when the music started to play' is labeled as both the argument of 'give' and of 'applause'. The question here is does the argument apply to both the giving and the applauding event? In other words, does the adjunct play an equal role in both passes?

Since it could be easily said that the temporal phrase applies to both the applauding and the giving of the applause events, this example may not be particularly compelling. However, what if a syntactic complement of the light verb is a semantic argument of the true predicate and the true predicate only? This is seen more frequently in the cases where the light verb is less bleached than in the case of 'give' above. Consider the following Arabic example.

أخذنا في [**الاعتبار**] في تحضيراتنا إمكان تكبدهم خسائر
took.we PREP DEF-consideration PREP prepertations.our possibility sustain.their losses
'We **took** into **[consideration]** during our preparations the possibility of them sustaining losses'

Here, even though the constituent 'of them sustaining losses' is the syntactic complement of the verb 'to take;' semantically, it modifies only the nominal object of the PP 'consideration.'

There are similar phenomena in Chinese light verb constructions. Syntactic modifiers of the light verb are semantic arguments of the true predicate, which is usually a nominalization that serves as its complement.

我们 正 对 这 个 问题 [进行] 讨论 。
we now regarding this CL issue [**conduct] discussion**.
lit."We are conducting a discussion on this issue."
"We are discussing this issue."

The prepositional phrase **对这个问题** 'regarding this issue' is a sister to the light verb but semantically it is an argument of the nominalized predicate **讨论** 'discussion'.

The logical next question would be: does the annotation of the arguments, adjuncts and modifiers have to be all or nothing? It could conceivably be possible to assign a selected set of arguments at the light verb or true predicate level. For example, in the Chinese sentence, the modifier 'regarding this CL issue', though a syntactic adjunct to the light verb, could be left out from the semantic annotation in Pass 1 and included only in the Pass 2.

However, the objection to this treatment comes from a more practical need. As mentioned above, in order to keep the manual annotation to a minimum, it would be necessary to keep Pass 3 completely deterministic. As is, with the unmodified Method 2, there would be the need to choose between Pass 1 or Pass 2 annotation to when doing the automatic Pass 3. If we modify Method 2 by annotating only a selected set of syntactic arguments for the light verb or the true predicate, then this issue is exacerbated. In such a case there we would have to develop with strict rules for which arguments of which pass should be included in Pass 3. Pass 3 would no longer be automatic, and should be done manually.

	Pass 1: Light Verb Identification	Pass 2: LVC Annotation	Pass 3: Deterministic relation merge
Relation	Light verb	True predicate	Light verb + true predicate
Arguments & Modifiers	- Predicating expression is annotated with ARG-PRX	- Arguments and modifiers of the LVCs are annotated	- Arguments and modifiers are taken from Pass 2
Frame File	<no Frame File needed>	LVC's Frame File	LVC's Frame File
Example	***"John took a brisk walk through the park."***		
	REL: took ARG-PRX: a brisk walk	ARG0: John ARG-MNR: brisk REL: walk ARGM-DIR: through the park	ARG0: John ARG-MNR: brisk REL: [took][walk] ARGM-DIR: through the park

Table 4. Final Annotation Scheme

6 Final Annotation Scheme

6.1 Semantic Fidelity

Many of the objections so far to Methods 1 and 2 have centered on the issue of semantic fidelity during the annotation of each of the two passes. The debate of whether both passes should be annotated and to what extent has practical implications for the third Pass, as described above. However, more importantly it comes down to whether or not the semantics of the final *light verb plus true predicate combination* is indeed distinct from the semantics of its parts (i.e. light verb and true predicate, separately). This may be a fascinating linguistic question, but it is not something our annotators can be debating for each and every instance.

Instead, we argue that the semantic argument structure of the *light verb plus true predicate combination* can in practice be different from that of the expressions taken independently as has been proposed by various studies (Butt, 2004; Rosen, 1997; Grimshaw & Mester, 1988). Thus, we resolve the cases in which the differences in argument roles as assigned by the light verb and the nominalization (Section 5.2) by handling the argument structure of the standard nominalization separately from that of the nominalization participating in the LVC. In the example 'Chen made another appearance in court', we annotate 'Chen' as the Agent (ARG0) of the full predicate '[make] [appearance]', which is different from the argument structure of the standard nominalization which would label 'Chen' to be the Patient argument (ARG1).

6.2 Method 3: Final Method

Our final method of light verb annotation reflects the notion that the noun, verb, or adjective as a true predicate within an LVC can have a different argument structure from that of the word alone. Table 4 shows the final annotation scheme for light verb construction.

During Pass 1, the LVCs and their predicating expressions are identified in the data. Instances identified as LVCs in Pass 1 are then manually annotated during Pass 2, annotating the arguments and adjuncts of the light verb and the true predicate with roles that reflect their semantic relationships to the *light verb plus true predicate*. In practice, Pass 1 becomes a way of simply manually identifying the light verb usages. It is in Pass 2 that we make the final choice of argument labels for all of the arguments. Thus in Pass 3, the light verb and the true predicate lemmas from Pass 1 and 2 are joined into a single unit (e.g., in the example found in Table 4, the light verb 'took' would be joined with the true predicate 'walk' into 'took+walk')[3]. In this final method, Pass 3 can be achieved completely deterministically.

The major difference in this annotation scheme from that of Methods 1 and 2 is that instead of annotating in terms of the semantics of the bare noun, adjective or verb, the argument structure is determined for the entire predicate or the full event: semantics of the *light verb plus the true predicate*. This means that for the sentences where the argument roles of the verb and the nominalization disagree like 'Chen' in 'Chen

3 The order of Pass 2 and Pass 3 as presented in Table 4 is arguably a product of how the annotation tools for PropBank are set up for Arabic, Chinese, and English. That is, the order of the Pass 2 and Pass 3 could potentially be flipped provided that the tools and procedures of annotation support it, as is the case for Hindi PropBank. After the LVC and ARG-PRX are identified in Pass 1, the light verb and the true predicate can be deterministically joined into a single relation in Pass 2, leaving the manual annotation of LVC for Pass 3. The advantage of this alternative ordering is that because the annotation of LVC is done around light verb plus the true predicate as a single relation, rather than the true predicate alone as in Table 4, the argument annotation may in actuality be more intuitive for annotators even with less training.

made another[4] appearance in court', we label the argument with the role that is consistent with the entire predicate (i.e. Agent, ARG0).

Frame Files: The final advantage to this method is that only one Frame File is needed. Since Pass 1 is an identification round, no Frame File is required. A single Frame File for LVC that includes the argument structure with respect to the *light verb plus true predicate combination* will suffice for Pass 2 and Pass 3.

7 Distinguishing LVCs from MWEs

As we have discussed in Section 2, we adapted our approach from Butt's (2004) definition of LVCs. That is, an LVC is characterized by a semantically bleached light verb and a true predicate. These elements combine as a single predicating unit, in such a way that the light verb plus its true predicate can be paraphrased by a verbal form of the true predicate without loss of the core meaning of the expression (e.g. 'lectured' for 'gave a lecture'). Also, as discussed in Section 6.1, our approach advocates the notion that the semantic argument structure of the *light verb plus true predicate* is different from that of the expressions taken independently (as also proposed by Butt, 2004; Rosen, 1997; Grimshaw & Mester, 1988 among others).

While these definitions are appropriate for the PropBank annotation task as we have presented it, there are still cases that merit closer attention. Even English with a rather limited set of verbs that are commonly cited as LVCs, includes a problematic mixture of what could arguably be termed either LVCs or idiomatic expressions: 'make exception', 'take charge'. This difficulty in part is the effect of frequency and entrenchment of particular constructions. The light verbs themselves do not diminish in form over time in a manner similar to auxiliaries (Butt, 2004), although the complements of common LVCs can change over time such that it is no longer clear that the complement is a predicating element.

In the case of English, the expressions 'take charge' may be more commonly found today as a LVC than independently in its verbal form. As we discovered with our annotators, native English speakers are uncomfortable using the verb 'charge' (i.e. to burden with a responsibility) as an independent matrix verb. A similar phenomenon can be seen in Arabic, where the predicate أطلق اسم lit. 'release name' exemplifies a prototypical LVC that means 'to name'. However, in our data we see cases in which the complement is missing, while the semantics of the LVC remains intact:

أو ما يطلق عليه "القطاع العام"
CONJ REL be released.he PREP-him/it
DEF-sector DEF-public
lit 'Or what is released to it "the public sector"'
'Or what is called/named "the public sector."'

This raises the question of: when does a construction that may have once been an LVC become more properly defined as an idiomatic expression due to such entrenchment? Idiomatic expressions can potentially be distinguished from LVCs through judgments of how fixed or syntactically variable a construction is, and on the basis of how semantically transparent or decomposable the construction is (Nunberg et. al., 1994). However, sometimes the dividing line is hard to draw.

A similar problem arises in determining whether a construction is a case of an LVC or simply a usage with a distinct sense of the verb. Take, for example, the following Arabic sentence.

تناول الغذاء
take.he DEF-food
lit. '(he) took food'
'he ate'

Here, the Arabic word غذاء 'food' is the noun derivation of the root shared by the verb تغذى 'to eat', in such a way that the sentence could be rephrased as تغذى '(he) ate'. This example falls neatly into the LVC category. However, further examples suggest that the example is a case of a distinct sense of 'to take orally' where the restrictions on the object are that the theme must be something that can be taken by mouth:

تناول الحساء	تناول الدواء
take.he DEF-soup	take.he DEF-medicine
'he took soup'	'he took medicine'

Finally, determining the appropriate criteria to distinguish between a truly semantically bleached verb and verbs that seem to be participating in complex predication but contribute more to the semantics of the construction is a challenge for all languages. For example, in English data, there are potential LVCs with verbs that are not often thought of as light verbs, such as 'produce an alteration' and

[4] The adjective 'another' is annotated as the modifier of the full predicate '[make][appearance]' as it can be interpreted to mean that the make appearance event happened a previous appearance has been made.

'issue a complaint'. Although most English speakers would agree that the verbs in these constructions do not contribute to the semantics of the construction (e.g. 'issue a complaint' can be paraphrased to 'to complain'), there are similar constructions such as 'register a complaint,' wherein the verb cannot be considered light. For the purposes of annotation, where it is necessary for annotators to understand clear criteria for distinguishing light verbs, such cases are highly problematic because there is no deterministic way to measure the extent to which the verbal element contributes to the semantics of the construction. In turn, there is not a good way to distinguish some of these borderline verbs from their normal, heavy usages.

Such problems can be resolved by establishing language-specific semantic or syntactic tests that can be used for taking care of the borderline cases of LVCs. However, there is one other plausible manner we have identified that could help in detecting such atypical LVCs. This can be done by focusing on the argument structures of predicating complements rather than focusing on the verbs themselves. Grimshaw & Mester (1988) suggest that the formation of LVCs involves argument transfer from the predicating complement to the verb, which is semantically bleached and thematically incomplete and assigns no thematic roles itself. Similarly, Stevenson *et al.* (2004) suggest that the acceptability of a potential LVC depends on the semantic properties of the complement. Thus, atypical LVCs, such as the English construction 'issue a complaint,' can potentially be detected during the annotation of eventive nouns, planned for all PropBank languages.

This process will make our treatment of LVCs more comprehensive. Used with our language-specific semantic and syntactic criteria relating to both the verb and the predicating complement, it will help us to more effectively capture as many types of LVCs as possible, including those of the V+ADJ and V+V varieties.

8 Usefulness of our Approach

Two basic approaches have previously been taken to handle all types of MWEs, including LVCs in natural language processing applications. The first is to treat MWEs quite simply as fixed expressions or long strings of words with spaces in between; the second is to treat MWEs as purely compositional (Sag et al., 2002). The words-with-spaces approach is adequate for handling fixed idiomatic expressions, but issues of lexical proliferation and flexibility quickly arise when this approach is applied to light verbs, which are syntactically flexible and can number in the tens of thousands for a given language (Stevenson et al., 2004; Sag et al., 2002). Nonetheless, large-scale lexical resources such as FrameNet (Baker et al., 1998) and WordNet (Fellbaum, 1999) continue to expand with entries that are MWEs.

The purely compositional approach is also problematic for light verbs because it is notoriously difficult to predict which light verbs can grammatically combine with other predicating elements; thus, this approach leads to problems of overgeneration (Sag et al., 2002). In order to overcome this problem, Stevenson et al. (2004) attempted to determine which nominalizations could form a valid complement to the English light verbs *take, give* and *make,* using Levin's (1993) verb classes to group similar nominalizations. This approach was rather successful for *take* and *give*, but inconclusive for the verb *make*.

Our approach can help to develop a resource that is useful whether one takes a words-with-spaces approach or a compositional approach. Specifically, for those implementing a words-with-spaces approach, the resulting PropBank annotation can serve as a lexical resource listing for LVCs. For those interested in implementing a compositional approach the PropBank annotation can serve to assist in predicting likely combinations. Moreover, information in the PropBank Frame Files can be used to generalize across classes of nouns that can occur with a given light verb with the help of lexical resources such as WordNet (Fellbaum, 1998), FrameNet (Baker et. al., 1998), and VerbNet (Kipper-Schuler, 2005) (in a manner similar to the approach of Stevenson et al. (2004)).

Acknowledgements

We also gratefully acknowledge the support of the National Science Foundation Grant CISE-CRI 0709167, Collaborative: A Multi-Representational and Multi-Layered Treebank for Hindi/Urdu, and a grant from the Defense Advanced Research Projects Agency (DARPA/IPTO) under the GALE program, DARPA/CMO Contract No HR0011-06-C-0022, subcontract from BBN, Inc.

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.

Reference

Alsina, A. 1997. Causatives in Bantu and Romance. In A. Alsina, J. Bresnan, and P. Sells eds. *Complex Predicates*. Stanford, California: CSLI Publications, p. 203-246.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In Proceedings of the 17th International Conference on Computational Linguistics (COLING/ACL-98), pages 86–90, Montreal. ACL.

Butt, M. 2004. 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, M. 1997. Complex Predicates in Urdu. In A. Alsina, J. Bresnan, and P. Sells eds. *Complex Predicates*. Stanford, California: CSLI Publications, p. 107-149.

Fellbaum, Christine, ed.: 1998, *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press.

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

Gildea, Daniel and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. Computational Linguistics 28:3, 245-288.

Goldberg, Adele E. 2003. "Words by Default: Inheritance and the Persian Complex Predicate Construction." In E. Francis and L. Michaelis (eds). Mismatch: Form-Function Incongruity and the Architecture of Grammar. CSLI Publications. 84-112.

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

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: Chicago Univ. Press.

Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank Project: An interim report. In Proceedings of the HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation, pages 24- 31, Boston, MA. pages 430–437, Barcelona, Spain.

Mohanan, T. 1997. Multidimensionality of Representation: NV Complex Predicates in Hindi. In A. Alsina, J. Bresnan, and P. Sells eds. *Complex Predicates*. Stanford, California: CSLI Publications, p. 431-471.

Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, Fei Xia, Hindi Syntax: Annotating Dependency, Lexical Predicate-Argument Structure, and Phrase Structure, In the Proceedings of the 7th International Conference on Natural Language Processing, ICON-2009, Hyderabad, India, Dec 14-17, 2009

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics, 31(1):71–106.

Pradhan, Sameer, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin, Daniel Jurafsky. 2004. Shallow Semantic Parsing using Support Vector Machines. University of Colorado Technical Report: TR-CSLR 2003-03.

Rosen, C. 1997. Auxiliation and Serialization: On Discerning the Difference. In A. Alsina, J. Bresnan, and P. Sells eds. *Complex Predicates*. Stanford, California: CSLI Publications, p. 175-202.

Sag, I., Baldwin, T. Bond, F., Copestake, A., Flickinger, D. 2002. Multiword expressions: A pain in the neck for NLP. In Proceedings of teh Third International Conference on Intelligent Text processing and Computatinal Linguistics (CICLING 2002), p. 1-15, Mexico City, Mexico. ACL.

Stevenson, S., Fazly, A., and North, R. (2004). Statistical measures of the semi-productivity of light verb constructions. In Proceedings of the ACL-04 Workshop on Multiword Expressions: Integrating Processing, p. 1–8.