Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures

Afsaneh Fazly
Department of Computer Science
University of Toronto
Toronto, Canada
afsaneh@cs.toronto.edu

Suzanne Stevenson
Department of Computer Science
University of Toronto
Toronto, Canada
suzanne@cs.toronto.edu

Abstract

We identify several classes of multiword expressions that each require a different encoding in a (computational) lexicon, as well as a different treatment within a computational system. We examine linguistic properties pertaining to the degree of semantic idiosyncrasy of these classes of expressions. Accordingly, we propose statistical measures to quantify each property, and use the measures to automatically distinguish the classes.

1 Motivation

Multiword expressions (MWEs) are widely used in written language as well as in colloquial speech. An MWE is composed of two or more words that together form a single unit of meaning, e.g., *frying pan*, *take a stroll*, and *kick the bucket*. Most MWEs behave like any phrase composed of multiple words, e.g., their components may be separated, as in *She took a relaxing stroll along the beach*. Nonetheless, MWEs are distinct from multiword phrases because they involve some degree of semantic idiosyncrasy, i.e., the overall meaning of an MWE diverges from the combined contribution of its constituent parts. Because of their frequency and their peculiar behaviour, MWEs pose a great challenge to the creation of natural language processing (NLP) systems (Sag et al., 2002). NLP applications, such as semantic parsing and machine translation should not only identify MWEs, but also should know how to treat them when they are encountered.

Semantic idiosyncrasy is a matter of degree (Nunberg et al., 1994). The idiom *shoot the breeze* is largely idiosyncratic, because its meaning ("to chat") does not have much to do with the meaning of *shoot* or *breeze*. MWEs such as *give a try* ("try") and *make a decision* ("decide") are semantically less idiosyncratic (more predictable). These are MWEs because the overall meaning of the expression diverges from the combined meanings of the constituents. Nonetheless, there is some degree of predictability in their meanings that makes them distinct from idioms. In these, the complement of the verb (here, a noun) determines the primary meaning of the overall expression. This class of expressions is referred to as light verb constructions (LVCs) in the linguistics literature (Miyamoto, 2000; Butt, 2003).

Clearly, a computational system should distinguish idioms and LVCs, both from each other, and from similar-on-the-surface (literal) phrases such as *shoot the bird* and *give a present*. Idioms are largely idiosyncratic; a computational lexicographer thus may decide to list idioms such as *shoot the breeze* in a lexicon along with their idiomatic meanings. In contrast, the meaning of MWEs such as *make a decision* can be largely predicted, given that they are LVCs. Table 1 shows the different underlying semantic structure of a sentence containing an idiom (*shoot the breeze*) and a sentence containing an LVC (*give a try*). As can be seen, such MWEs should also be treated differently when translated into another language. Note that in contrast to a literal combination, such as *shoot the bird*, for idioms and LVCs, the number of arguments expressed syntactically may differ from the number of the semantic participants.

Many NLP applications also need to distinguish another group of MWEs that are less idiosyncratic

Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, pages 9–16,
Prague, June 2007.

Class	English sentence	Semantic representation	French translation
Literal	*Jill and Tim shot the bird.*	(event/SHOOT :agent ("*Jill ∧ Tim*") :theme ("*bird*"))	*Jill et Tim ont abattu l'oiseau.* *Jill and Tim shot down the bird.*
Abstract	*Jill makes a living singing in pubs.*	(event/EARN-MONEY :agent ("*Jill*"))	*Jill gagne sa vie en chantant dans des bars.* *Jill makes a living by singing in the pubs.*
LVC	*Jill gave the lasagna a try.*	(event/TRY :agent ("*Jill*") :theme ("*lasagna*"))	*Jill a essayé le lasagne.* *Jill tried the lasagna.*
Idiom	*Jill and Tim shot the breeze.*	(event/CHAT :agent ("*Jill ∧ Tim*"))	*Jill et Tim ont bavardé.* *Jill and Tim chatted.*

Table 1: Sample English MWEs and their translation in French.

than idioms and LVCs, but more so than literal combinations. Examples include *give confidence* and *make a living*. These are idiosyncratic because the meaning of the verb is a metaphorical (abstract) extension of its basic physical semantics. Moreover, they often take on certain connotations beyond the compositional combination of their constituent meanings. They thus exhibit behaviour often attributed to collocations, e.g., they appear with greater frequency than semantically similar combinations. For example, searching on Google, we found much higher frequency for *give confidence* compared to *grant confidence*. As can be seen in Table 1, an abstract combination such as *make a living*, although largely compositional, may not translate word-for-word. Rather, it should be translated taking into account that the verb has a metaphorical meaning, different from its basic semantics.

Here, we focus on a particular class of English MWEs that are formed from the combination of a verb with a noun in its direct object position, referred to as verb+noun combinations. Specifically, we provide a framework for identifying members of the following semantic classes of verb+noun combinations: (i) literal phrases (LIT), (ii) abstract combinations (ABS), (iii) light verb constructions (LVC), and (iv) idiomatic combinations (IDM). Section 2 elaborates on the linguistic properties related to the differences in the degree of semantic idiosyncrasy observed in members of the above four classes. In Section 3, we propose statistical measures for quantifying each of these properties, and use them as features for type classification of verb+noun combinations. Section 4 and Section 5 present an evaluation of our proposed measures. Section 6 discusses the related studies, and Section 7 concludes the paper.

2 Semantic Idiosyncrasy: Linguistic Properties

Linguists and lexicographers often attribute certain characteristics to semantically idiosyncratic expressions. Some of the widely-known properties are institutionalization, lexicosyntactic fixedness, and non-compositionality (Cowie, 1981; Gibbs and Nayak, 1989; Moon, 1998). The following paragraphs elaborate on each property, as well as on its relevance to the identification of the classes under study.

Institutionalization is the process through which a combination of words becomes recognized and accepted as a semantic unit involving some degree of semantic idiosyncrasy. IDMs, LVCs, and ABS combinations are institutionalized to some extent.

Lexicosyntactic fixedness refers to some degree of lexical and syntactic restrictiveness in a semantically idiosyncratic expression. An expression is lexically fixed if the substitution of a semantically similar word for any of its constituents does not preserve its original meaning (e.g., compare *spill the beans* and *spread the beans*). In contrast to LIT and ABS combinations, IDMs and LVCs are expected to exhibit lexical fixedness to some extent.

An expression is syntactically fixed if it cannot undergo syntactic variations and at the same time retain its original semantic interpretation. IDMs and LVCs are known to show strong preferences for the syntactic patterns they appear in (Cacciari and Tabossi, 1993; Brinton and Akimoto, 1999). E.g., compare

Joe gave a groan with ?*A groan was given by Joe*, and *Tim kicked the bucket* with **Tim kicked the buckets* (in the idiom reading). Nonetheless, the type and degree of syntactic fixedness in LVCs and IDMs are different. For example, most LVCs prefer the pattern in which the noun is introduced by the indefinite article *a* (as in *give a try* and *make a decision*), whereas this is not the case with IDMs (e.g., *shoot the breeze* and *kick the bucket*). IDMs and LVCs may also exhibit preferences with respect to adjectival modification of their noun constituent. LVCs are expected to appear both with and without an adjectival modifier, as in *give a (loud) groan* and *make a (wise) decision*. IDMs, on the other hand, mostly appear either with an adjective, as in *keep an open mind* (cf. ?*keep a mind*), or without, as in *shoot the breeze* (cf. ?*shoot the fun breeze*).

Non-compositionality refers to the situation where the meaning of a word combination deviates from the meaning emerging from a word-by-word interpretation of it. IDMs are largely non-compositional, whereas LVCs are semi-compositional since their meaning can be mainly predicted from the noun constituent. ABS and LIT combinations are expected to be largely compositional.

None of the above-mentioned properties are sufficient criteria by themselves for determining which semantic class a given verb+noun combination belongs to. Moreover, semantic properties of the constituents of a combination are also known to be relevant for determining its class (Uchiyama et al., 2005). Verbs may exhibit strong preferences for appearing in MWEs from a particular class, e.g., *give*, *take* and *make* commonly form LVCs. The semantic category of the noun is also relevant to the type of MWE, e.g., the noun constituent of an LVC is often a predicative one. We hypothesize that if we look at evidence from all these different sources, we will find members of the same class to be reasonably similar, and members of different classes to be notably different.

3 Statistical Measures of Semantic Idiosyncrasy

This section introduces measures for quantifying the properties of idiosyncratic MWEs, mentioned in the previous section. The measures will be used as features in a classification task (see Sections 4–5).

3.1 Measuring Institutionalization

Corpus-based approaches often assess the degree of institutionalization of an expression by the frequency with which it occurs. Raw frequencies drawn from a corpus are not reliable on their own, hence association measures such as pointwise mutual information (PMI) are also used in many NLP applications (Church et al., 1991). PMI of a verb+noun combination $\prec v,\, n \succ$ is defined as:

$$\begin{aligned} \mathrm{PMI}\,(v,\, n) &\doteq \log \frac{P(v,\, n)}{P(v)\, P(n)} \\ &\approx \log \frac{f(*,\, *) f(v,\, n)}{f(v, *)\, f(*,\, n)} \end{aligned} \tag{1}$$

where all frequency counts are calculated over verb–object pairs in a corpus. We use both frequency and PMI of a verb+noun combination to measure its degree of institutionalization. We refer to this group of measures as `INST`.

3.2 Measuring Fixedness

To measure fixedness, we use statistical measures of lexical, syntactic, and overall fixedness that we have developed in a previous study (Fazly and Stevenson, 2006), as well as some new measures we introduce here. The following paragraphs give a brief description of each.

$\mathrm{Fixedness}_{\mathrm{lex}}$ quantifies the degree of lexical fixedness of the target combination, $\prec v, n \succ$, by comparing its strength of association (measured by PMI) with those of its lexical variants. Like Lin (1999), we generate lexical variants of the target automatically by replacing either the verb or the noun constituent by a semantically similar word from the automatically-built thesaurus of Lin (1998). We then use a standard statistic, the z-score, to calculate $\mathrm{Fixedness}_{\mathrm{lex}}$:

$$\mathrm{Fixedness}_{\mathrm{lex}}(v,\, n) \doteq \frac{\mathrm{PMI}(v,\, n) - \overline{\mathrm{PMI}}}{std} \tag{2}$$

where $\overline{\mathrm{PMI}}$ is the mean and std the standard deviation over the PMI of the target and all its variants.

$\mathrm{Fixedness}_{\mathrm{syn}}$ quantifies the degree of syntactic fixedness of the target combination, by comparing its behaviour in text with the behaviour of a typical verb–object, both defined as probability distributions over a predefined set of patterns. We use a standard information-theoretic measure, relative entropy,

v	det:NULL	n_{sg}	v	det:NULL	n_{pl}
v	det:*a/an*	n_{sg}			
v	det:*the*	n_{sg}	v	det:*the*	n_{pl}
v	det:DEM	n_{sg}	v	det:DEM	n_{pl}
v	det:POSS	n_{sg}	v	det:POSS	n_{pl}
v	det:OTHER	$n_{sg,pl}$	det:ANY $n_{sg,pl}$ be $v_{passive}$		

Table 2: Patterns for syntactic fixedness measure.

to calculate the divergence between the two distributions as follows:

$$\begin{aligned}&\mathrm{Fixedness_{syn}}\,(v,\ n)\\ &\quad\doteq\ D(P(pt|v,n)\,||\,P(pt))\\ &\quad=\ \sum_{pt_k\in\mathcal{P}} P(pt_k|\ v,\ n)\log\frac{P(pt_k|\ v,\ n)}{P(pt_k)}\end{aligned}\qquad(3)$$

where $\mathcal{P}$ is the set of patterns (shown in Table 2) known to be relevant to syntactic fixedness in LVCs and IDMs. $P(pt|\ v,\ n)$ represents the syntactic behaviour of the target, and $P(pt)$ represents the typical syntactic behaviour over all verb–object pairs.

$\mathrm{Fixedness_{syn}}$ does not show which syntactic pattern the target prefers the most. We thus use an additional measure, $\mathrm{Pattern_{dom}}$, to determine the dominant pattern for the target:

$$\mathrm{Pattern_{dom}}(v,\ n)\ \doteq\ \underset{pt_k\in\mathcal{P}}{\mathrm{argmax}}\, f(v,\ n,\ pt_k)\qquad(4)$$

In addition to the individual measures of fixedness, we use $\mathrm{Fixedness_{overall}}$, which quantifies the degree of overall fixedness of the target:

$$\begin{aligned}&\mathrm{Fixedness_{overall}}\,(v,\ n)\\ &\quad\doteq\ \alpha\ \mathrm{Fixedness_{syn}}\,(v,\ n)\\ &\qquad+(1-\alpha)\ \mathrm{Fixedness_{lex}}\,(v,\ n)\end{aligned}\qquad(5)$$

where α weights the relative contribution of lexical and syntactic fixedness in predicting semantic idiosyncrasy.

$\mathrm{Fixedness_{adj}}$ quantifies the degree of fixedness of the target combination with respect to adjectival modification of the noun constituent. It is similar to the syntactic fixedness measure, except here there are only two patterns that mark the presence or absence of an adjectival modifier preceding the noun:

$$\mathrm{Fixedness_{adj}}(v,\ n)\ \doteq\ D(P(a_i|v,n)\,||\,P(a_i))\qquad(6)$$

where $a_i \in \{\mathrm{present}, \mathrm{absent}\}$. $\mathrm{Fixedness_{adj}}$ does not determine which pattern of modification the target combination prefers most. We thus add another measure—the odds of modification—to capture this:

$$\mathrm{Odds_{adj}}(v,\ n)\ \doteq\ \frac{P(a_i=\mathrm{present}|v,n)}{P(a_i=\mathrm{absent}|v,n)}\qquad(7)$$

Overall, we use six measures related to fixedness; we refer to the group as `FIXD`.

3.3 Measuring Compositionality

Compositionality of an expression is often approximated by comparing the "context" of the expression with the contexts of its constituents. We measure the degree of compositionality of a target verb+noun combination, $t = \prec v, n \succ$, in a similar fashion.

We take the context of the target (t) and each of its constituents (v and n) to be a vector of the frequency of nouns cooccurring with it within a window of ± 5 words. We then measure the "similarity" between the target and each of its constituents, $\mathrm{Sim_{dist}}\,(t,\ v)$ and $\mathrm{Sim_{dist}}\,(t,\ n)$, using the cosine measure.[1]

Recall that an LVC can be roughly paraphrased by a verb that is morphologically related to its noun constituent, e.g., *to make a decision* nearly means *to decide*. For each target t, we thus add a third measure, $\mathrm{Sim_{dist}}\,(t,\ rv)$, where rv is a verb morphologically related to the noun constituent of t, and is automatically extracted from WordNet (Fellbaum, 1998).[2]

We use abbreviation `COMP` to refer to the group of measures related to compositionality.

3.4 The Constituents

Recall that semantic properties of the constituents of a verb+noun combination are expected to be relevant to its semantic class. We thus add two simple feature groups: (i) the verb itself (`VERB`); and (ii) the semantic category of the noun according to WordNet (`NSEM`). We take the semantic category of a noun to be the ancestor of its first sense in the hypernym hierarchy of WordNet 2.1, cut at the level of the children

[1] Our preliminary experiments on development data from Fazly and Stevenson (2006) revealed that the cosine measure and a window size of ± 5 words resulted in the best performance.

[2] If no such verb exists, $\mathrm{Sim_{dist}}\,(t,\ rv)$ is set to zero. If more than one verb exist, we choose the one that is identical to the noun or the one that is shorter in length.

of ENTITY (which will include PHYSICAL ENTITY and ABSTRACT ENTITY).[3]

4 Experimental Setup

4.1 Corpus and Experimental Expressions

We use the British National Corpus (BNC),[4] automatically parsed using the Collins parser (Collins, 1999), and further processed with TGrep2.[5] We select our potential experimental expressions from pairs of verb and direct object that have a minimum frequency of 25 in the BNC and that involve one of a predefined list of basic (transitive) verbs. Basic verbs, which in their literal uses refer to states or acts central to human experience (e.g., *give* and *put*), commonly form MWEs in combination with their direct object argument (Cowie et al., 1983). We use 12 such verbs ranked highly according to the number of different nouns they appear with in the BNC. Here are the verbs in alphabetical order:

bring, find, get, give, hold, keep, lose, make, put, see, set, take

To guarantee that the final set of expressions contains pairs from all four classes, we pseudo-randomly select them from the initial list of pairs extracted from the BNC as explained above. To ensure the inclusion of IDMs, we consult two idioms dictionaries (Cowie et al., 1983; Seaton and Macaulay, 2002). To ensure we include LVCs, we select pairs in which the noun has a morphologically related verb according to WordNet. We also select pairs whose noun is not morphologically related to any verb to ensure the inclusion of LIT combinations.

This selection process resulted in 632 pairs, reduced to 563 after annotation (see Section 4.2 for details on annotation). Out of these, 148 are LIT, 196 are ABS, 102 are LVC, and 117 are IDM. We randomly choose 102 pairs from each class as our final experimental expressions. We then pseudo-randomly divide these into training (TRAIN), development (DEV), and test (TEST) data sets, so that each set has an equal number of pairs from each class. In addition, we ensure that pairs with the same verb that belong to the same class are divided equally among the three sets. Our final TRAIN, DEV, and TEST sets contain 240, 84, and 84 pairs, respectively.

4.2 Human Judgments

We asked four native speakers of English with sufficient linguistic background to annotate our experimental expressions. The annotation task was expected to be time-consuming, hence it was not feasible for all the judges to annotate all the expressions. Instead, we asked one judge to be our primary annotator, PA henceforth. (PA is an author of this paper, but the other three judges are not.)

First, PA annotated all the 632 expressions selected as described in Section 4.1, and removed 69 of them that could be potential sources of disagreement for various reasons (e.g., if an expression was unfamiliar or was likely to be part of a larger phrase). Next, we divided the remaining 563 pairs into three equal-sized sets, and gave each set to one of the other judges to annotate. The judges were given a comprehensive guide for the task, in which the classes were defined solely in terms of their semantic properties. Since expressions were annotated out of context (type-based), we asked the judges to annotate the predominant meaning of each expression.

We use the annotations of PA as our gold standard for evaluation, but use the annotations of the others to measure inter-annotator agreement. The observed agreement (p_o) between PA and each of the other three annotators are 79.8%, 72.2%, and 67%, respectively. The kappa (κ) scores are .72, .62, and .56. The reasonably high agreement scores confirm that the classes are coherent and linguistically plausible.

4.3 Classification Strategy and Features

We use the decision tree induction system C5.0 as our machine learning software, and the measures proposed in Section 3 as features in our classification experiments.[6] We explore the relevance of each feature group in the overall classification, as well as in identifying members of each individual class.

5 Experimental Results

We performed experiments on DEV to find features most relevant for classification. These experiments

[3] Experiments on development data show that looking at all senses of a noun degrades performance.

[4] http://www.natcorp.ox.ac.uk.

[5] http://tedlab.mit.edu/~dr/Tgrep2.

[6] Experiments on DEV using a Support Vector Machine algorithm produced poorer results; we thus do not report them.

revealed that removing $\mathrm{Sim}_{\mathrm{dist}}(t, v)$ resulted in better performance. This is not surprising given that basic verbs are highly polysemous, and hence the distributional context of a basic verb may not correspond to any particular sense of it. We thus remove this feature (from COMP) in experiments on TEST. Results presented here are on the TEST set; those on the DEV set have similar trends. Here, we first look at the overall performance of classification in Section 5.1. Section 5.2 presents the results of classification for the individual classes.

5.1 Overall Classification Performance

Table 3 presents the results of classification—in terms of average accuracy ($\%Acc$) and relative error reduction ($\%RER$)—for the individual feature groups, as well as for all groups combined. The baseline (chance) accuracy is 25% since we have four equal-sized classes in TEST. As can be seen, INST features yield the lowest overall accuracy, around 36%, with a relative error reduction of only 14% over the baseline. This shows that institutionalization, although relevant, is not sufficient for distinguishing among different levels of semantic idiosyncrasy. Interestingly, FIXD features achieve the highest accuracy, 50%, with a relative error reduction of 33%, showing that fixedness is a salient aspect of semantic idiosyncrasy. COMP features achieve reasonably good accuracy, around 40%, though still notably lower than the accuracy of FIXD features. This is especially interesting since much previous research has focused solely on the non-compositionality of MWEs to identify them (McCarthy et al., 2003; Baldwin et al., 2003; Bannard et al., 2003). Our results confirm the relevance of this property, while at the same time revealing its insufficiency. Interestingly, features related to the semantic properties of the constituents, VERB and NSEM, overall perform comparably to the compositionality features. However, a closer look at their performance on the individual classes (see Section 5.2) reveals that, unlike COMP, they are mainly good at identifying items from certain classes. As hypothesized, we achieve the highest performance, an accuracy of 58% and a relative error reduction of 44%, when we combine all features.

Only the features in group	$\%Acc$	($\%RER$)
INST	35.7	(14.3)
FIXD	50	(33.3)
COMP	40.5	(20.7)
VERB	42.9	(23.9)
NSEM	39.3	(19.1)
ALL	**58.3**	(**44.4**)

Table 3: Accuracy ($\%Acc$) and relative error reduction ($\%RER$) over TEST pairs, for the individual feature groups, and for all features combined.

All features except those in group	$\%Acc$	($\%RER$)
INST	53.6	(38.1)
FIXD	47.6	(30.1)
COMP	56	(41.3)
VERB	48.8	(31.7)
NSEM	46.4	(28.5)
ALL	**58.3**	(**44.4**)

Table 4: Accuracy ($\%Acc$) and relative error reduction ($\%RER$) over TEST pairs, removing one feature group at a time.

Table 4 displays classification performance, when we use all the feature groups except one. These results are more or less consistent with those in Table 3 above, except some differences which we discuss below. Removing FIXD features results in a drastic decrease in performance (10.7%), while the removal of INST and COMP features cause much smaller drops in performance (4.7% and 2.3%, respectively). Here again, we can see that features related to the semantics of the verb and the noun are salient features. Removing either of these results in a substantial decrease in performance—9.5% and 11.9%, respectively—which is comparable to the decrease resulting from removing FIXD features. This is an interesting observation, since VERB and NSEM features, on their own, do not perform nearly as well as FIXD features. It is thus necessary to futher investigate the performance of these groups on larger data sets with more variability in the verb and noun constituents of the expressions.

5.2 Performance on Individual Classes

We now look at the performance of the feature groups, both separately and combined, on the individual classes. For each combination of class and feature group, the F-measures of classification are given in Table 5, with the two highest F-measures for each class shown in boldface.[7] These results show that the combination of all feature groups yields the best or the second-best performance on all four classes. (In fact, in only one case is the performance

[7]Our F-measure gives equal weights to precision and recall.

Class	Only the features in group INST	FIXD	COMP	VERB	NSEM	ALL
LIT	.48	.42	.51	.54	**.57**	**.60**
ABS	.40	.32	.17	.27	**.49**	**.46**
LVC	.21	**.58**	.47	.55	-	**.68**
IDM	.33	**.67**	.42	0	-	**.56**

Table 5: F-measures on TEST pairs, for individual feature groups and all features combined.

	ANNOTATOR$_1$		ANNOTATOR$_2$		ANNOTATOR$_3$	
Class	$\%p_o$	κ	$\%p_o$	κ	$\%p_o$	κ
LIT	93.6	.83	88.3	.67	91.4	.78
ABS	83	.63	76.6	.46	78	.52
LVC	91	.71	83	.54	87.7	.61
IDM	92	.73	87.2	.63	87.2	.59

Table 6: Per-class observed agreement and kappa score between PA and each of the three annotators.

of ALL features notably smaller than the best performance achieved by a single feature group.)

Looking at the performance of ALL features, we can see that we get reasonably high F-measure for all classes, except for ABS. The relatively low values of p_o and κ on this class, as shown in Table 6, suggest that this class was also the hardest to annotate. It is possible that members of this class share properties with other classes. The extremely poor performance of the COMP features on ABS also reflects that perhaps members of this class are not coherent in terms of their degree of compositionality (e.g, compare *give confidence* and *make a living*). In the future, we need to incorporate more coherent membership criteria for this class into our annotation procedure.

According to Table 5, the most relevant feature group for identifying members of the LIT and ABS classes is NSEM. This is expected since NSEM is a binary feature determining whether the noun is a PHYSICAL ENTITY or an ABSTRACT ENTITY.[8] Among other feature groups, INST features also perform reasonably well on both these classes. The most relevant feature group for LVC and IDM is FIXD. (Note that for IDM, the performance of this group is notably higher than ALL). On the other hand, INST features have a very poor performance on these classes, reinforcing that IDMs and LVCs may not necessarily appear with significantly high frequency of occurrence in a given corpus. Fixedness features thus prove to be particularly important for the identification of highly idiosyncratic MWEs, such as LVCs and IDMs.

6 Related Work

Much recent work on classifying MWEs focuses on determining different levels of compositionality in verb+particle combinations using a measure of distributional similarity (McCarthy et al., 2003; Baldwin et al., 2003; Bannard et al., 2003). Another group of research attempts to classify a particular MWE subtype, such as verb-particle constructions (VPCs) or LVCs, according to some fine-grained semantic criteria (Wanner, 2004; Uchiyama et al., 2005; Cook and Stevenson, 2006). Here, we distinguish subtypes of MWEs that are defined according to coarse-grained distinctions in their degree of semantic idiosyncrasy.

Wermter and Hahn (2004) recognize the importance of distinguishing MWE subtypes that are similar to our four classes, but only focus on separating MWEs as one single class from literal combinations. For this, they use a measure that draws on the limited modifiability of MWEs, in addition to their expected high frequency. Krenn and Evert (2001) attempt to separate German idioms, LVCs, and literal phrases (of the form verb+prepositional phrase). They treat LVCs and idioms as institutionalized expressions, and use frequency and several association measures, such as PMI, for the task. The main goal of their work is to find which association measures are particularly suited for identifying which of these classes. Here, we look at properties of MWEs other than their institutionalization (the latter we quantify using an association measure).

The work most similar to ours is that of Venkatapathy and Joshi (2005). They propose a minimally-supervised classification schema that incorporates a variety of features to group verb+noun combinations according to their level of compositionality. Their work has the advantage of requiring only a small amount of manually-labeled training data. However, their classes are defined on the basis of compositionality only. Here, we consider classes that are linguistically salient, and moreover need special treatment within a computational system. Our work is also different in that it brings in a new group of features, the fixedness measures, which prove to be very effective in identifying particular classes of MWEs.

[8] Since this is a binary feature, it can only distinguish two classes. In the future, we need to include more semantic classes.

7 Conclusions

We have provided an analysis of the important characteristics pertaining to the semantic idiosyncrasy of MWEs. We have also elaborated on the relationship between these properties and four linguistically-motivated classes of verb+noun combinations, falling on a continuum from less to more semantically idiosyncratic. On the basis of such analysis, we have developed statistical, corpus-based measures that quantify each of these properties. Our results confirm that these measures are effective in type classification of the MWEs under study. Our class-based results look into the interaction between the measures (each capturing a property of MWEs) and the classes (which are defined in terms of semantic idiosyncrasy). Based on this, we can see which measures—or properties they relate to—are most or least relevant for identifying each particular class of verb+noun combinations. We are currently expanding this work to investigate the use of similar measures in token classification of verb+noun combinations in context.

Acknowledgements

We thank Eric Joanis for providing us with NP-head extraction software. We thank Saif Mohammad for the CooccurrenceMatrix and the DistributionalDistance packages.

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proc. of ACL-SIGLEX Wkshp. on Multiword Expressions*, 89–96.

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proc. of ACL-SIGLEX Wkshp. on Multiword Expressions*, 65–72.

Laurel J. Brinton and Minoji Akimoto, eds. 1999. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. John Benjamins.

Miriam Butt. 2003. The light verb jungle. Workshop on Multi-Verb Constructions.

Cristina Cacciari and Patrizia Tabossi, eds. 1993. *Idioms: Processing, Structure, and Interpretation*. Lawrence Erlbaum Associates, Publishers.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1991. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, 115–164.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, UPenn.

Paul Cook and Suzanne Stevenson. 2006. Classifying particle semantics in English verb-particle constructions. In *Proc. of COLING-ACL'06 Wkshp. on Multiword Expressions*, 45–53.

Anthony P. Cowie, Ronald Mackin, and Isabel R. McCaig. 1983. *Oxford Dictionary of Current Idiomatic English*, volume 2. OUP.

Anthony P. Cowie. 1981. The treatment of collocations and idioms in learner's dictionaries. *Applied Linguistics*, II(3):223–235.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proc. of EACL'06*, 337–344.

Christiane Fellbaum, editor. 1998. *WordNet, An Electronic Lexical Database*. MIT Press.

Raymond W., Jr. Gibbs and Nandini P. Nayak. 1989. Psychololinguistic studies on the syntactic behaviour of idioms. *Cognitive Psychology*, 21:100–138.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proc. of ACL'01 Wkshp. on Collocations*, 39–46.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. of COLING-ACL'98*, 768–774.

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

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

Tadao Miyamoto. 2000. *The Light Verb Construction in Japanese: the Role of the Verbal Noun*. John Benjamins.

Rosamund Moon. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford University Press.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proc. of CICLing'02*, 1–15.

Maggie Seaton and Alison Macaulay, eds. 2002. *Collins COBUILD Idioms Dictionary*. HarperCollins.

Kiyoko Uchiyama, Timothy Baldwin, and Shun Ishizaki. 2005. Disambiguating Japanese compound verbs. *Computer Speech and Language*, 19:497–512.

Sriram Venkatapathy and Aravind Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proc. of HLT-EMNLP'05*, 899–906.

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

Joachim Wermter and Udo Hahn. 2004. Collocation extraction based on modifiability statistics. In *Proc. of COLING'04*, 980–986.