In Proceedings of CICLING 2006

Automatically Determining Allowable Combinations of a Class of Flexible Multiword Expressions

Afsaneh Fazly, Ryan North, and Suzanne Stevenson

Department of Computer Science
University of Toronto
Toronto, ON M5S 3H5
Canada
{afsaneh,ryan,suzanne}@cs.toronto.edu

Abstract. We develop statistical measures for assessing the acceptability of a frequent class of multiword expressions. We also use the measures to estimate the degree of productivity of the expressions over semantically related nouns. We show that a linguistically-inspired measure outperforms a standard measure of collocation in its match with human judgments. The measure uses simple extraction techniques over non-marked-up web data.

1 Light Verb Constructions

Recent work in NLP has recognized the challenges posed by the rich variety of multiword expressions (MWEs) (e.g., Sag et al., 2002). One unsolved problem posed by MWEs is how they should be encoded in a computational lexicon. Many MWEs are syntactically flexible; for these it is inappropriate to treat the full expression as a single word. However, fully compositional techniques can lead to overgeneralization, because flexible MWEs are often *semi*-productive: new expressions can only be formed from limited combinations of semantically and syntactically similar component words. In order to achieve accurate lexical acquisition methods, we must determine computational mechanisms for capturing the allowable combinations of such MWEs.

Our focus here is on light verb constructions (LVCs); these are largely compositional and semi-productive MWEs having a high frequency of occurrence across many diverse languages (Karimi, 1997; Miyamoto, 2000; Butt, 2003). LVCs combine a member of a restricted set of light verbs, such as *give*, *take*, and *make* among others in English, with a wide range of complements of varying syntactic categories. We consider a common class of LVCs, in which the complement is a noun generally used with an indefinite article, as in (a–c) below:

a. Priya *took a walk* along the beach.
b. Allene *gave a smile* when she saw us.
c. Randy *made a joke* to his friends.

d. Priya *walked* along the beach.
e. Allene *smiled* when she saw us.
f. Randy *joked* to his friends.

Moreover, the complement nouns in these expressions, such as *walk*, *smile*, and *joke* in (a–c), have a stem form identical to a verb. Because the light verb is "semantically bleached" to some degree (Butt, 2003), most of the meaning of these LVCs comes from the complement. The predicative nature of the complement is illustrated by the fact that the noun complements in (a–c) contribute the verbs of the corresponding paraphrases in (d–f).

These LVCs are of interest because they are very frequent, and moreover, their productivity appears to be patterned (Kearns, 2002; Wierzbicka, 1982). 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 and on the light verb itself. For example, in contrast to *take*, we observe ?*give a walk*, ?*give a stroll*, ?*give a run*, compared to *give a groan*, *give a smile*, *give a wink*. Thus, these constructions provide a rich testbed for exploring computational means for capturing the range of allowable combinations of semi-productive MWEs.

We approach this problem by developing and evaluating acceptability measures for LVCs, which draw on their linguistic properties. Furthermore, we investigate the hypothesis that the acceptability of a candidate LVC depends on the semantic class of its complement. Two semantic classifications for the potential complements are compared, to explore the impact of different semantic similarity criteria. Human acceptability judgments of each candidate expression are gathered, and used as the standard against which our statistical measures are evaluated.

Our results indicate a high level of compatibility between our computational measures of acceptability and human judgments. Moreover, we find that the measures can be used to quantify the productivity of a class of complements, i.e., the extent to which the semantic class forms acceptable expressions with a particular light verb. Automatically assessing both acceptability of individual expressions, and productivity across semantic classes, enables us to take a first step toward adequate representation of LVCs in a computational lexicon. Since such semi-productive behaviour arises frequently (e.g., in verb-particle constructions and other phrasal verbs), we believe our approach yields insights for the automatic extraction and representation of MWEs more generally.

2 Acceptability Measures

We present three statistical measures for a continuously valued assessment of LVC acceptability. These measures capture in differing ways the association between a light verb (LV) and a noun complement. Since the complement of such LVCs contributes event semantics to the expression (as illustrated in (a–f) above), the noun must be a predicative noun (PN)—i.e., a noun that has an argument structure. Because the PN is preceded by an indefinite determiner (*a* or *an*) in these expressions, we refer to the complement in our formulas below as *aPN*.[1]

2.1 The $\mathrm{PMI}_{\mathrm{LVC}}$ Measure

Following Stevenson et al. (2004), our first measure uses pointwise mutual information (PMI), a standard measure of collocation (Church et al., 1991), to assess the strength of

[1] Since LVCs are somewhat flexible (*give it a try*, *take a nice walk*), we allow other intervening words between the LV and PN in some of our counts, as described in detail in §3.3.

association between a given LV and PN:[2]

$$\mathrm{PMI}_{\mathrm{LVC}}(LV;aPN) = \log \frac{n\, f(LV,aPN)}{f(LV)\, f(aPN)},$$

where n is the corpus size. Higher values of $\mathrm{PMI}_{\mathrm{LVC}}$ reveal a greater degree of association between the LV and PN, which can be interpreted as an indication of LVC acceptability.

2.2 The $\mathrm{Prob}_{\mathrm{LVC}}$ Measure

While common LVCs typically appear as good collocations, the $\mathrm{PMI}_{\mathrm{LVC}}$ measure fails to incorporate other important properties of LVCs. Here, we propose a linguistically-motivated measure, $\mathrm{Prob}_{\mathrm{LVC}}$, which captures the likelihood that a given LV and PN form an acceptable LVC, i.e., $\mathrm{Pr}(LV,PN,LVC)$. The joint probability is factored as:

$$\mathrm{Pr}(PN)\ \mathrm{Pr}(LVC|PN)\ \mathrm{Pr}(LV|PN,LVC).$$

The first factor, $\mathrm{Pr}(PN)$, reflects the observation that higher frequency words are more likely to be used as complements in LVCs (Wierzbicka, 1982). We estimate this probability by $f(PN)/n$, where n is the number of words in the corpus.

The $\mathrm{Pr}(LVC|PN)$ factor captures the general tendency of the PN to form LVCs with *any* light verb. This factor is estimated by the number of times we observe the prototypical LVC pattern “LV a PN” with this PN across possible LVs:

$$\mathrm{Pr}\,(LVC|PN) \approx \frac{\sum_{i=1}^{v} f(LV_i, aPN)}{f(aPN)},$$

where v is the number of LVs in our study. Since we are only counting usages of the PN in the context of an indefinite determiner in the numerator, we normalize over counts of *aPN*. Note that simply counting “LV a PN” as an LVC is an overestimate, since we cannot determine which of such usages are indeed LVCs, as opposed to literal usages of the LV as in *give a present*. However, we expect that true predicative nominals will have a higher probability of usage in LVCs than other nouns, since the noun complement must contribute an argument structure to the LVC.

Finally, $\mathrm{Pr}(LV|PN,LVC)$ reflects the specific tendency of the PN to form an LVC with this particular light verb, LV. We similarly estimate this factor with counts of the LV and PN in the typical LVC pattern: $f(LV,aPN)/f(aPN)$.

Combining the estimation of the three factors results in the following formula:

$$\mathrm{Prob}_{\mathrm{LVC}}\,(LV,PN) \approx \frac{f(PN)}{n} \times \frac{\sum_{i=1}^{v} f(LV_i, aPN)}{f(aPN)} \times \frac{f(LV,aPN)}{f(aPN)},$$

where v is the number of LVs and n the corpus size.

[2] PMI has some limitations with very low frequency items, but since we use the web as our corpus (see §3), we do not expect counts of such low frequency.

2.3 The Freq_{LVC} Measure

We also propose here an additional measure, Freq_{LVC}, for which the primary goal is inexpensive extraction from noisy but plentiful data. Freq_{LVC} assesses the acceptability of a candidate LVC simply according to its raw frequency in the corpus. But from that raw frequency, we subtract an estimate of the amount of noise affecting the candidate expression, in order to better approximate the frequency of "true" LVC usage.

Our estimate of noise is based on the intuition that the likelihood of seeing an LVC with m internal modifiers—intervening words between the LV and PN—approaches zero as m increases. For example, we expect to find *take a walk* more often than *take a long walk*, which in turn is more probable than *take a long relaxing walk*, etc. We assume there exists a threshold, t, at which the likelihood of producing an LVC involving $m \geq t$ words of internal modification is negligible. At this threshold, any results found must be noise—i.e., cooccurrences of the LV and PN that are unrelated to LVC usage.

Let $f_m(LV, aPN)$ be the frequency of the string "LV a *word*^{m} PN". As we increase the value of m from 0 to t, the number of actual LVC usages included in $f_m(LV, aPN)$ gradually decreases. Under the assumption that the amount of noisy results remains roughly constant, we can use $f_t(LV, aPN)$ as the estimate of noise for each count. Thus if $f_0(LV, aPN)$ is the count of "LV a PN", including both actual LVCs and noise, then if we subtract from it the estimate of noise, $f_t(LV, aPN)$, we have an estimate of the actual LVC usage when $m = 0$.

The assumption that noise remains constant does not hold in practice (as actual LVC usage decreases, noise increases). However, we find that by taking an average of $f_t(LV, aPN)$ across a range of values of t, we achieve a useful estimate of noise. The resulting measure is defined as:

$$\text{Freq}_{\text{LVC}}(LV, PN) = f_0(LV, aPN) - \underset{t}{\text{mean}}\ f_t(LV, aPN).$$

In our experiments, t is in the range [6,10], empirically established through experiments on the development data.

3 Materials and Methods

3.1 Light Verbs Used

Linguists have identified a small set of verbs which, crosslinguistically, are commonly used as light verbs. We focus on three frequent light verbs in English, *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* is less clear. We expect then that *make* will show contrasting behaviour.

3.2 Experimental Expressions

Experimental expressions—i.e., potential LVCs using *take*, *give*, and *make*—are formed by combining the three light verbs with predicative nouns from (i) selected semantic verb classes of (Levin, 1993); or (ii) generated WordNet classes. In each case, some

Table 1. Seed words selected according to acceptability trends identified for each Levin test class.

	Acceptability Trend			
Levin Class	*take*	*give*	*make*	Seed Word
Hit, 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 that are not vehicle names.

classes are used as development data, and some classes as test data. The following paragraphs explain the selection of Levin classes, and the process of generating corresponding classes using WordNet.

Selection of Levin classes It may seem odd to use a verb classification for noun complements. However, recall that an important property of the type of LVCs we are considering is that the complement noun has an argument structure, and is identical in stem form to a verb. The verb classes of Levin (1993), defined by similarity of argument structure, therefore provide natural similarity sets to consider. As long as we only use verbs identical in form to a noun, we are assured that such complements are PNs.

Our three development and four test classes from Levin are taken from Stevenson et al. (2004). These classes reflect a range of productivity in combination with the three light verbs. For classes with more than 35 verbs (30 for development classes), a random subset of that size is selected for experimentation.

Generation of WordNet classes 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). We therefore also use semantic classes generated from both the noun and verb hierarchies of WordNet 2.0 (Fellbaum, 1988). In determining these WordNet-derived classes, it is important that they are comparable to each of the Levin classes, so that we can relate performance of our measures across the corresponding classes from the two classifications. We achieve this by generating a WordNet set that is semantically similar to a representative word from a given Levin class.

Specifically, for each Levin class, we first determine the general pattern of acceptability of that class with the different light verbs. As described in §3.4 below, human ratings are put into buckets of ‘poor’, ‘fair’, and ‘good’. We then determine the predominant bucket for each class and light verb, and select a representative PN seed that best reflects the typical ratings across the three light verbs (see Table 1). For each seed, we 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 form to a verb.[3] A random selection of 35 of the remaining words forms a WordNet class, which we refer to by “WN-” plus the seed word (e.g., WN-*knock*).

[3] In contrast to the Levin expressions, we also filter rare PNs, whose frequency as a verb in the British National Corpus is less than 50.

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

3.3 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 very frequently attested in a traditional corpus. We therefore decided to use the web (the subsection indexed by Google) to estimate frequency counts required by the statistical measures. Each count is calculated via an exact-phrase search; counts including LVs are collapsed across separate searches using three tenses of the verb: base, present, and simple past. The number of hits is used as the frequency of the string searched for. 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.[4] Such web-based frequency estimates have been successfully used in many NLP applications (Turney, 2001; Villavicencio, 2003), and have been shown to highly correlate with frequency counts from a balanced corpus (Keller and Lapata, 2003).

Most LVCs allow their noun component to be modified, as in *take a long walk*. To capture such uses, we use the '*' wildcard (as in "*take a * walk*"), which matches exactly one word. Moreover, many LVCs using the light verb *give* frequently appear in the dative form; 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 intervening between the LV and PN components. The estimated frequency of an expression is the sum over its bare, adjectival, and dative forms. (The additional searches are not run for Freq_{LVC}, as it is designed to explore rating LVCs using little information.)

3.4 Human Acceptability Judgments

Two expert native speakers of English rated the acceptability of each experimental expression. The ratings range from 1 (unacceptable) to 4 (completely natural), by 0.5 increments. On Levin test expressions, the two sets of ratings yield kappa values of .72, .39, and .44, for *take*, *give*, and *make*, respectively, and .53 overall. (We use linearly weighted kappa, since our ratings are ordered.) Wide differences in ratings typically arose when one rater missed a possible meaning for an expression; these were corrected in the 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

[4] 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.

.71 overall. Ratings were reconciled to within one point difference, and then averaged 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.

4 Experimental Results

We evaluate our measures by comparing their acceptability scores with the consensus human ratings: Spearman rank correlation coefficient (r_s) is used to compare the rankings provided by the two (§4.1); linearly weighted observed agreement (p_o) is used to examine their agreement at the coarser level of the acceptability buckets (§4.2). The acceptability buckets are further used to determine the appropriateness of our measures for predicting the productivity of a class with respect to LVC formation (§4.3). We focus on the performance on unseen test data; trends are similar on development data.

4.1 Correlation between Human Ratings and Statistical Measures

We perform separate correlation tests between the consensus human ratings and the three measures over each of the three LVs in combination with each of the four test classes within the two classifications (Levin and WordNet). In Figure 1, we show the results graphically, so that patterns are easier to see. Each rectangle represents a separate correlation calculation. 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 increasingly darker rectangles.[5] We discuss the results in terms of the measures, the light verbs, and the two classifications, in turn.

The Prob_{LVC} measure is the most consistent of the three measures, performing best overall and achieving good correlations in most cases. The PMI_{LVC} measure does surprisingly well, as a simple measure of collocation; it even performs comparably to Prob_{LVC} on the WordNet classes. Freq_{LVC} has reasonably good performance on the Levin classes, but relatively poor performance on WordNet classes. It is the most knowledge-poor of the measures, and also the most inconsistent performer, indicating that such simple methods are inappropriate for fine-grained acceptability measures in this task.

Examining the patterns in Figure 1 by light 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 §3.4).

Now looking at the patterns across the two semantic classifications, we note that the performance of Prob_{LVC} is overall comparable across the two, while PMI_{LVC} shows a marked improvement with WordNet, and Freq_{LVC} a marked decline. A closer look at the WordNet and Levin expressions reveals an interesting difference between the two: the average frequency of PNs in the WordNet classes is significantly higher than that of

[5] We used a significance cut-off of $p < .07$, since some tests achieved reasonably good correlations that were marginally significant at this level. Numerical r_s values are available in an unpublished TR at the authors’ website.

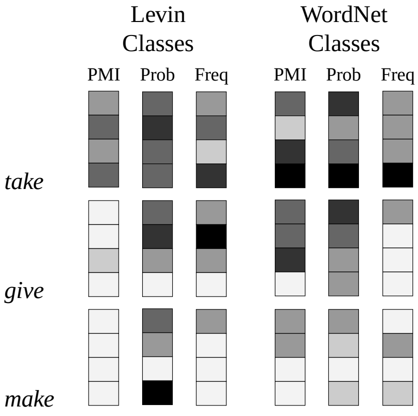

Fig. 1. Greyscale representation of the correlation coefficients (r_s) of each measure, across the three light verbs, for the four Levin and WordNet test classes.

PNs in the corresponding Levin classes (26M vs. 8M, respectively). This observation provides evidence for the robustness of Prob_{LVC}, which appears to be less sensitive to frequency factors than PMI_{LVC} or Freq_{LVC}.

The effect of semantic classification on the measures also interacts with the specific light verb being used. We see that PMI_{LVC} is particularly inferior on Levin classes with *give* and *make*. It seems that expressions with *give* and *make* are less treatable as straightforward collocations, especially with lower frequency items.

4.2 Agreement between Human Ratings and Statistical Measures

We now inspect the performance of our statistical measures when the coarser level of acceptability—'poor', 'fair', or 'good'—is considered. For each measure, thresholds for dividing the continuous ratings into the discrete buckets are chosen such that the bucket sizes for development data match as closely as possible those of the human ratings. We then compare the measures on unseen test data to a chance baseline using the (linearly weighted) proportion of observed agreement with the "bucketized" human ratings.[6] For most LV–class pairs, 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'.

Table 2 presents the observed agreement scores (p_o) averaged across classes in each classification (Levin or WordNet); values of p_o above the baseline are in boldface. On Levin and WordNet expressions with *take* and *give*, both Prob_{LVC} and Freq_{LVC} mostly outperform the baseline, with Prob_{LVC} performing the best. On expressions involving

[6] 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.

Table 2. Weighted observed agreement (p_o) between statistical measures and human judgments, over Levin and WordNet test expressions.

Light Verb	Class Type	Chance Agreement	PMI_{LVC}	Prob_{LVC}	Freq_{LVC}
take	Levin	.78	.77	**.85**	**.80**
	WordNet	.81	**.88**	**.86**	**.82**
give	Levin	.80	.59	.77	**.86**
	WordNet	.75	.74	**.80**	.73
make	Levin	.87	.81	.82	.82
	WordNet	.85	.80	.74	.73

make, however, none of the measures perform better than the 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_{LVC} does better on WordNet classes, and Freq_{LVC} on Levin classes.

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

4.3 Predicting Class Productivity

The ultimate goal of this study is to devise statistical measures that are good indicators of the semi-productivity of LVC formation for a semantic class of predicative nouns, given a particular light verb. One aspect of this is our proposed measures of the individual acceptability of a particular LV and PN combination as an LVC. We also want to assess the overall acceptability of a class of semantically related PNs, which indicates the productivity of the class with respect to the LV. Such class knowledge can be useful in extending our measures of acceptability to new or low frequency PNs. For example, if our measure predicts that the class of sound emission nouns, such as *groan* and *yell*, productively forms acceptable LVCs with *give*, the acceptability of an unseen LVC such as *give a moan* should be promoted.

The productivity level of a class is indicated by the proportion of PNs that form acceptable LVCs with the given LV. We consider an acceptable LVC to be one that is either ‘fair’ or ‘good’ according to human judgments. Thus, to investigate the appropriateness of each proposed measure as an indicator of class productivity, we compare (for each combination of LV and semantic class of PNs) that measure’s proportions of PNs in the ‘fair’ and ‘good’ buckets with those 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 LV (*take* and *give*) and semantic class. As an example, Table 3 presents the productivity of each WordNet test class for *take*, as determined by human

Table 3. Proportion of acceptable expressions (those rated ‘fair’ or ‘good’) for *take* and each WordNet test class, as determined by human ratings and the statistical measures.

Class	Human	PMI_{LVC}	Prob_{LVC}	Freq_{LVC}
WN-*knock*	.26	.40	.26	.40
WN-*check*	.14	.09	.26	.34
WN-*ring*	.09	.17	.23	.20
WN-*sail*	.46	.40	.37	.34

Table 4. Divergence of the productivity levels assessed by each of the statistical measures from those determined by human judges, averaged across Levin or WordNet classes.

Class	SSE × 100		
Type	PMI_{LVC}	Prob_{LVC}	Freq_{LVC}
Levin	22.0	**9.0**	12.0
WordNet	5.7	**3.5**	8.1

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 phenomenon.

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 all classes and light verbs. The divergence is measured as the sum of squared errors (SSE) between the two sets of numbers, averaged over all light verbs and classes. Table 4 shows the average SSE values for each measure and each classification (Levin or WordNet). The lowest SSE (best match to human judgments) is shown in bold. For both classifications, Prob_{LVC} gives the closest predictions, i.e., the lowest SSEs. Notably, here we see overall better performance with WordNet than with Levin classes across all three measures.

4.4 Summary of Results

Our results indicate that Prob_{LVC}, the measure that incorporates more linguistic knowledge about LVCs, performs well at assessing acceptability at both the fine- and coarse-grained levels, according to the observed r_s and p_o values, respectively. Prob_{LVC} also accurately predicts the degree of productivity of a semantic class of complements with a light verb, according to the reported SSE values. PMI_{LVC} achieves reasonably good performance at both tasks when using WordNet classes, while Freq_{LVC} shows inconsistent performance across the tasks and the classifications.

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 work.

5 Related Work

Compared to other types of MWEs, such as verb particle constructions, LVCs have not been studied computationally in great detail. Grefenstette and Teufel (1995) and Dras and Johnson (1996) examine the problem of choosing the best support verb (similar to an LV) for a given deverbal noun complement (similar to a PN). This is too restrictive for our purposes, since the same complement may form acceptable LVCs with different light verbs. Like us, Moirón (2004) links surface syntactic behaviour of LVCs to their underlying semantics; however, her approach requires a great deal of manual analysis.

Sag et al. (2002) address the lexical encoding of LVCs more directly, but consider the selection of complements by an LV mainly idiosyncratic. Although they mention the use of selectional restrictions for LVs, they do not give an explicit means for determining the allowable combinations of semi-productive LVCs. Stevenson et al. (2004) particularly focus on the issue of semi-productivity of LVCs using Levin classes, but lack a clear proposal for extending their PMI-based acceptability measure to assess productivity. Here we propose a measure, Prob_{LVC}, that captures linguistic properties of LVCs relevant to their acceptability in a more appropriate manner, and explore its effectiveness across WordNet classes as well. We also show that Prob_{LVC} fits well with the human judgments on predicting the productivity level of both types of classes.

Our study of semantic classes is related to the idea of substitutability in other types of MWEs, i.e., substituting part of an MWE with a semantically similar word to determine the productivity of the expression (McCarthy et al., 2003; Lin, 1999; Villavicencio, 2003). However, the approach in this work differs not only in focusing on LVCs, but also in its goal of quantifying degree of acceptability of an expression in order to more precisely assess productivity. Moreover, a contribution of this paper is the investigation of different classifications and their impact on performance of our measures.

6 Conclusions

We have developed three statistical measures of the acceptability of light verb constructions, for use in automatically determining the allowable complements of a light verb. In comparisons against human judgments, we find that the Prob_{LVC} measure, which incorporates some linguistic insight within a probabilistic formulation, performs best and most consistently overall, for both fine- and coarse-grained assessment of acceptability. The results demonstrate that LVCs are best treated as more than simple collocations. Moreover, estimation of the Prob_{LVC} measure requires only simple extraction techniques over non-marked-up web data.

Our findings also show that Prob_{LVC} yields an accurate assessment of the productivity of a class of semantically related nouns as potential complements of a light verb. Due to the semi-productive nature of LVCs, such an assessment is crucial for generalizing the knowledge in a computational lexicon to previously unseen potential complements.

Given the crosslinguistic prominence of light verb constructions, our future work aims to extend these techniques to similar constructions in languages other than English. Moreover, while we have focused here on LVCs, we believe that similar techniques can be useful in dealing with other semi-productive MWEs, especially other types of phrasal verbs which are crosslinguistically frequent as well.

Bibliography

Miriam Butt. 2003. The light verb jungle. http://edvarda.hf.ntnu.no/ling/tross/Butt.pdf.

Kenneth W. Church, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In U. Zernik, editor, *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, pages 115–164. Lawrence Erlbaum Associates.

Mark Dras and Mike Johnson. 1996. Death and lightness: Using a demographic model to find support verbs. In *Proceedings of the 5th ICCSNLP*, pages 165–172.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low Kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Christiane Fellbaum, editor. 1988. *WordNet: An Electronic Lexical Database*. The MIT Press.

Gregory Grefenstette and Simone Teufel. 1995. Corpus-based method for automatic identification of support verbs for nominalizations. In *Proceedings of the 7th EACL*, pages 98–103.

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

Kate Kearns. 2002. Light verbs in English. Manuscript.

Frank Keller and Maria Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Beth Levin. 1993. *English Verb Classes and Alternations, A Preliminary Investigation*. University of Chicago Press.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of ACL-99*, pages 317–324.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions*, pages 73–80.

Tadao Miyamoto. 2000. *The Light Verb Construction in Japanese: the role of the verbal noun*. John Benjamins Publishing Company.

M. Begoña Villada Moirón. 2004. Discarding noise in an automatically acquired lexicon of support verb constructions. In *Proceedings of the 4th LREC*.

John Newman. 1996. *Give: A Cognitive Linguistic Study*. Mouton de Gruyter.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd CICLING*, pages 1–15.

Suzanne Stevenson, Afsaneh Fazly, and Ryan North. 2004. Statistical measures of the semi-productivity of light verb constructions. In *Proceedings of the ACL'04 Workshop on Multiword Expressions*, pages 1–8.

Peter D. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th ECML*, pages 491–502.

Aline Villavicencio. 2003. Verb-particle constructions in the World Wide Web. In *Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their use in Computational Linguistics Formalisms and Applications*.

Anna Wierzbicka. 1982. Why can you Have a Drink when you can't *Have an Eat? In *The Semantics of Grammar*, pages 293–358. John Benhamins Publishing Co.