Automatically learning semantic knowledge about multiword predicates

Article *in* Language Resources and Evaluation · October 2007

DOI: 10.1007/s10579-007-9017-9 · Source: DBLP

Afsaneh Fazly
VerticalScope, Inc.

Lang Resources & Evaluation (2007) 41:61–89
DOI 10.1007/s10579-007-9017-9

Automatically learning semantic knowledge about multiword predicates

Afsaneh Fazly · Suzanne Stevenson · Ryan North

Published online: 20 June 2007

Abstract Highly frequent and highly polysemous verbs, such as *give*, *take*, and *make*, pose a challenge to automatic lexical acquisition methods. These verbs widely participate in multiword predicates (such as light verb constructions, or LVCs), in which they contribute a broad range of figurative meanings that must be recognized. Here we focus on two properties that are key to the computational treatment of LVCs. First, we consider the degree of figurativeness of the semantic contribution of such a verb to the various LVCs it participates in. Second, we explore the patterns of acceptability of LVCs, and their productivity over semantically related combinations. To assess these properties, we develop statistical measures of figurativeness and acceptability that draw on linguistic properties of LVCs. We demonstrate that these corpus-based measures correlate well with human judgments of the relevant property. We also use the acceptability measure to estimate the degree to which a semantic class of nouns can productively form LVCs with a given verb. The linguistically-motivated measures outperform a standard measure for capturing the strength of collocation of these multiword expressions.

Keywords Lexical acquisition · Corpus-based statistical measures · Verb semantics · Multiword predicates · Light verb constructions

A. Fazly (✉) · S. Stevenson · R. North
Department of Computer Science, University of Toronto, 6 King's College Road,
Toronto, ON, Canada M5S 3G4
e-mail: afsaneh@cs.toronto.edu

S. Stevenson
e-mail: suzanne@cs.toronto.edu

R. North
e-mail: ryan@cs.toronto.edu

1 Highly polysemous verbs

People are presumed to have a marked cognitive priority for concrete, easily visualized entities over more abstract ones. Hence, abstract notions are often expressed in terms of more familiar, concrete things and situations, giving rise to a widespread use of figurative language (Lakoff and Johnson 1980; Johnson 1987; Numberg et al. 1994; Newman 1996). In particular, it is common across languages for certain verbs to easily undergo a range of figurative meaning extensions (Pauwels 2000; Newman and Rice 2004). In their literal uses, these highly polysemous verbs typically refer to states or acts that are central to human experience (e.g., *cut*, *give*, *put*, *take*), hence they are often referred to as *basic verbs*. In their extended uses, basic verbs often combine with various types of complements to form multiword predicates (MWPs) to which the verb contributes (possibly different) figurative meanings, as in 1(a–d):

1.
   (a) cut in line, cut sb. a break
   (b) give a speech, give a groan
   (c) put one's finger on sth., put sth. to rest
   (d) take a walk, take care

As with other multiword expressions, MWPs pose a challenge to computational lexicographers: specifically, how should they be encoded in a computational lexicon (Sag et al. 2002)? On the one hand, MWPs show the internal semantic cohesion attributed to lexical units. On the other hand, they retain some of their identity as phrases since they are formed by rules of verb phrase formation. Because of this, the constituents of an MWP may exhibit flexibility to some extent (for example, some MWPs may be passivized, as in *Every care was taken to ensure that the information was accurate*). Despite the superficial similarity between an MWP such as *give a groan*, and a verb phrase such as *give a present*, they should be distinguished from each other for several reasons.[1] For one, MWPs involve a certain degree of semantic idiosyncrasy. In addition, unlike verb phrases, the flexibility of MWPs is restricted, e.g., not all MWPs can undergo passivization. Furthermore, MWPs are *semi*-productive: new expressions can be formed from only limited combinations of syntactically and semantically similar component words (e.g., ?*give a gripe*, in contrast to *give a groan/cry/yell*). Note that explicitly storing MWPs in a lexicon is not a solution, since such an approach does not capture useful generalizations regarding the particular syntactic and semantic behaviour of MWPs.

In this article, we address some of the above-mentioned issues regarding the lexical representation of MWPs in a computational lexicon. More specifically, we focus on a common subclass of MWPs, called light verb constructions (LVCs). An LVC is formed around a highly polysemous basic verb, such as *give*, *make*, or *take*, as in *give a groan*, *make a decision*, and *take a walk*. The verb constituent of an

[1] Throughout this paper, we use the term *verb phrase* to refer to a syntactic combination of a verb and its arguments. We use the term *multiword predicate* (MWP) to refer to a verb phrase that has been lexicalized.

LVC—called a *light verb* because it is assumed to have lost its literal semantics to some degree (Butt 2003)–contributes a figurative meaning that is an extension of its literal semantics. The complement of the light verb in an LVC can be a verb, a noun, an adjective, or a prepositional phrase, which contributes to the overall predicative meaning of the LVC. As mentioned above for MWPs in general, an LVC is semantically idiosyncratic, i.e., it takes on a (predicative) meaning beyond the simple composition of the meanings contributed by its two constituents.

Light verb constructions are frequently and productively used in languages as diverse as English (Kerns 2002), French (Desbiens and Simon 2003), Spanish (Alba-Salas 2002), Persian (Karimi 1997), Urdu (Butt 2003), Chinese (Lin 2001), and Japanese (Miyamoto 2000). In this study, we focus on a broadly-documented subclass of English LVCs, in which the complement is an indefinite, non-referential predicative noun—i.e., a noun that has an argument structure. The noun constituent of such an LVC in its canonical form appears as a bare noun, or with an indefinite article, as shown in 2(a–c):

2.
(a) Priya *took a walk* along the beach.
(b) Allene *gave* her some *help*.
(c) The Minister has to *make a decision* about his resignation.

In such LVCs, the predicative noun is often morphologically related to a verb, and is the primary source of semantic predication (Wierzbicka 1982). The predicative nature of the noun constituent of the LVCs in 2(a–c) is illustrated by the fact that they contribute the verbs of the corresponding paraphrases in 3(a–c):

3.
(a) Priya *walked* along the beach.
(b) Allene *helped* her some.
(c) The Minister has to *decide* about his resignation.

Throughout this article, we will continue to use the term LVC to refer to this particular class of light verb construction. We investigate these LVCs because they are frequent across many different languages; in addition, they have interesting properties with respect to their syntactic and semantic flexibility, as well as their productivity.

We propose computational methods for the acquisition of lexical knowledge about LVCs. Specifically, we develop automatic techniques for separating LVCs from literal phrases, as well as for distinguishing different types of LVCs. We also provide automatic means for the organization of semantically-related LVCs in a computational lexicon. The next section expands on our proposal for tackling these problems.

2 Meaning extensions in LVCs: our proposal

In this study, we first tackle the problem of identifying LVCs by looking into the semantic contribution of the verb constituent. We propose automatic means for distinguishing expressions that have highly figurative uses of a basic verb—and hence are likely to be LVCs (e.g., *give a speech* and *give a groan*)—from those that have less figurative uses of the verb and hence are likely to be verb phrases (e.g., *give a present* and *give an idea*). We then set about the semi-productivity problem, by looking into patterns of acceptability of LVCs across semantic classes of complements. Section 2.1 further elaborates on the importance of the distinction among less to more figurative usages of basic verbs for the development of plausible natural language processing (NLP) systems. Section 2.2 expounds on the notion of semi-productivity of LVCs by describing the role of semantically similar classes of complements in refining the figurative meanings of a basic verb. (A preliminary version of this work where we first explained the relationship between the two problems was presented in Fazly et al. (2005).)

2.1 LVCs vs. verb phrases

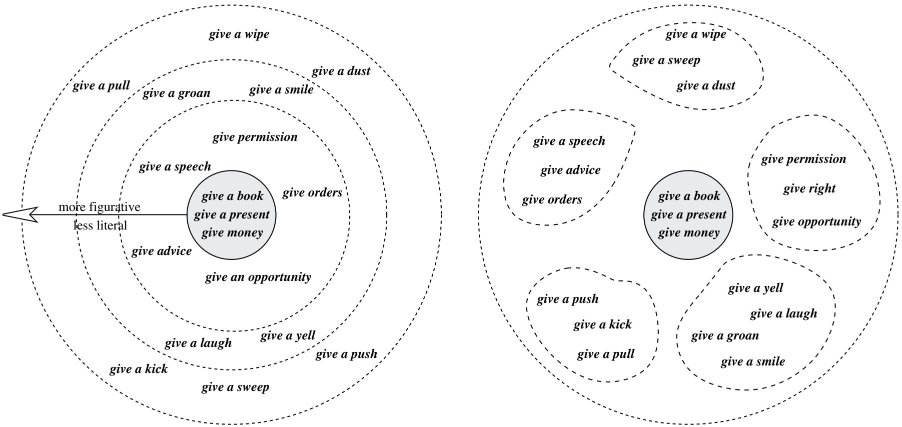
Basic verbs may contribute a literal meaning to the phrase they appear in. For example, in *give a present*, *give* refers to the ‘‘transfer of possession’’ of a physical object (*a present*) to a RECIPIENT. A more figurative use of the basic verb may also contribute its meaning to a verb phrase. For example, in *give an idea*, *give* indicates ‘‘transfer’’ of an abstract entity to a RECIPIENT. In an LVC, a basic verb contributes an even more figurative meaning, while the noun takes on more of the predicative burden of the MWP, as noted above. Moreover, different LVCs involve different levels of figurative usage of a verb: *give* in *give a speech* indicates ‘‘transfer’’ but not ‘‘possession’’, while in *give a groan*, the notions of ‘‘transfer’’ and ‘‘possession’’ are both diminished to a large extent, and a RECIPIENT is not possible.

While expressions with varying degrees of figurativeness of the verb are superficially similar, they exhibit different semantic and syntactic behaviour (as we will explain in detail in Sect. 3.1). For example, *give a present*, *give an idea*, *give a speech*, and *give a groan* all conform to the grammar rules of verb phrase formation. Nonetheless, they involve different meanings of the verb *give*; moreover, whereas *give a present* and *give an idea* are verb phrases, *give a speech* and *give a groan* are multiword predicates (LVCs). Hence, it is essential for an NLP application to distinguish different levels of figurative usages of a basic verb, and to treat them differently. As an example, Table 1 illustrates the importance of such a distinction for a machine translation system: an LVC such as *give a groan* should be translated as a single unit of meaning, whereas this is not necessarily the case for a verb phrase such as *give a present*. In the long run, finer-grained distinctions of figurativeness among LVCs (as in *give a speech* vs. *give a groan*) could also help in computational lexicography in determining the predicative properties of the resulting MWP.

To determine thelevel of figurativeness of a basic verb usage, we focus on two salient characteristics of figurative language, i.e., conventionalization and syntactic fixedness (Moon 1998). Section 3 expounds on such properties for LVCs, as well as on how these relate to the degree of figurativeness of the verb. We propose a statistical measure which incorporates these properties to place verb usages on a continuum of meaning from less to more figurative, as depicted in Fig. 1(a). Our

Table 1 Sentences with literal and figurative usages of *give*

English sentence (Intermediate semantics)	French translation
Azin gave *Sam a present.* (e1/*give* :agent (a1/‘‘*Azin*’’) :theme (p1/‘‘*present*’’) :recipient (s1/‘‘*Sam*’’))	*Azin a* donné *un cadeau à Sam.* *Azin gave a present to Sam.*
Azin gave a groan (e2/*give-a-groan* ≈ *groan* :agent (a1/‘‘*Azin*’’))	*Azin a* gémi. *Azin groaned.*



(a) The literal–figurative continuum

(b) A semantic grouping of figurative usages

Fig. 1 Two possible partitioning of the semantic space of the verb *give*

hypothesis is that most LVCs tend to appear towards the more figurative end of the figurativeness spectrum. In contrast, verb phrases are expected to appear close to the less figurative end of the continuum. A measure of figurativeness can thus be used to separate LVCs from similar-on-the-surface verb phrases to a large extent. Results of our evaluation, presented in Sect. 4, show that the continuum as determined by our statistical measure correlates well with the judgments of human experts.

2.2 Basic verbs and semantically similar complements

Another interesting property of basic verbs is that, in their figurative usages, they tend to have similar patterns of cooccurrence with semantically similar complements. Moreover, each similar group of complement nouns can be viewed as a possible meaning extension for the verb (Wierzbicka 1982; Sag et al. 2002;

Newman 1996). For example, in *give advice*, *give permission*, *give a speech*, etc., *give* contributes a notion of ‘‘abstract transfer’’, while in *give a cry*, *give a groan*, *give a moan*, etc., *give* contributes a notion of ‘‘emission’’.

There is much debate on whether such verbs should be represented in a lexicon as having one underspecified meaning, further determined by the context, or as a network of identifiable (related) subsenses (Pustejovsky 1995; Newman 1996). Under either view, it is important to determine the sets of complements that a particular subsense can occur with. In the long run, we would like to try to capture both the semantic generalizations and semantic restrictions that lead to a particular pattern of use, and explain why, e.g., one can *give a groan/cry/yell*, but not ?*give a gripe*. In this regard, it is essential to look at both the acceptability of individual expressions and the patterns of LVC acceptability across semantic classes of complements, to fully understand the semi-productivity of the LVC formation process.

A long-term goal of this work is to divide the space of figurative uses of a basic verb into semantically coherent segments, as shown in Fig. 1(b). Section 5 describes our hypothesis on the class-based nature of LVCs, i.e., their semi-productivity. At this point we cannot spell out the different figurative meanings of the verb associated with such classes of complements. Instead, we take a step forward by proposing a statistical measure for predicting the individual acceptability of a given combination of a basic verb and a noun as an LVC. Such a measure can also be used to predict the collective acceptability of a class of nouns in forming LVCs when combined with a given verb. Our evaluation as presented in Sect. 6 reveals the class-based tendency of verbs in forming LVCs; it also demonstrates the appropriateness of the proposed measure in predicting such behaviour.[2]

3 Figurativeness of basic verbs

3.1 Conventionalization and syntactic fixedness

It is widely observed that the underlying semantic properties of an expression largely determine its surface (lexical and syntactic) behaviour. As mentioned above, we are particularly interested in how the semantic properties of an expression using a basic verb influence its degree of conventionalization and syntactic flexibility. We hypothesize that expressions involving highly figurative usages of a basic verb (e.g., LVCs) have a greater tendency to be conventionalized—i.e., to become accepted as a semantic unit. Conventionalization also involves the distinction of a particular instantiation of a concept as favoured relative to others. For example, *make a decision* is highly favoured over ?*create a decision*. We then expect LVCs to show a high degree of association between the two component words (the light verb and the noun).

The syntactic behaviour of a multiword expression is also known to be influenced by its figurativeness. Linguists have looked at the issue of figurativeness from a

[2] Our first approach to address the class-based pattern of LVC formation is described in Stevenson et al. (2004). The material in Sects. 5–6 of this article is an updated presentation of that in Fazly et al. (2006).

number of different perspectives (Cruse 1986; Gibbs and Nayak 1989; Cacciari 1993; Nunberg 1994). Nonetheless, the evidence seems to converge on a relation between the degree of syntactic fixedness that an expression exhibits and its level of figurativeness. In particular, LVCs, which involve highly figurative uses of basic verbs, enforce restrictions on the syntactic freedom of their noun constituents (Kearns 2002).

For example, in some LVCs, the noun constituent has little or no syntactic freedom:

4.

(a) Azin *gave a groan* just now.
(b) ?? Azin *gave the groan* just now.
(c) ? Azin *gave* a couple of *groans* last night.
(d) ?? *A groan* was *given* by Azin just now.
(e) ?? *The groan* that Azin *gave* was very long.
(f) ?? *Which groan* did Azin *give*?
(g) * Azin *gave* his partner *a groan* just now.

In others, the noun may be introduced by a definite article, pluralized, passivized, relativized, or even *wh*-questioned, as in 5(b–f). Note, however, that the dative use, as in 5(g), is still questionable.[3]

5.

(a) Azin *gave a speech* to a few students.
(b) Azin *gave the speech* just now.
(c) Azin *gave* a couple of *speeches* last night.
(d) *A speech* was *given* by Azin just now.
(e) *The speech* that Azin *gave* was brilliant.
(f) *Which speech* did Azin *give*?
(g) * Azin *gave* the students *a speech* just now.

The degree to which an LVC has restricted syntactic freedom, as in these examples, is related to the degree to which the light verb has lost its literal semantics. Recall that *give* in expressions such as *give a groan* (cf. 4) is presumed to be a more figurative usage than *give* in expressions such as *give a speech* (cf. 5). By contrast, less figurative phrases, such as *give an idea* and *give a present*, which are verb phrases, exhibit virtually complete syntactic freedom, generally allowing all the constructions in these examples.

The linguistic explanation for this spectrum of behaviour relies on properties of the relation between the basic verb and the noun. When the verb is used more literally, the noun has an independent semantic identity as the complement of the verbal predicate; in this case, the noun exhibits syntactic freedom (Gibbs 1993). As the sentences in 5 above show, LVCs whose noun constituent can be treated, possibly figuratively, as the complement of the light verb also show syntactic

[3] It is important to note that these judgments are subject to individual differences. The point here is that the patterns specified by ‘‘?’’ and ‘‘??’’ (and to some extent those specified by ‘‘*’’) are less-preferred for the given expression. We do not claim here that these are impossible, rather that they are expected to be less natural, and less common, compared to the preferred pattern(s).

flexibility to a large extent. However, in highly figurative LVCs, as in 4, the relation between the noun and verb can no longer be construed as one of argument to a predicate, and the noun is then much more restricted. To summarize, the more figurative the meaning of the light verb in an LVC, the less ‘‘object-like’’ its noun constituent and the less flexibly the latter can be expressed.

These observations concerning conventionalization and syntactic fixedness motivate our proposed statistical measure described in the following subsection. This measure can be used to separate LVCs from similar-on-the-surface verb phrases, and also to distinguish different types of LVCs.

3.2 A statistical measure of figurativeness

We propose a statistical measure that quantifies the degree of figurativeness of the basic verb constituent of an expression by tapping into the notions of conventionalization and syntactic fixedness as described in Sect. 3.1. The measure assigns a score to an expression involving a verb (V) and a noun (N) by examining the degree of association between V and N, as well as their frequency in any of a set of relevant syntactic patterns, such as those in examples 4 and 5 above. The measure is defined as:

$$\textsc{Figness}\ (\mathrm{V,N}) \doteq \textsc{Assoc}\ (\mathrm{V;N}) + \textsc{Diff}\ (\textsc{Assoc}_{\mathrm{pos}}, \textsc{Assoc}_{\mathrm{neg}}) \tag{1}$$

whose components are explained in turn in the following paragraphs.

The first component, Assoc(V; N), measures the strength of the association between the verb and the complement noun. This is expected to reflect the degree to which these two components are bound together within a single unit of meaning, i.e., the degree to which the combination is conventionalized. This component is calculated using a standard information-theoretic measure, pointwise mutual information or PMI (Church et al. 1991):

$$\begin{aligned} \textsc{Assoc}\ (V;N) &\doteq \mathrm{PMI}(V;N) \\ &\doteq \log \frac{\Pr(V,N)}{\Pr(V)\Pr(N)} \\ &\approx \log \frac{n \times f(V,N)}{f(V,*)f(*,N)} \end{aligned} \tag{2}$$

where n is the total number of verb–object pairs in the corpus, $f(V,N)$ is the frequency of V and N cooccurring as a verb–object pair, $f(V,*)$ is the frequency of V with any object noun, and $f(*, N)$ is the frequency of N in the object position of any verb.

The second component of the Figness measure, Diff, estimates the degree of syntactic rigidity of the expression formed from V and N, by examining their association within different syntactic patterns. Assoc$_{pos}$ measures the strength of association between the expression and $\mathcal{PS}_{pos}$, the pattern set that includes syntactic patterns preferred by (more figurative) LVCs. Similarly, Assoc$_{neg}$ measures the strength of association between the expression and $\mathcal{PS}_{neg}$, representing patterns that are less preferred by LVCs.

In our current formulation, the two sets $\mathcal{PS}_{pos}$ and $\mathcal{PS}_{neg}$ contain syntactic patterns encoding the following attributes: the voice of the extracted expression (active or passive); the type of the determiner introducing N (definite or non-definite, the latter including the indefinite determiner *a/an* as well as no determiner); and the number of N (singular or plural). These attributes were identified (manually) by looking into the linguistic studies on the syntactic and semantic behaviour of LVCs (see Sect. 3.1). Note that this formulation is flexible and could be expanded to incorporate more attributes if necessary. As shown in Table 2, $\mathcal{PS}_{pos}$ consists of a single pattern with values for these attributes of active, non-definite, and singular; $\mathcal{PS}_{neg}$ has all the patterns with at least one of these attributes having the alternative value.

To measure the strength of association of an expression with a set of patterns, e.g., $\mathcal{PS}_{neg}$, we use the PMI between the expression and the set, as shown in Eq. 3 below. (Assoc$_{pos}$ is calculated similarly, by replacing $\mathcal{PS}_{neg}$ with $\mathcal{PS}_{pos}$.)

$$\begin{aligned} \textsc{Assoc}_{neg} &\doteq \mathrm{PMI}(V, N; \mathcal{PS}_{neg}) \\ &\doteq \log \frac{\Pr(V, N, \mathcal{PS}_{neg})}{\Pr(V, N)\Pr(\mathcal{PS}_{neg})} \\ &\approx \log \frac{n \times f(V, N, \mathcal{PS}_{neg})}{f(V, N, *) f(*, *, \mathcal{PS}_{neg})} \\ &= \log \frac{n \sum_{pt_j \in \mathcal{PS}_{neg}} f(V, N, pt_j)}{f(V, N, *) \sum_{pt_j \in \mathcal{PS}_{neg}} f(*, *, pt_j).} \end{aligned} \tag{3}$$

Our calculations of the PMI values use maximum likelihood estimates of the true probabilities. This results in PMI values with different levels of confidence (since different syntactic patterns have different frequencies of occurrence in text). Thus, directly comparing the two association strengths, Assoc$_{pos}$ and Assoc$_{neg}$, is subject to a certain degree of error. Following Lin (1999), we estimate the difference more accurately, by comparing the two confidence intervals surrounding the calculated association strength values, at a confidence level of 95%. Like Lin (1999) and Dunning (1993), we assume the estimates of the probabilities (e.g., as in Eq. 3 above) are normally distributed. We form confidence intervals around the estimates of $\Pr(V, N, \mathcal{PS}_{neg})$ and $\Pr(V, N, \mathcal{PS}_{pos})$, reflecting the possible ranges of the true probabilities. We use these ranges to form confidence intervals for the corresponding PMI values. We take the minimum distance between the two intervals as a

Table 2 Pattern sets used in measuring the syntactic rigidity of a given V + N combination, along with examples for each pattern

$\mathcal{PS}_{pos}$ = {“V_{active} det$_{nondef}$ N_{sing}” }	*give a groan, give permission*
$\mathcal{PS}_{neg}$ = {“V_{active} det$_{nondef}$ N_{plur}” ,	*?give groans*
“V_{active} det_{def} $N_{sing,plur}$”,	*?give the groan(s)*
“$det_{def,nondef}$ $N_{sing,plur}$ $V_{passive}$” }	*?a/the groan was given*

conservative estimate of the true difference, as depicted in Fig. 2, and shown in Eq. 4 below:

$$\textsc{Diff}(\textsc{Assoc}_{pos}, \textsc{Assoc}_{neg}) \doteq (\textsc{Assoc}_{pos} - \Delta\textsc{Assoc}_{pos}) - (\textsc{Assoc}_{neg} + \Delta\textsc{Assoc}_{neg}) \quad (4)$$

where $\Delta\textsc{Assoc}_{pos}$ $(\Delta\textsc{Assoc}_{neg})$ equals half of the interval surrounding $\textsc{Assoc}_{pos}$ $(\textsc{Assoc}_{neg})$. We expect that estimating the difference between the two PMI values in this way—i.e., using confidence intervals—lessens the effect of differences that are not statistically significant. Recall that low frequencies result in less reliable PMI values, hence they are expected to correspond to larger confidence intervals. Thus it is possible that the difference between two unreliable PMI values is high, but if we look at the difference between their corresponding intervals, we may find small differences or none at all.

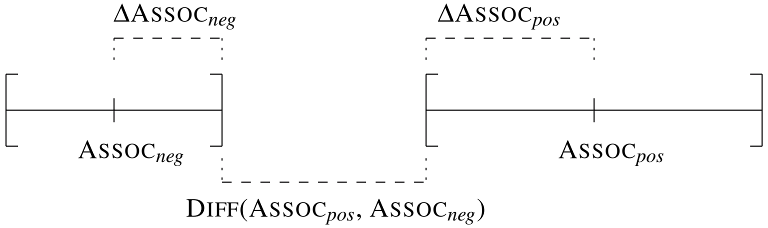
To summarize, the stronger the association between *V* and *N*, and the greater the rigidity of their use together (as measured by the difference between their association with positive and negative syntactic patterns), the more figurative the meaning of the verb, and the higher the score given by Figness(*V,N*).

4 Evaluation of the figurativeness measure

To determine how well our proposed measure, Figness, captures the degree of figurativeness of a basic verb usage, we compare the ratings it assigns over a list of test expressions with those assigned by human judges. Section 4.1 describes the selection of the experimental expressions, and the corpus we use to estimate frequency counts required by the measure. In Sect. 4.2, we elaborate on our approach in collecting consensus human ratings of figurativeness for the experimental expressions. Finally, Sect. 4.3 presents the evaluation results.

4.1 Materials and methods

Common basic verbs in English include *give*, *take*, *make*, *get*, *have*, and *do*, among others (Quirk et al. 1985; Brinton and Akimoto 1999). In the evaluation of our figurativeness measure, we focus on two of these, *give* and *take*, which are frequently and productively used in light verb constructions (Claridge 2000). These



$\Delta\textsc{Assoc}_{neg}$
$\Delta\textsc{Assoc}_{pos}$
$\textsc{Assoc}_{neg}$
$\textsc{Assoc}_{pos}$
$\textsc{Diff}(\textsc{Assoc}_{pos}, \textsc{Assoc}_{neg})$

Fig. 2 Approximating the difference between two PMI values as the minimum distance between the two corresponding confidence intervals

verbs are highly polysemous: the number of different WordNet senses for *give* and *take* are 44 and 42, respectively (Fellbaum 1998). They are also highly frequent: in the British National Corpus (BNC Reference Guide 2000), these verbs are among the transitive verbs with the highest frequency. These are important considerations for us since we need expressions that cover a wide range of possible meaning extensions of a particular verb.[4]

We use the British National Corpus, both as a source for extracting experimental expressions, and as a corpus for estimating the frequency counts required by the figurativeness measure.[5] We automatically parse the BNC using the Collins parser (Collins 1999), and further process it using TGrep2 (Rohde 2004) and NP-head extraction software based on heuristics from collins (1999).

Our experimental expressions are pairs of the form of a basic verb (*give* or *take*) plus a noun in direct object position. The list of expressions was randomly extracted from the BNC, subject to the constraint that each noun is morphologically related to a verb according to WordNet. The constraint on the noun ensures that our candidate list includes LVCs, which require a predicative noun. However, it also results in the exclusion of most literal combinations, which biases the set of experimental expressions to those that involve a figurative use of the verb. To perform a plausible evaluation, we need development and test data sets that cover a wide range of figurative and literal usages of the two verbs under study. To achieve a full spectrum of literal to figurative usages, we augmented the original list with literal expressions, such as *give a book* and *take a bowl*. Because these expressions were judged to be clearly literal by the authors, they were not subject to the procedure for rating figurativeness (described in the next subsection).[6] In addition to providing ratings on the original expressions, we also requested our judges to provide short paraphrases of each; in the final experiments, we only include those expressions for which a majority of the judges expressed the same sense.

The list of expressions is divided into a development set, DEV, and a test set, TST. In total, we have 150 development expressions and 70 test expressions, of which 114 involve the verb *give* and 106 involve *take*.

4.2 Human judgments of figurativeness

To provide human judgments on figurativeness, three native speakers of English with sufficient linguistic knowledge answered several yes/no questions about each of the experimental expressions. The questions were devised so that they indirectly

[4] We do not include *get*, *have*, and *do* because of their frequent use as auxiliaries; we did not include *make* in this experiment since, compared to *give* and *take*, it seemed to be more difficult to distinguish between literal and figurative usages of this verb. Our ongoing work focuses on expanding the set of verbs (see Fazly 2007).

[5] We also evaluated our figurativeness measure, FIGNESS, using web data as in our experiments for acceptability presented in Sect. 6. We found that since the estimation of FIGNESS requires more sophisticated linguistic knowledge, using a smaller but cleaner corpus (i.e., the parsed BNC) provides substantially better results.

[6] Note that since the initial sets were missing expressions that were rated as ''literal'' by the human annotators, the distributions of figurative and literal expressions in them were not representative of their ''true'' distribution.

Table 3 Questions asked of the human judges

	Questions for expressions with *give*	Answers
	As a result of the *event* expressed by the *expression*:	
I.	Does ‘‘Subj **transfer** a physical object to AP[a]’’?	y, n, m, ?[b]
II.	Does ‘‘Subj **transfer** something (non-physical) to AP’’?	y, n, m, ?
III.	Does ‘‘Subj **emit** something (non-physical)’’?	y, n, m, ?
	Questions for expressions with *take*	Answers
	As a result of the *event* expressed by the *expression*:	
I.	Does ‘‘Subj **take in** a physical object’’, or	
	‘‘AP **transfer** a physical object to Subj’’?	y, n, m, ?
II.	Does ‘‘Subj **move**’’?	y, n, m, ?
III.	Does ‘‘AP **transfer** something (non-physical) to Subj’’?	y, n, m, ?
IV.	Does ‘‘Subj **take in** or **adopt** something (non-physical)’’?	y, n, m, ?

[a] An Active Participant in the event, other than the Agent

[b] y: *yes*, n: *no*, m: *maybe*, ?: do not know

capture the degree to which aspects of the literal meaning of the verb is retained in the meaning of an expression. There are two sets of questions, one for each verb under study, as given in Table 3.

Each possible combination of answers to these questions is transformed to a numerical rating, ranging from 4 (largely literal) to 0 (highly figurative).[7] For example, the combination (**yes**, **no**, **no**) for an expression with *give* translates to a figurativeness rating of 4 (e.g., *give a dose*); the combination (**no**, **no**, **yes**) translates to a rating of 1 (e.g., *give a cry*); and the combination (**no**, **no**, **no**) to a rating of 0 (e.g., *give a go*). The complete list of all possible combinations of answers to these questions, as well as the numerical rating each combination translates to, are given in the Appendix. The numerical ratings are then averaged to form a consensus set to be used for final evaluation. Note that since we average the values, the consensus rating for an expression may be a non-integer value.

On the final set of experimental expressions (including both development and test expressions), the three sets of human ratings yield linearly weighted kappa values (Cohen 1968) of .34 and .70 for *give* and *take*, respectively.[8] (We use linearly weighted kappa since our ratings are ordered.)

The literal expressions added to the list of rated expressions are assigned a value of 5 (completely literal). Table 4 shows the distribution of the full lists of experimental expressions across three intervals of figurativeness level, ‘high’

[7] In order to maintain simplicity of both the questions and the process of translating their answers to numerical ratings, some fine-grained distinctions were lost. For example, under this scheme, *give an idea* and *give a speech* would receive the same rating. To distinguish such cases, we could also ask judges about the possibility of paraphrasing a given expression with a verb morphologically related to the noun constituent, which is a strong indicator of an LVC.

[8] We realize that a kappa value of .34 (for expressions with *give*) is low. In the future, we intend to resolve this problem, e.g., by providing the judges with more training, or more appropriate questions. The fact that expressions with *take*, which were annotated after those with *give*, have a much higher kappa reflects that more training may lead to more consistent annotations, and hence higher interannotator agreements.

Table 4 Distribution of Dev and Tst expressions according to human figurativeness ratings, along with examples

Verb	Figurativeness level	Dev	Tst	Example
give	'high'	20	10	*give a squeeze*
	'medium'	34	16	*give help*
	'low'	24	10	*give a dose*
	Total	78	36	
take	'high'	36	19	*take a shower*
	'medium'	9	5	*take a course*
	'low'	27	10	*take a bottle*
	Total	72	34	

(human ratings $\leq$1), 'medium' (1 < ratings < 3), and 'low' (ratings $\geq$3). The table also contains sample expressions for each figurativeness level. (Note that we do not perform any evaluation on these ''bucketized'' data sets. This is only to give the reader a feel for the distribution of the experimental expressions with respect to their figurativeness level.)

4.3 Figurativeness results

We use the Spearman rank correlation coefficient, r_s, to compare the ratings assigned by our figurativeness measure to the consensus human ratings. We also compare the ''goodness'' of Figness (as determined by the correlation tests) with that of an informed baseline, PMI_{LVC}.[9] PMI_{LVC} measures the strength of the association between the two constituents in particular syntactic configurations: i.e., $\text{PMI}_{\text{LVC}} = \text{PMI}(V; N, \mathcal{PS}_{pos})$. PMI_{LVC} is a baseline since it considers a given combination of a verb and a noun simply as a collocation. It is informed because it draws on linguistic properties of LVCs, by considering occurrences of the verb and noun in syntactic patterns preferred by LVCs—i.e., $\mathcal{PS}_{pos}$.

Table 5 displays the correlation scores between the human figurativeness ratings and those assigned by each statistical measure: PMI_{LVC} and Figness. Scores for the measure with the highest correlations are shown in boldface. In all cases the correlations are statistically significant ($p \ll .01$); we thus omit p values from the table. We report correlation scores not only on our test set (Tst), but also on development and test data combined (Dev+Tst) to get more data points and hence more reliable correlation scores. As noted above, there are two different types of experimental expressions: those with an indefinite determiner, e.g., *give a kick*, and those without a determiner, e.g., *give guidance*. Despite shared properties, the two types of expressions may differ with respect to syntactic flexibility, due to differing

[9] PMI is known to be unreliable when used with low frequency data. Nonetheless, in our preliminary experiments on development data, we found that PMI performed better than two other association measures, Dice and Log Likelihood. Other research has also shown that PMI performs better than or comparable to many other association measures (Inkpen 2003; Mohammad and Hirst 2006). We also alleviate the problem of sparse data by: (i) using large corpora, the 100-million-word BNC and the Web, and (ii) focusing on expressions with a minimum frequency of 5 (Dunning 1993).

Table 5 Correlations between human figurativeness ratings and the statistical measures

Verb	Data set	(size)	r_s	
			PMI$_{\text{LVC}}$	Figness
give	Tst	(36)	.62	**.66**
	Dev+Tst	(114)	.68	**.70**
	Dev+Tst/a	(79)	.68	**.77**
take	Tst	(34)	.51	**.57**
	Dev+Tst	(106)	.52	**.56**
	Dev+Tst/a	(68)	.63	**.68**

semantic properties of the noun complements in the two cases.[10] We thus calculate correlation scores for expressions with the indefinite determiner only; to have a sufficient number of data points, we use expressions from both development and test data (Dev+Tst/a).

Our proposed measure, Figness, shows notable improvements over the baseline on all data sets—Tst, Dev+Tst, and Dev+Tst/a. The results also show that Figness has higher correlation scores (with large improvements over the baseline) when tested on expressions with an indefinite determiner only, i.e., Dev+Tst/a. (Note that the correlation scores are highly significant—very small p values—on both data sets, Dev+Tst and Dev+Tst/a.)

These results confirm our hypothesis that the degree of figurativeness of a basic verb usage can be determined by looking into the conventionalization and syntactic fixedness of the expression containing the verb. Recall that LVCs tend to appear towards the more figurative end of the literal–figurative continuum. By setting a threshold, we can thus use our figurativeness measure to identify LVCs, i.e., to separate them from similar-on-the-surface verb phrases. Moreover, the measure can be used to distinguish between semantically (and syntactically) different LVCs, such as *give a speech* and *give a groan*. Given the differing predicative properties of such expressions (as discussed in Sect. 2.1), this distinction could be useful in (semi-) automatically determining their argument structures.

5 LVC acceptability across semantic classes

5.1 Class-based productivity

In this aspect of our work, we narrow our focus onto a subclass of LVCs that have a predicative noun constituent identical (in stem form) to a verb. We also consider only those expressions in which the noun is typically preceded by an indefinite determiner, e.g., *take a walk* and *give a smile*. These LVCs are of interest because they are very common, and moreover, their productivity appears to be patterned

[10] The use of an indefinite determiner or no determiner in an LVC relates to semantic characteristics such as the aspectual properties of the state or event expressed by the predicative noun (Wierzbicka 1982). The detailed discussion of their differences, however, is outside the scope of this study.

(Wierzbicka 1982; Kearns 2002). For example, one can *take a walk*, *take a stroll*, and *take a run*, but it is less natural to ?*take a groan*, ?*take a smile*, or ?*take a wink*. These patterns of semi-productivity depend on both the semantics of the complement as well as on the light verb itself; for example, in contrast to *take*, we observe ?*give a walk*, ?*give a stroll*, ?*give a run*, but *give a groan*, *give a smile*, *give a wink*.

Our hypothesis is that semantically similar LVCs—i.e., those formed from a light verb plus any of a set of semantically similar nouns—distinguish a figurative subsense of the verb. In the long run, if this is true, it could be exploited by using class information to extend our knowledge of observed LVCs and their likely meaning to unseen LVCs (cf. such an approach to verb-particle constructions by Villavicencio (2003, 2005)).

As a first step to achieving this long-term goal, we must devise an acceptability measure which determines, for a given verb, which nouns it successfully combines with to form an LVC. We can then examine whether this measure exhibits differing behaviour across semantic classes of potential complements, matching the behaviour as predicted by human judgments.

5.2 A statistical measure of acceptability

We propose a measure that captures the likelihood of a basic verb (V) and a noun (N) forming an acceptable LVC. We define our acceptability measure to be the joint probability of the V, the N, and these elements being used in an LVC:

$$\begin{aligned}&\textsc{Accept}_{\textsc{LVC}}(V,N)\\&\doteq \Pr(V,N,LVC)\\&=\Pr(N)\Pr(LVC|N)\Pr(V|N,LVC)\end{aligned} \tag{5}$$

We discuss each of the three factors in the following paragraphs.

The first factor, Pr(*N*), reflects the linguistic observation that higher frequency nouns are more likely to be used as LVC complements (Wierzbicka 1982). We estimate this factor by *f*(*N*)/*n*, where *n* is the number of words in the corpus.

The probability that a given V and N form an acceptable LVC further depends on how likely it is that the N combines with *any* basic verb to form an LVC ($\Pr(LVC|N)$). This is expected to be greater for true predicative nouns, since an argument structure must be contributed from the noun in the LVC. The frequency with which a noun forms LVCs is estimated as the number of times we observe it in the prototypical ‘‘V a/an N’’ pattern across basic verbs. (Note that such counts are an overestimate, since some of these occurrences may be literal uses of the verb.) Since these counts consider the noun only in the context of an indefinite determiner, we normalize over counts of ‘‘a/an N’’ (noted as *aN*):

$$\Pr(LVC|N) \approx \frac{\sum_{i=1}^{v} f(V_i, aN)}{f(aN)} \tag{6}$$

where *v* is the number of basic verbs considered in this study.

The third factor, $\Pr(V|N, LVC)$, reflects that different basic verbs have varying degrees of acceptability when used with a given noun in an LVC. We similarly estimate this factor with counts of the given V and N in the typical LVC pattern: $f(V,aN)/f(aN)$.

Combining the estimates of the three factors yields:

$$\textsc{Accept}_{\text{LVC}}(V,N) \doteq \frac{f(N)}{n} \times \frac{\sum_{i=1}^{v} f(V_i, aN)}{f(aN)} \times \frac{f(V,aN)}{f(aN)}. \tag{7}$$

6 Evaluation of the acceptability measure

To determine whether our measure, $\textsc{Accept}_{\text{LVC}}$, appropriately captures LVC acceptability, we compare its ratings to human judgments. We have two goals in evaluating $\textsc{Accept}_{\text{LVC}}$: one is to demonstrate that the measure is indeed indicative of the level of acceptability of an individual LVC, and the other is to explore whether it helps to indicate class-based patterns of LVC formation.

Section 6.1 explains our approach in selecting experimental expressions, and the corpus we use to approximate frequency counts required by our acceptability measure. In Sect. 6.2, we describe our collection of a consensus human rating of LVC acceptability on the experimental expressions. Last, in Sect. 6.3, we present the results of comparing the two sets of ratings: those given by our measure, and those assigned by the human judges.

6.1 Materials and methods

6.1.1 Experimental expressions

In the evaluation of our acceptability measure, we include three common English basic verbs, *take*, *give*, and *make*. *Take* and *give* have nearly opposite, but highly related, semantics, while *make* differs from both. Also, the line between light and literal uses of *make* appears to be less clear.[11] We expect then that *make* will show contrasting behaviour. Experimental expressions are formed by combining the three verbs with predicative nouns from (i) selected semantic verb classes of Levin (1993) (henceforth, Levin); or (ii) generated WordNet classes (Fellbaum 1998). In each case, some classes are used as development data, and some classes as test data.

It may seem odd to use a verb classification as a source of noun complements. However, recall that an important property of the type of LVCs we are considering is that the complement is a predicative noun (one with an argument structure), and is identical in stem form to a verb. The verb classes of Levin (1993), defined on the basis of argument structure similarity, therefore provide natural similarity sets to

[11] This was an observation made by the judges who later rated the acceptability of the experimental expressions as LVCs. The extent to which this observation holds for *make* or for other verbs in general is outside the scope of this study.

consider. As long as we only use verbs identical in form to a noun, we are assured that such complements are predicative nouns.

Although the use of Levin verb classes has linguistic motivation, it may be that semantic classes which also incorporate nominal similarity are more appropriate for this task (Newman 1996). Therefore, we also use semantic classes generated from both the noun and the verb hierarchies of WordNet 2.0. In determining these WordNet-derived classes, it is important that they are comparable to each of our Levin classes, so that we can relate performance of our acceptability measure across the two classifications. We achieve this by generating each WordNet-derived class as a set of words that are semantically similar to a representative word from a corresponding Levin class.

In the following paragraphs, we explain our criteria for the selection of experimental classes from Levin, and our algorithm for generating corresponding classes using WordNet.

Selection of Levin classes: Three Levin classes are used as development data, and four classes as (unseen) test data. The development classes are *Wipe* Verbs (#10.4.1), *Throw* Verbs (#17.1), and *Run* Verbs (#51.3.2). The test classes include *Hit* and *Swat* Verbs (#18.1,2), *Peer* Verbs (#30.3), Sound Emission Verbs (#43.2), and a subclass of Verbs of Motion (#51.4.2). The classes are chosen such that they reflect a range of LVC productivity in combination with the three verbs under study. Recall that we only include verbs that are identical in stem form to a noun. For classes with more than 35 verbs (30 for development classes), we select a random subset of that size, due to the manual effort needed for their annotation.

Generation of WordNet classes: For each Levin class, we first determine the general pattern of LVC acceptability with the three verbs under study. As described in Sect. 6.2 below, human ratings of expressions as acceptable LVCs are put into buckets of ‘poor’, ‘fair’, and ‘good’. We then determine the predominant bucket for each class and verb, and manually select a representative seed from each class that most closely matches the typical ratings across the three verbs (see Table 6). For most Levin classes, there was only one such noun; if there was more than one, we arbitrarily picked one as the seed. For each seed, we automatically examine both the noun and verb hypernym hierarchies of WordNet, and select all words which have a parent in common with the seed. We filter from this set those words which do not appear in both hierarchies, thereby excluding items which are not nouns identical in

Table 6 Seed words selected according to acceptability trends identified for each Levin test class and verb

Levin class	Acceptability trend			Seed word
	take	*give*	*make*	
Hit and *Swat* Verbs	fair	good	fair	*knock*
Peer Verbs	fair	fair	poor	*check*
Verbs of sound emission	poor	good	fair	*ring*
Verbs of motion using a vehicle[a]	good	fair	poor	*sail*

[a] The subset that are verbs which are not vehicle names

form to a verb. (In contrast to the Levin expressions, we also filter rare predicative nouns, whose frequency as a verb in the British National Corpus is less than 50.) A random selection of 35 of the remaining words forms a WordNet class, which we refer to by ''WN-'' plus the seed verb (e.g., WN-*knock*).

Our final experimental data consists of 195 nouns in the development set (90 from Levin classes and 105 from WordNet classes), and 238 nouns in the test set (98 from Levin classes and 140 from WordNet classes). These nouns are combined with each of the three verbs to yield 585 development expressions, and 714 test expressions, all of the form ''*give/take/make a/an* N''.

6.1.2 Corpus and data extraction

LVCs of the type we consider are, as a class, very frequent. Interestingly, however, individual expressions may be highly acceptable but not attested in any particular corpus. We decided therefore to use the web—the subsection indexed by Google—to estimate frequency counts required by our acceptability measure. Each count is calculated via an exact-phrase search; the number of hits is used as the frequency of the string searched for. Counts including verbs are collapsed across three tenses of the verb: base, present, and simple past. The size of the corpus, *n*, is estimated at 5.6 billion, the number of hits returned in a search for ''*the*''. Note that frequency counts for candidate expressions are likely underestimated, as a phrase may occur more than once in a single web page; we make the simplifying assumption that this affects all counts similarly.[12] Such frequency estimates have been successfully used in many NLP applications (e.g., Turney 2001; Villavicencio 2005). Moreover, they have been shown to correlate highly with frequency counts from a balanced corpus (Keller and Lapata 2003).

Most LVCs allow their noun constituent to be modified, as in *take a long walk*. To capture such cases, we used the '*' wildcard (as in ''*take a * walk*''), which at the time we performed our Google searches matched exactly one word. Moreover, many LVCs using the light verb *give* frequently appear in the dative form, and some of these can only appear in this form. For example, one can *give NP a try*, but typically not ?*give a try to NP*. To address this, we perform individual searches for each of a set of 56 common object pronouns—e.g., *them*, *each*—intervening between the verb and the noun. Note that this only captures a subset of dative uses since we only consider cases where the NP is a pronoun. The final estimated frequency of an expression is the sum over the approximated frequencies of its bare, modified, and dative forms.

6.2 Human judgments of acceptability

To provide human judgments of acceptability, two expert native speakers of English rated the acceptability of each candidate ''V a/an N'' expression as an LVC. A

[12] This is clearly not the case for the estimate of the corpus size, since ''*the*'' likely occurs frequently within each page. However, in our formulas, this value appears as a constant, thus all scores are equally affected.

candidate was not rated highly if it was an acceptable literal or idiomatic expression, but not an LVC. For example, even though *give a sink*, *take a fall*, and *make a face* are all acceptable expressions, only *take a fall* should receive a high rating as an acceptable LVC: *take a fall* roughly means *fall*, whereas *give a sink* is acceptable only as a literal expression, and *make a face* is acceptable only as an idiom. The ratings range from 1 (unacceptable) to 4 (completely natural), by 0.5 increments.

On Levin test expressions, the two sets of ratings yield linearly weighted kappa values of .72, .39, and .44, for *take*, *give*, and *make*, respectively, and .53 overall. Wide differences in ratings typically arose when one rater missed a possible meaning for an expression; these were corrected in a reconciliation process. Discussion of disagreements when rating Levin expressions led to more consistency in ratings of WordNet expressions, which yield linearly weighted kappa values of .79, .66, and .69, for *take*, *give*, and *make*, respectively, and .71 overall. These ratings were also reconciled to within one point difference. For each set of expressions, we then average the two ratings to form a single consensus rating. We also place the consensus ratings in buckets of ‘poor’ (range [1–2)), ‘fair’ (range [2–3)), and ‘good’ (range 3 and higher) for coarser-grained comparison.

6.3 Acceptability results

The following subsections describe different aspects of the evaluation of our acceptability measure, $\textsc{Accept}_{\mathrm{LVC}}$. We use the Spearman rank correlation coefficient, r_s, to compare the ratings provided by $\textsc{Accept}_{\mathrm{LVC}}$ to the human acceptability judgments (Sect. 6.3.1). Linearly weighted observed agreement, p_o, is used to examine the agreement between the statistical measure and humans at the coarser level of the acceptability buckets (Sect. 6.3.2). The acceptability buckets are further used to determine the appropriateness of our measure for predicting the productivity of a class with respect to LVC formation (Sect. 6.3.3). In each case, we compare the ‘‘goodness’’ of $\textsc{Accept}_{\mathrm{LVC}}$ (as determined by r_s or p_o) with that of a baseline. We use the same baseline as in the evaluation of the figurativeness measure, i.e., $\mathrm{PMI}_{\mathrm{LVC}}$. Higher values of $\mathrm{PMI}_{\mathrm{LVC}}$ reveal a greater degree of association between the verb and the noun, which can be interpreted as an indication of LVC acceptability. In the presentation of our results, we focus on the analysis on unseen test data; trends are similar on development data.

6.3.1 Correlation between $\textsc{Accept}_{\mathrm{LVC}}$ and human ratings

We perform separate correlation tests between the human judgments and the two measures (our proposed acceptability measure, and the informed baseline) over each of the three verbs in combination with each of the four test classes within the two classifications, Levin and WordNet. That is, we perform a total of 24 correlation tests for each measure—12 for each classification. In Fig. 3, we show the results graphically, so that patterns are easier to see; numerical r_s values are available in the Appendix. Each rectangle in Fig. 3 represents the result of the correlation test on a single test class. Values of r_s which are not significant are shown as the lightest rectangles; significant values from .30 to over .70 (by deciles) are shown as

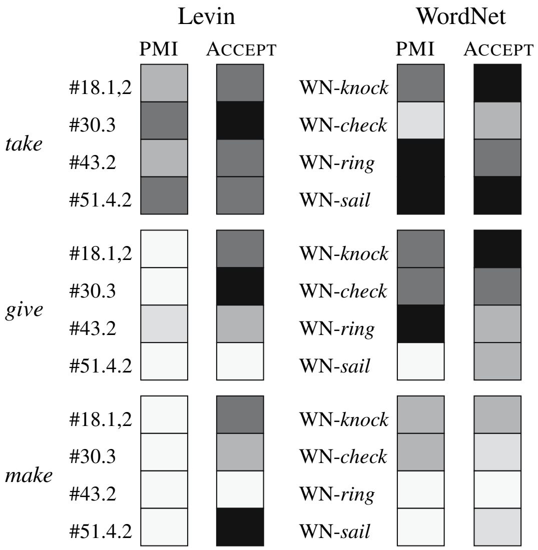


Fig. 3 Greyscale representation of the correlation scores (r_s) for Accept$_{\text{LVC}}$ and PMI$_{\text{LVC}}$, across the 3 verbs and the 4 Levin and WordNet test classes. Levin classes are specified by number; WordNet classes are referred to by ''WN-'' plus the seed verb

increasingly darker rectangles. We used a significance cut-off of $p < .07$, since some tests achieved reasonably good correlations that were marginally significant at this level. In what follows, we discuss the results in terms of the statistical measures, the three verbs, and the two classifications.

The Accept$_{\text{LVC}}$ measure is more consistent than the baseline, performing best overall and achieving good correlations in most cases. The PMI$_{\text{LVC}}$ measure does surprisingly well, as a simple measure of collocation; it even performs comparably to Accept$_{\text{LVC}}$ on the WordNet classes.

Examining the patterns in Fig. 3 by verb, we see that *take* achieves the best correlations on both Levin and WordNet expressions, followed by *give*, then *make*, which has particularly poor results. The poorer correlations with *give* and *make* may be partly due to the difficulty in rating them; note the lower interannotator agreement on expressions involving *give* and *make* (see Sect. 6.2).

Now looking at the patterns across the two semantic classifications, we note that the performance of Accept$_{\text{LVC}}$ is overall comparable across the two, while PMI$_{\text{LVC}}$ shows a marked improvement with the WordNet classes. A closer look at the WordNet and Levin expressions reveals an interesting difference between the two: the average frequency of nouns in the WordNet classes is significantly higher than that of nouns in the corresponding Levin classes (26M vs. 8M, respectively). Accept$_{\text{LVC}}$ appears to be less sensitive to frequency factors than the simple PMI-based measure.

The effect of semantic classification on the measures also interacts with the specific verb being used. We see that PMI$_{\text{LVC}}$ is particularly inferior on Levin classes with *give* and *make*. In addition to the possible problem with interannotator

Table 7 Weighted observed agreement (p_o) for statistical measures applied to Levin and WordNet test expressions

Verb	Class type	Chance agreement	p_o	
			PMI$_{\text{LVC}}$	Accept$_{\text{LVC}}$
take	Levin	.78	.77	**.85**
	WordNet	.81	**.88**	**.86**
give	Levin	.80	.59	.77
	WordNet	.75	.74	**.80**
make	Levin	.87	.81	.82
	WordNet	.85	.80	.74

agreement mentioned above, it seems that expressions with *give* and *make* are less treatable as straightforward collocations, especially with lower frequency items.

6.3.2 Agreement between Accept$_{LVC}$ and human ratings

We now inspect the performance of the Accept$_{\text{LVC}}$ measure when the coarser level of acceptability—'poor', 'fair', or 'good'—is considered. For both Accept$_{\text{LVC}}$ and PMI$_{\text{LVC}}$, we divide the continuous ratings into the discrete buckets, by setting thresholds. Thresholds are chosen such that the bucket sizes (i.e., number of expressions in each bucket) for development data match as closely as possible those of the human ratings. These thresholds are then used in dividing the test expressions into the buckets. We then calculate the (observed) agreement between each measure and the human judges in assigning the test expressions to the buckets. The agreement, p_o, is estimated as the (linearly weighted) proportion of the items that are assigned to the same bucket.[13] For comparison, we also calculate the uninformed baseline given by chance agreement. For most pairs of verb and class, our chance baseline considers all items to be labelled 'poor', since that is the largest bucket size in the human ratings. The one exception is *take* with the Levin class of Verbs of Motion, in which the baseline assignment is 'good'.

Observed agreement scores are shown in Table 7; values of p_o above the chance baseline are in boldface. On Levin and WordNet expressions with *take* and *give*, Accept$_{\text{LVC}}$ mostly outperforms both the chance baseline and the informed baseline, PMI$_{\text{LVC}}$. On expressions involving *make*, however, neither Accept$_{\text{LVC}}$ nor PMI$_{\text{LVC}}$ perform better than the chance baseline, reinforcing our initial hypothesis that *make* has differing properties from the other two light verbs. This coarser-grained level of acceptability shows a similar pattern across Levin and WordNet classes to that revealed by the correlation scores. Here again, PMI$_{\text{LVC}}$ does better on WordNet classes, and Accept$_{\text{LVC}}$ performs more consistently across the two.

[13] Because our ratings are skewed toward low values, slight changes in observed agreement cause large swings in kappa values (the ''paradox'' of low kappa scores with high observed agreement; Feinstein and Cicchetti 1990). Since we are concerned with comparison to a baseline, observed agreement better reveals the patterns.

We look next at the productivity of these classes with the different verbs. Because accurate assessment of class productivity depends on a measure having a reasonable level of agreement with the human ratings, we exclude *make* from the consideration of productivity.

6.3.3 Predicting class productivity

Our probabilistic measure achieves good performance in determining the level of acceptability of an individual ‘‘V a/an N’’ combination as an LVC. Still, a further goal is to devise statistical indicators of the productivity of LVC formation over a class of semantically related nouns with a given light verb. This is required for the adequate treatment of LVCs in a computational system. Knowledge about the collective tendency of a semantic class in forming LVCs with a given verb can be extended to unattested, semantically similar nouns. For example, if the class of sound emission nouns (e.g., *groan*, *moan*) is known to productively form LVCs with *give*, the assessed acceptability of an unseen or low frequency LVC, such as *give a rasp*, should be promoted.

The productivity of a class with respect to a light verb is indicated by the proportion of nouns in that class that form acceptable LVCs with the verb. We consider an acceptable LVC to be one that is either ‘fair’ or ‘good’ according to human judgments. Thus, to investigate the appropriateness of a measure as an indicator of class productivity, we compare (for each combination of verb and semantic class of nouns) the measure’s proportion of nouns in the ‘fair’ and ‘good’ buckets with that of the human judgments. The better the match between the two proportions, the better the measure at assessing class productivity.

Using the bucket thresholds described above, we determine the productivity level of each combination of verb (*take* or *give*) and semantic class (Levin or WordNet classes). As an example, Table 8 presents the productivity of each WordNet test class for *take*, as determined by human judges and by each of the statistical measures. The variability across the classes according to the human judgments clearly shows that LVC acceptability is a class-based effect.

We quantify the goodness of each measure for predicting productivity by calculating the divergence of its assessed productivity levels from those of the human judges, across the experimental classes and verbs. The divergence is measured as the sum of squared errors (SSE) between the two sets of numbers, averaged over the verbs and classes. Table 9 shows the average SSE values for each

Table 8 Proportion of expressions rated ‘fair’ or ‘good’ for *take* and each WordNet test class, as determined by human ratings and the statistical measures

Class	Human	$\mathrm{PMI}_{\mathrm{LVC}}$	$\textsc{Accept}_{\mathrm{LVC}}$
WN-*knock*	.26	.40	.26
WN-*check*	.14	.09	.26
WN-*ring*	.09	.17	.23
WN-*sail*	.46	.40	.37

Table 9 Divergence between productivity assessments of the statistical measures and human judgments, expressed as the sum of squared errors (SSE), averaged across Levin or WordNet classes

Class type	PMI$_{\mathrm{LVC}}$	Accept$_{\mathrm{LVC}}$
Levin	.220	**.093**
WordNet	.057	**.035**

measure and each classification, Levin or WordNet. The lowest SSE (best match to human judgments) is shown in bold. For both classifications, Accept$_{\mathrm{LVC}}$ gives the closest predictions, i.e., the lowest SSEs. Notably, here we see overall better performance with WordNet than with Levin classes for both measures.

6.3.4 Summary of results

Our results indicate that Accept$_{\mathrm{LVC}}$ is a good measure of acceptability at both the fine- and coarse-grained levels, according to the observed r_s and p_o values, respectively. Accept$_{\mathrm{LVC}}$ also accurately predicts the level of productivity of a semantic class of complements with a light verb, according to the reported SSE values.

In general, the classes generated from WordNet seem most useful in our tasks, especially when considering generalization of knowledge of possible LVC complements. Whether this is due to their higher item frequency noted above, or to the fact that our generation process draws on both nominal and verbal similarity, is an issue for future explanation.

7 Discussion and concluding remarks

Recently there has been a growing awareness of the need for the appropriate handling of multiword expressions (MWEs) (Sag et al. 2002). Much of the previous research on MWEs has concentrated on their automatic extraction (Melamed 1997; Baldwin and Villavicencia 2002; Seratan et al. 2003). Moreover, research focusing on the acquisition of deeper knowledge about MWEs has mainly covered certain classes, such as verb-particle constructions (McCarthy et al. 2003; Bannard et al. 2003; Baldwin et al. 2003). Our work focuses on the acquisition of syntactic and semantic knowledge about MWEs involving basic verbs, which are both highly frequent and highly polysemous. Specifically, we investigate the use of basic verbs in light verb constructions (LVCs), a class of cross-linguistically frequent MWEs that has been granted relatively little attention within the computational linguistics community (though see Grefenstette and Teufel 1995; Dras and Johnson 1996; Krenn and Evert 2001; Moirón 2004).

Previous work on MWE semantics has concentrated on computational methods for determining the degree to which the components of an MWE contribute compositionally to the semantics of the full expression. Most research in this vein examines the distributional similarity between an expression and its individual constituents (McCarthy et al. 2003; Bannard et al. 2003; Baldwin et al. 2003). Such techniques depend on a potential contrast between a constituent within an MWE and on its own. This approach is inappropriate for basic verbs, whose frequent use within LVCs and other figurative expressions makes it difficult to determine usages outside LVCs. Krenn and Evert (2001) attempt to distinguish light (support) verb constructions from expressions with different levels of compositionality, i.e., idioms and literal phrases. In contrast to our work, they treat LVCs purely as (conventionalized) collocations, and use frequency and several association measures, such as PMI, for the task. Lin (1999) and Wermter and Hahn (2005) look into another property of MWEs that is inversely correlated with their compositionality, i.e., their lexical fixedness. Venkatapathy and Joshi (2005) combine aspects of the above-mentioned work by incorporating measures of lexical fixedness, collocation, and distributional similarity into a classifier for determining the level of compositionality of verb–noun combinations. We instead relate the semantic properties of MWEs to their syntactic, and not just lexical, behaviour.

Our work also differs from previous studies in considering a different aspect of semantic contribution of the constituents of an MWE. Specifically, we are concerned with the degree to which the semantic contribution of the verb constituent of an LVC lies along the continuum from less to more figurative. We combine evidence from two sources: the degree of conventionalization of LVCs, and the extent to which they exhibit syntactic fixedness, the latter of which is a salient but mostly overlooked characteristic of LVCs. By examining the degree to which a basic verb usage is syntactically ‘‘similar’’ to the prototypical LVC, we provide an inverse indicator of the degree to which the verb retains aspects of its literal semantics. In particular, the more syntactically fixed the target expression, the more figurative the use of the basic verb. Our proposed figurativeness measure, Figness, correlates well with the literal–figurative spectrum represented in human judgments, supporting such an approach.

Work indicating acceptability of MWEs is largely limited to collocational analysis using simple frequency-based measures (Dras and Johnson 1996; Lin 1999; Stevenson et al. 2004). We instead use a probability formula that enables flexible integration of linguistic properties of LVCs. In a similar vein, Grefenstette and Teufel (1995) use LVC-specific knowledge to guide the extraction of relevant evidence about the best choice of light (support) verb for a given predicative noun. Their study, however, lacks a comprehensive evaluation and provides only subjective assessment of the results. Here, we show that our $\text{Accept}_{\text{LVC}}$ measure yields good correlations with human acceptability judgments.

A long-term goal of this study is to determine fine-grained distinctions among the figurative usages of a basic verb. In most cases, such distinctions appear to relate to the semantic properties of the complement that combines with a light verb to form an LVC. In other words, not only does a light verb tend to combine with semantically similar complements, it tends to contribute a similar figurative meaning to the resulting LVC. Semantic class knowledge thus may enable us to further refine the semantic space of a verb by elucidating its relation with complements of different semantic types.

Wanner (2004) attempts to classify verb–noun combinations into predefined groups, each corresponding to a particular semantic relation between the two constituents. His approach, however, requires manually-labelled training data. Uchiyama et al. (2005) propose a statistical approach to classifying Japanese LVCs (of the form verb–verb). They acknowledge the importance of the semantic properties of the complement for this task; however, they do not explicitly use such information. Moreover, the classes are broad, identified based on possible semantic contributions of the light verb (spatial, aspectual, or adverbial), and hence do not account for fine-grained distinctions among LVCs. Villavicencio (2005) uses class-based knowledge to extend a lexicon of verb-particle constructions (VPCs), but assumes that an unobserved VPC is not acceptable. We instead believe that more robust application of class-based knowledge can be achieved with a better estimate of the acceptability level of various expressions. Our $\text{Accept}_{\text{LVC}}$ measure also reflects patterns across semantic classes of complement nouns, similar to those reflected in the human judgments.

The work presented here is the first we are aware of that aims not only at distinguishing literal and figurative usages of a certain class of highly polysemous verbs, but also at refining the figurative senses. Our work ties together the two issues of figurativeness of basic verbs and LVC acceptability, and relates them to the notion of class-based meaning extensions of these polysemous verbs. Nonetheless, there are limitations that need to be addressed. In the future, we need to provide more and cleaner annotated expressions to conduct a more comprehensive evaluation of the suggested techniques. Moreover, while we have focused here on light verb constructions, we believe that similar techniques can be useful in dealing with related types of MWEs (as shown by Fazly and Stevenson 2006). Our ongoing work focuses on expanding the set of basic verbs, as well as on broadening the scope of the study to multiword predicates (MWPs) other than LVCs. Currently, we are also looking at other characteristics of figurative multiword expressions, in addition to syntactic fixedness, in order to recognize different classes of MWPs (see Fazly 2007).

Acknowledgements We thank Anne-Marie Brousseau, for the enlightening discussions regarding the human judgments on figurativeness; Eric Joanis, for providing us with NP-head extraction software; and our judges, who made the evaluation of our ideas possible. We are also grateful of the Natural Sciences and Engineering Research Council of Canada (NSERC), the Ontario Graduate Scholarship program (OGS), and the University of Toronto for the financial support.

Appendix

This appendix contains information on the procedure for interpreting the human judgments for the development and test expressions used in the experiments of Sect. 4.3. It also contains the numerical r_s values of the results presented in Sect. 6.3.1.

Tables 10 and 11 show how the judges' answers to the questions (given in Table 3 on page 13) are translated into numerical ratings ranging from 0 to 4. Higher numerical ratings express higher degrees of literalness, hence lower degrees of figurativeness. Expressions for which no numerical rating is listed in the tables are removed from the final set of experimental expressions. These were expressions that

Table 10 Interpretation of answers to the questions for expressions with *give*

Q(I)	Q(II)	Q(III)	Rating
yes	no	no	4
yes/maybe	yes/maybe	no	3
no	yes	no	2
no	no/maybe	yes	1
no	no	no	0

Table 11 Interpretation of answers to the questions for expressions with *take*

Q(I)	Q(II)	Q(III)	Q(IV)	Rating
yes/maybe	no	no	no	4
yes/maybe	–	yes/maybe	no	3
maybe	–	no	maybe	3
no	–	yes/maybe	no	2
no	–	no/maybe	yes/maybe	1
maybe	–	no	yes	1
no	–	no	no	0
yes/maybe	yes	no	no	0

Table 12 Correlation scores corresponding to Fig. 3

	Levin			WordNet		
	Class no.	PMI	Accept	Class name	PMI	Accept
take	#18.1,2	.47	.54	WN-*knock*	.55	.69
	#30.3	.56	.60	WN-*check*	.38	.46
	#43.2	.43	.51	WN-*ring*	.63	.59
	#51.4.2	.54	.55	WN-*sail*	.78	.74
give	#18.1,2	.26	.54	WN-*knock*	.57	.63
	#30.3	.28	.62	WN-*check*	.57	.51
	#43.2	.39	.45	WN-*ring*	.65	.49
	#51.4.2	.16	.25	WN-*sail*	.23	.42
make	#18.1,2	.29	.52	WN-*knock*	.44	.45
	#30.3	.26	.43	WN-*check*	.40	.34
	#43.2	.09	.17	WN-*ring*	.13	.14
	#51.4.2	.32	.73	WN-*sail*	.27	.38

were considered unacceptable or ambiguous by a majority of the annotators. (This resulted in the removal of 11 expressions in total.) Table 12 contains the correlation scores (r_s) for Accept$_{\mathrm{LVC}}$ and PMI$_{\mathrm{LVC}}$ across the three verbs (*take*, *give*, and *make*) and the Levin and WordNet test classes. (These are the numbers used in creating the greyscale representation shown in Fig. 3.)

References

Alba-Salas, J. (2002). Light verb constructions in Romance: A syntactic analysis. PhD thesis, Cornell University.

Baldwin, T., Bannard, C., Tanaka, T., & Widdows, D. (2003). An empirical model of multiword expression decomposability. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 89–96.

Baldwin, T., & Villavicencio, A. (2002). Extracting the unextractable: A case study on verb-particles. In *Proceedings of the Sixth Conference on Computational Natural Language Learning (CoNLL'02)*, pp. 98–104.

Bannard, C., Baldwin, T., & Lascarides, A. (2003). A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 65–72.

BNC Reference Guide (2000). Reference guide for the British National Corpus (World Edition). Second edition.

Brinton, L. J., & Akimoto, M. (Eds.) (1999). *Collocational and idiomatic aspects of composite predicates in the history of English*. John Benjamins Publishing Company.

Butt, M. (2003). The light verb jungle. Manuscript.

Cacciari, C. (1993). The place of idioms in a literal and metaphorical world. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure, and interpretation* (pp. 27–53). Lawrence Erlbaum Associates.

Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: Exploiting on-line resources to build a lexicon* (pp. 115–164). Lawrence Erlbaum.

Claridge, C. (2000). *Multi-word verbs in early modern English: A corpus-based study*. Amsterdam, Atlanta: Rodopi B.V.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Collins, M. (1999). Head-driven statistical models for natural language parsing. PhD thesis, University of Pennsylvania.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Desbiens, M. C., & Simon, M. (2003). Déterminants et locutions verbales. Manuscript.

Dras, M., & Johnson, M. (1996). Death and lightness: Using a demographic model to find support verbs. In *Proceedings of the Fifth International Conference on the Cognitive Science of Natural Language Processing*.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61–74.

Fazly, A. (2007). Automatic acquisition of lexical knowledge about multiword predicates. PhD thesis, University of Toronto.

Fazly, A., North, R., & Stevenson, S. (2005). Automatically distinguishing literal and figurative usages of highly polysemous verbs. In *Proceedings of the ACL'05 Workshop on Deep Lexical Acquisition*, pp. 38–47.

Fazly, A., North, R., & Stevenson, S. (2006). Automatically determining allowable combinations of a class of flexible multiword expressions. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'06)*, pp. 81–92.

Fazly, A., & Stevenson, S. (2006). Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pp. 337–344.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa:I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549.

Fellbaum, C. (Ed.) (1998). *WordNet, an electronic lexical database*. The MIT Press.

Gibbs, R. W. (1993). Why idioms are not dead metaphors. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure, and interpretation* (pp. 57–77). Lawrence Erlbaum Associates.

Gibbs, R., & Nayak, N. P. (1989). Psychololinguistic studies on the syntactic behaviour of idioms. *Cognitive Psychology, 21*, 100–138.

Glucksberg, S. (1993). Idiom meanings and allusional content. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure, and interpretation* (pp. 3–26). Lawrence Erlbaum Associates.

Grefenstette, G., & Teufel, S. (1995). Corpus-based method for automatic identification of support verbs for nominalization. In *Proceedings of the Seventh Meeting of the European Chapter of the Association for Computational Linguistics (EACL'95).*

Inkpen, D. (2003). Building a lexical knowledge-base of near-synonym differences. PhD thesis, University of Toronto.

Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason*. The University of Chicago Press.

Karimi, S. (1997). Persian complex verbs: Idiomatic or compositional? *Lexicology, 3*(1), 273–318.

Kearns, K. (2002). Light verbs in English. Manuscript.

Keller, F., & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics, 29*, 459–484.

Krenn, B., & Evert, S. (2001). Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL'01 Workshop on Collocations*, pp. 39–46.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. The University of Chicago Press.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. The University of Chicago Press.

Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pp. 317–324.

Lin, T. -H. (2001). Light verb syntax and the theory of phrase structure. PhD thesis, University of California, Irvine.

McCarthy, D., Keller, B., & Carroll, J. (2003). Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment.*

Melamed, I. D. (1997). Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the Second Conference on Empirical Methods for Natural Language Processing (EMNLP'97).*

Miyamoto, T. (2000). *The light verb construction in Japanese: The role of the verbal noun*. John Benjamins Publishing Company.

Mohammad, S., & Hirst, G. (2006). Determining word sense dominance using a thesaurus. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pp. 121–128.

Moirón, M. B. V. (2004). Discarding noise in an automatically acquired lexicon of support verb constructions. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC).*

Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press.

Newman, J. (1996). *Give: A cognitive linguistic study*. Mouton de Gruyter.

Newman, J., & Rice, S. (2004). Patterns of usage for English SIT, STAND, and LIE: A cognitively inspired exploration in corpus linguistics. *Cognitive Linguistics, 15*(3), 351–396.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. *Language, 70*(3), 491–538.

Pauwels, P. (2000). *Put, set, lay and place: A cognitive linguistic approach to verbal meaning*. LINCOM EUROPA.

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Rohde, D. L. T. (2004). TGrep2 User Manual.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP'. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'02)*, pp. 1–15.

Seretan, V., Nerima, L., & Wehrli, E. (2003). Extraction of multi-word collocations using syntactic bigram composition. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'03).*

Stevenson, S., Fazly, A., & 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*, pp. 1–8

Turney, P. D. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning (ECML'01)*, pp. 491–502.

Uchiyama, K., Baldwin, T., & Ishizaki, S. (2005). Disambiguating Japanese compound verbs. *Computer Speech and Language, 19*, 497–512.

Venkatapathy, S., & Joshi, A. (2005). Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods for Natural Language Processing (HLT-EMNLP'05)*, pp. 899–906.

Villavicencio, A. (2003). Verb-particle constructions and lexical resources. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp. 57–64.

Villavicencio, A. (2005). The availability of verb-particle constructions in lexical resources: How much is enough? *Computer Speech and Language, 19*, 415–432.

Wanner, L. (2004). Towards automatic fine-grained semantic classification of verb-noun collocations. *Natural Language Engineering, 10*(2), 95–143.

Wermter, J., & Hahn, U. (2005). Paradigmatic modifiability statistics for the extraction of complex multi-word terms. In *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods for Natural Language Processing (HLT-EMNLP'05)*, pp. 843–850.

Wierzbicka, A. (1982). Why can you have a drink When you can't *Have an eat? *Language, 58*(4), 753–799.