A Distributional Account of the Semantics of Multiword Expressions

Afsaneh Fazly and Suzanne Stevenson
Department of Computer Science, University of Toronto
6 King's College Road, Toronto, ON M5S 3H5, Canada
{afsaneh,suzanne}@cs.toronto.edu

Abstract

The lexical status of multiword expressions (MWEs), such as *make a decision* and *shoot the breeze*, has long been a matter of debate. Although MWEs behave much like phrases on the surface, it has been argued that they should be treated like words because their components together form a single unit of meaning. However, MWEs are not a homogeneous lexical category, but rather can have distinct semantic and syntactic properties. For example, the overall meaning of an MWE may vary in how much it diverges from the combined contribution of its constituent parts, with *make a decision*, e.g., having a strong relation to *decide*, while *shoot the breeze* is entirely idiomatic. In order to understand whether and how MWEs should be represented in a (computational) lexicon, it is necessary to look into the relationship between the underlying semantic properties of these expressions and their surface behaviour. We examine several properties of MWEs pertaining to their semantic idiosyncrasy, and relate them to the distributional behaviour of MWEs in their actual usages. Accordingly, we propose statistical measures for quantifying each property, which we then use for separating different types of MWEs that require different treatment within a lexicon.

1 Introduction

Multiword expressions (MWEs) are linguistic constructions formed from the combination of multiple words that together convey a single new meaning. For example, *make a decision* is a light verb construction which roughly means "to decide", and *shoot the breeze* is an idiomatic expression meaning "to chat idly". MWEs, such as idioms and light verb constructions, are of great interest to linguists, psycholinguists, and lexicographers, because of their plenitude in language, their peculiar syntactic and semantic behaviour, and their unclear lexical status (Jackendoff, 1997; Moon, 1998; Pauwels, 2000; Fellbaum, 2006). On the one hand, MWEs are

semantically idiosyncratic—i.e., they have a meaning that diverges from the combined contribution of their constituent parts—hence they should be included in a lexicon along with their idiosyncratic meaning. On the other hand, MWEs are often morphologically and/or syntactically flexible—i.e., they behave like any syntactic phrase composed of multiple words—and thus cannot be simply listed in a lexicon as "words with spaces" (Sag et al., 2002). More importantly, MWEs do not form a homogeneous category, rather they can have varying semantic and syntactic properties. Particularly, different MWEs may involve different degrees of semantic idiosyncrasy (e.g., cf. *make a decision* and *shoot the breeze*), and/or exhibit varying degrees of syntactic flexibility (e.g., cf. *spill the beans* and *kick the bucket*).[1]

It is clear that not all MWEs can be treated the same when it comes to their representation in a lexicon. Instead, we need to look closely into the distinctive properties of different types of MWEs that might lead to different lexical representations for them. This article attempts to address some of the issues surrounding the lexical representation of MWEs, which also affect their treatment within a computational system. Specifically, we look into the relationship between the surface behaviour of MWEs in use and their underlying semantic properties, and accordingly develop techniques for automatically determining the type (class) of a given expression.

The article is organized as follows: First, in §2, we identify several classes of MWEs on the basis of their degree of semantic idiosyncrasy, and argue that each class requires a different encoding in a (computational) lexicon. Next, in §3, we expound on some of the linguistic properties that are known to be related to the semantic idiosyncrasy of a multiword expression, and hence are expected to be useful in determining its semantic class. In this section, we also propose techniques for modelling these properties with patterns of usage of the expressions (i.e., their distributional behaviour) derived from text. §4 provides a multi-faceted evaluation of the proposed statistical usage-based measures, and §5 concludes the paper.

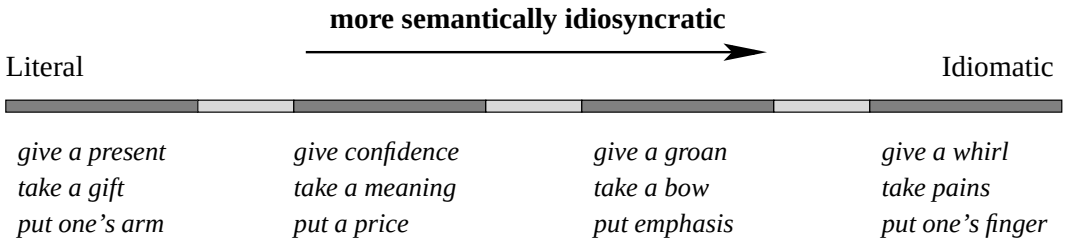

Figure 1: The projection of the four semantic classes of expressions on the semantic transparency continuum.

2 Semantic Classes of Expressions

Expressions composed of multiple words involve different degrees of semantic idiosyncrasy. A literal combination, such as *give a present*, has a fully transparent meaning, whereas an idiomatic expression, such as *give a whirl* (meaning "to try") has a largely opaque semantics. There are also expressions with in-between levels of semantic idiosyncrasy/transparency, such as *give confidence* (referring to an abstract transfer, as opposed to a physical transfer), and *give a groan* (roughly meaning "to groan"). Although semantic idiosyncrasy is a matter of degree, linguists have often identified coherent classes of expressions that have shared properties within each class and differing properties across the classes. Many properties contribute to the degree of semantic idiosyncrasy of an expression, e.g., degree of figurativeness and/or degree of semantic compositionality of the expression. The space of MWEs can thus be viewed as a multidimensional space, with each dimension corresponding to one such property. For the ease of exposition, Figure 1 presents a linear projection of this rather complex space, where the classes are arranged along a "continuum" from fully transparent literal expressions to largely opaque idiomatic combinations. In the following paragraphs, we elaborate on the specific properties of members of each of these classes, and argue for their distinct treatment within a lexicon.

One broadly-documented class of semantically idiosyncratic expressions is that of idioms (such as *give a whirl*) whose meanings are typically not directly related to the meanings of their constituents. Thus a lexicographer may decide that idioms are best treated as (multiword) lexical units that should be listed in a lexicon along with their idiomatic meaning. As we will

see below in §3, in addition to having highly idiosyncratic meanings, idioms are also known to be largely non-productive and to be at least somewhat restricted with respect to the syntactic forms they appear in. For a computational system to appropriately understand and use idioms, information about the lexical and syntactic restrictions of each idiom must also be included in its lexical representation.

Another linguistically well-known class of MWEs is the class of light verb constructions or LVCs, such as *give a groan*. LVCs are also considered to be semantically idiosyncratic because the verb component does not contribute much of its "basic" meaning — e.g., in *give a groan*, *give* does not mean "transfer of possession". Nonetheless, LVCs differ from idioms in that they are semantically more transparent because of a strong semantic connection to the noun constituent — e.g., *give a groan* can be roughly paraphrased by *groan*. Even though the meaning of LVCs is somewhat predictable, they are often considered as multiword predicates with special argument structure (influenced by the argument structures of both constituents). Such information is necessary for understanding and appropriately using LVCs, and hence can be seen as an important part of an LVC's lexical representation in a computational system.

Idioms and LVCs are clearly distinct from similar-on-the-surface literal phrases such as *give a present*, which has a fully transparent meaning constructed from a compositional combination of its constituent semantics. Due to the compositional meaning of literal phrases, and also their high degree of productivity, these phrases are not included in a lexicon.

Still, there are many expressions that exhibit some degree of semantic idiosyncrasy, such as *give confidence* and *put a price (on something)*. These expressions often take on extra connotations beyond the fully compositional combination of their constituent meanings. For example, in *give confidence*, *give* contributes an abstract meaning that is different from (though metaphorically related to) its basic "transfer of possession" meaning. Speakers may understand such expressions by analogy and through establishing metaphorical connections between the basic and the extended meanings of the constituent words (Lakoff & Johnson, 1980; Newman, 1996)—here, "transfer of control over a psychological feature" in analogy with "transfer of possession of a physical entity". Many such expressions are thus considered as "collocations"

which should be treated specially when it comes to their representation and automatic processing in a computational system. As in a collocation, in an abstract combination such as *give confidence*, a base word (here, *confidence*) combines with a collocate selected from a restricted set of words (here, *give* but not, e.g., *grant*).[2]

To summarize, we have identified four classes of expressions on the continuum of semantic idiosyncrasy, and have argued that they need different encodings in a lexicon. The classes are: idiomatic expressions (IDM), light verb constructions (LVC), abstract combinations (ABS), and literal phrases (LIT).[3]

Many of these types of MWEs are formed from the combination of a verb with one or more of its arguments. It is especially common for certain highly frequent verbs to combine with a noun in their direct object position to form MWEs such as the ones in Figure 1 above (Cowie et al., 1983; Nunberg et al., 1994; Pauwels, 2000; Newman & Rice, 2004; Fellbaum, 2007). In our study, we thus focus on such MWEs (in English), which we refer to as verb+noun combinations. All analyses presented in this article are on expression types, as opposed to their specific usages (tokens) in context. We assume that each expression (type) has a dominant meaning shared by many speakers of the language. In the rest of the article, we provide evidence that people can successfully identify verb+noun types as belonging to one of the four identified semantic classes, solely on the basis of their semantic properties. We also show that despite their surface similarities, it is possible to (automatically) distinguish among members of these classes by looking at statistics over their distributional behaviour in text.

3 Linguistic Properties and Statistical Measures

In this section, we discuss some of the linguistic properties of MWEs that are known to be connected to the semantic idiosyncrasy of these expressions. We also propose simple statistical measures that model each property based on frequencies collected from actual usages of the expressions in text.

3.1 Institutionalization

Institutionalization is the process through which a combination of words becomes accepted as a conventional semantic unit (with some degree of semantic idiosyncrasy). In contrast to LIT phrases, IDMs, LVCs, and ABS combinations are expected to be institutionalized to some extent. Moreover, we expect there to be a positive correlation between the degree of semantic idiosyncrasy of MWEs and their degree of institutionalization.

Many corpus-based approaches assess the degree of institutionalization of an expression using its frequency of occurrence in text. In the case of MWEs, observed frequencies are not reliable on their own, as many word sequences may appear frequently by chance, simply due to their components being highly frequent. In many natural language processing (NLP) applications, measures of strength of association between the constituents of an expression are used instead. Here, we use both the frequency of occurrence and an association measure, the pointwise mutual information or PMI (Church et al., 1991). PMI of a verb+noun combination $v+n$ measures the degree of association between v and n, by calculating the proportion of their joint co-occurrence probability (as observed in text) relative to the probability of the two appearing together due to chance (given their individual probabilities of occurrence), as in:

$$\begin{aligned} \mathrm{PMI}(v,n) &\doteq \log \frac{P(v,n)}{P(v)\,P(n)} \\ &\approx \log \frac{\mathit{freq}(\mathit{any\_verb}+\mathit{any\_noun}) \times \mathit{freq}(v+n)}{\mathit{freq}(v+\mathit{any\_noun}) \times \mathit{freq}(\mathit{any\_verb}+n)} \end{aligned} \tag{1}$$

where $\mathit{freq}(\mathit{any\_verb}+\mathit{any\_noun})$ is the total frequency of all verb+object combinations, $\mathit{freq}(v+n)$ is the frequency of v and n co-occurring in a verb–object relation, $\mathit{freq}(v+\mathit{any\_noun})$ is the frequency of v co-occurring with any noun in its direct object position, and $\mathit{freq}(\mathit{any\_verb}+n)$ is the frequency of n co-occurring as the direct object of any verb. All frequency counts are calculated using the British National Corpus (Burnard, 2000) containing 100 million words in total.

3.2 Lexicosyntactic Fixedness

Lexicosyntactic fixedness refers to some degree of lexical or syntactic restrictiveness in a semantically idiosyncratic expression. Below we explain various kinds of fixedness and propose measures for each. For a more comprehensive discussion on fixedness and its relation to semantic idiosyncrasy, see Fazly (2007).

Lexical Fixedness. An MWE is lexically fixed if the substitution of a semantically similar word for any of its constituents does not preserve its original (idiosyncratic) meaning, e.g., *shoot the wind* does not have an idiomatic meaning like the similar expression *shoot the breeze*, even though *wind* and *breeze* are semantically similar. This is in contrast to literal expressions which are lexically flexible, e.g., *give a gift* and *give a present* have very similar meanings.[4] There is evidence that semantically idiosyncratic expressions, such as IDMs, LVCs, and ABS combinations, exhibit lexical fixedness to some extent. Moreover, the more idiosyncratic an expression, the more the degree of lexical fixedness it exhibits (Gibbs et al., 1989; Nunberg et al., 1994).

Ideally, we want a measure that compares the degree of idiosyncrasy of a target verb+noun combination with the idiosyncrasy of semantically similar expressions that result from the substitution of the verb or the noun (referred to as the target's "lexical variants"). But semantic idiosyncrasy is what we are trying to measure using surface cues such as institutionalization and fixedness. As a simplification, we thus assume a target $v+n$ to be lexically fixed if it generally occurs much more frequently than its lexical variants. Accordingly, we propose a measure, $\text{Fixedness}_{\text{lex}}(v+n)$, which quantifies the degree of lexical fixedness of $v+n$ by comparing its strength of association, measured by PMI, with the average PMI of its variants:

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

$\text{Fixedness}_{\text{lex}}(v+n)$ indicates how far and in which direction the PMI of the target $(v+n)$ deviates from the average PMI of the target and all its variants ($\overline{\text{PMI}}$), expressed in units of the standard deviation of the PMI values (*std*, which measures the spread of the values). Note that, here, one can say that the PMI of a verb+noun combination is used as an indirect clue

to its degree of idiosyncrasy (as also done by Lin, 1999). We look at the difference between the PMI of the target, and the average PMI, so we rely on the collective evidence rather than individual cases. We normalize this difference by dividing it by the standard deviation, so we get scores that are comparable across verb+noun combinations. The variants of the target $v+n$ are automatically generated by replacing either v or n with a semantically (and syntactically) similar word taken from the automatically-built thesaurus of Lin (1998). Examples of automatically generated variants for the combination *spill+bean* are: *pour+bean, stream+bean, spill+corn,* and *spill+rice.*

Syntactic Fixedness. An MWE is syntactically fixed if it cannot undergo syntactic variations and at the same time retain its original semantic interpretation. IDMs are known to show strong preferences for the syntactic patterns they appear in (Cacciari & Tabossi, 1993) — e.g., compare *Tim kicked the bucket* with **Tim kicked the buckets* (in the idiom reading). Many LVCs also have a tendency of mostly appearing in preferred syntactic forms (Brinton & Akimoto, 1999) — e.g., compare *Joe gave a groan* with ?*A groan was given by Joe*. As a simplification of the above definition for syntactic fixedness, we assume a target $v+n$ to be syntactically fixed if it occurs mainly in a few fixed syntactic forms (as is the case with most IDMs and many LVCs). Our proposed measure, $\text{Fixedness}_{\text{syn}}$, thus quantifies the degree of syntactic fixedness of a verb+noun combination, by comparing its behaviour in text with the behaviour of a typical verb+object combination, both defined as probability distributions over a predefined set of syntactic patterns. We use a standard information-theoretic measure, KL-divergence, to calculate the divergence between the two distributions as follows:

$$\begin{aligned} \text{Fixedness}_{\text{syn}}(v+n) &\doteq D(P(pt|v+n)\,||\,P(pt)) \\ &= \sum_{pt_k \in \mathcal{P}} P(pt_k|v+n) \log \frac{P(pt_k|v+n)}{P(pt_k)} \end{aligned} \tag{3}$$

In the above formulation, $P(pt|v+n)$ represents the syntactic behaviour of the target $v+n$ — that is, the distribution of occurrence of $v+n$ over a set of syntactic patterns $pt \in \mathcal{P}$. $P(pt)$ represents the "typical" syntactic behaviour — that is, the distribution of occurrence of any verb+noun combination over the same set of syntactic patterns. The set of patterns, $\mathcal{P}$, contains

Table 1: Patterns used in the syntactic fixedness measure, along with examples for each. v_{act} and v_{pass} stand for active and passive usages of the verb v, respectively; n_{sg} and n_{pl} stand for singular and plural forms of the noun n, respectively; det specifies the determiner type, e.g., det:NULL specifies that no determiner is preceding the noun, and det:DEM and det:POSS specify that a demonstrative or a possessive determiner is used to introduce the noun.

Pattern Signature			Example
v_{act}	det:NULL	n_{sg}	*give money*
v_{act}	det:*a/an*	n_{sg}	*give a book*
v_{act}	det:*the*	n_{sg}	*give the book*
v_{act}	det:DEM	n_{sg}	*give this book*
v_{act}	det:POSS	n_{sg}	*give my book*
v_{act}	det:NULL	n_{pl}	*give books*
v_{act}	det:*the*	n_{pl}	*give the books*
v_{act}	det:DEM	n_{pl}	*give those books*
v_{act}	det:POSS	n_{pl}	*give my books*
v_{act}	det:OTHER	$n_{sg,pl}$	*give many books*
v_{pass}	det:ANY	$n_{sg,pl}$	*a/the/this/my book(s) was/were given*

11 manually-identified syntactic patterns (shown in Table 1) known to be relevant to syntactic fixedness in LVCs and IDMs; see Fazly et al. (to appear) for more discussion on the selection of the patterns. In brief, we consider three types of syntactic variation, namely, passivization, determiner type, and the number of the noun constituent (singular or plural).[5] KL-divergence is a standard information-theoretic measure of the difference between two probability distributions, widely-used in natural language processing applications.

Overall Fixedness. Highly idiosyncratic MWEs, such as IDMs, are known to be both lexically and syntactically fixed for the most part (as also shown in our previous work, Fazly & Stevenson, 2006). Thus, in addition to the lexical and syntactic fixedness measures, we use a measure of the overall fixedness of a verb+noun, $\text{Fixedness}_{\text{overall}}(v+n)$, that combines the two types of fixedness into a single measure, and is defined as:

$$\text{Fixedness}_{\text{overall}}(v+n) \doteq \alpha\,\text{Fixedness}_{\text{syn}}(v+n) + (1-\alpha)\,\text{Fixedness}_{\text{lex}}(v+n) \quad (4)$$

where α weights the relative contribution of lexical and syntactic fixedness in predicting semantic idiosyncrasy.[6]

Dominant Pattern. Even for MWEs which are syntactically fixed, different types of MWEs may prefer different syntactic patterns. For example, most LVCs are known to prefer the pattern in which the noun is introduced by the indefinite article *a/an* (as in *give a groan* and *make a decision*), whereas this is not the case with IDMs (e.g., *shoot the breeze* and *pull one's weight*). We thus use the dominant (most frequent) pattern of occurrence of a target verb+noun, $\text{Pattern}_{\text{dom}}(v+n)$, as a clue to its class membership.

Fixedness and Bias in Adjectival Modification. Semantic idiosyncrasy in verb+noun combinations is also argued to affect the modifiability of the noun constituent (Cacciari & Tabossi, 1993; Brinton & Akimoto, 1999). IDMs, for example, are known to consistently appear either with an adjective, as in *keep an open mind* (cf. ?*keep a mind*), or without one, as in *shoot the breeze* (cf. ?*shoot the fun breeze*). This is in contrast to most LIT, LVC, and ABS combinations, which tend to appear both with and without adjectival modifiers. We define a measure, $\text{Fixedness}_{\text{adj}}$, which quantifies the degree of fixedness of a target verb+noun with respect to adjectival modification of the noun constituent. As in the syntactic fixedness measure, here we use the KL-divergence between two distributions, reflecting the modifiability of a target verb+noun and that of a typical verb+noun, respectively, as in:

$$\begin{aligned} \text{Fixedness}_{\text{adj}}(v+n) &\doteq D(P(a|v+n)\,||\,P(a)) \\ &= \sum_{a_i \in \mathcal{A}} P(a_i|v+n) \log \frac{P(a_i|v+n)}{P(a_i)} \end{aligned} \tag{5}$$

where $\mathcal{A}$ is a set containing two patterns that mark the presence (PRES) or absence (ABS) of an adjectival modifier preceding the noun: "$v_{\{act,pass\}}$ det:ANY adj:PRES $n_{\{sg,pl\}}$" (e.g., *give a/the/this/my red book(s)*), and "$v_{\{act,pass\}}$ det:ANY adj:ABS $n_{\{sg,pl\}}$" (e.g., *give a/the/this/my book(s)*), respectively.

LVCs, although not fixed with respect to adjectival modification, are often argued to have a tendency of frequently appearing with an adjective modifying their noun constituent (Nickel,

1968; Brinton & Akimoto, 1999). To capture this tendency, we devise a measure that looks into the proportion of the likelihoods of a verb+noun appearing with and without an adjective:

$$\text{Tendency}_{\text{adj}}(v+n) \doteq \frac{P(a_i = \text{PRES}|v+n)}{P(a_i = \text{ABS}|v+n)} \quad (6)$$

3.3 Non-compositionality

Many MWEs are non-compositional to some extent, i.e., the meaning of the expression deviates from the meaning emerging from a word-by-word interpretation of it. Nonetheless, different types of MWEs involve different degrees of (non-)compositionality: IDMs are largely non-compositional; LVCs are semi-compositional since their meaning can be mainly predicted from the noun constituent; ABS and LIT combinations are expected to be compositional for the most part.

To automatically measure the degree of compositionality of an expression, we need a way of approximating the meaning of words (and word sequences) from their usages in text. Often, the context of a word (or an expression) is taken to be highly informative about its meaning (Firth, 1957; McDonald & Ramscar, 2001). Indeed, the context of a word or expression is known to affect (modulate) the meaning of the word/expression, reflecting its usefulness in determining similarity in meaning: words/expressions are considered to have similar or related meanings if they appear in similar contexts (the "distributional hypothesis"; Harris, 1954).

By definition, the meaning of a compositional expression is mainly derived from the meanings of its component parts. We thus expect the context of a compositional expression to be similar to those of its individual constituents. The degree of compositionality of a target $v+n$ can thus be measured by comparing its context to those of its constituents v and n (as in McCarthy et al., 2003; Bannard et al., 2003). We take a similar approach here, where we define the context of a word (or an expression) to be a vector of the frequency of the nouns co-occurring with it within a window of ± 5 words. We then measure the similarity between $v+n$ and each of its constituents (v and n separately) by measuring the proximity of the corresponding context vectors,[7] and refer to them as $\text{Sim}_{\text{dist}}(v+n, v)$ and $\text{Sim}_{\text{dist}}(v+n, n)$, respectively.

Recall that an LVC can be roughly paraphrased by a verb that is morphologically related to

Table 2: Statistical measures capturing institutionalization, fixedness, and compositionality.

Measure Group	Abbreviated Name	Individual Measures
Institutionalization	INST	$\text{Freq}(v+n)$
		$\text{PMI}(v, n)$
Fixedness	FIXD	$\text{Fixedness}_{\text{lex}}(v+n)$
		$\text{Fixedness}_{\text{syn}}(v+n)$
		$\text{Fixedness}_{\text{overall}}(v+n)$
		$\text{Pattern}_{\text{dom}}(v+n)$
		$\text{Fixedness}_{\text{adj}}(v+n)$
		$\text{Tendency}_{\text{adj}}(v+n)$
Compositionality	COMP	$\text{Sim}_{\text{dist}}(v+n, v)$
		$\text{Sim}_{\text{dist}}(v+n, n)$
		$\text{Sim}_{\text{dist}}(v+n, rv)$

its noun constituent, e.g., *to make a decision* roughly means *to decide*. For each target $v+n$, we thus add a third measure, $\text{Sim}_{\text{dist}}(v+n, rv)$, where rv is a verb morphologically related to n, and is automatically extracted from WordNet (Fellbaum, 1998).[8]

Table 2 summarizes the three groups of measures introduced in Sections 3.1– 3.3, respectively: institutionalization (INST), fixedness (FIXD), and compositionality (COMP).

4 Evaluation of the Statistical Measures

In §3, we have discussed several linguistic properties of MWEs pertaining to their semantic idiosyncrasy, and have proposed statistical measures for capturing each property. Our ultimate goal is to use such measures for the automatic acquisition of syntactic and semantic knowledge about different classes of expressions, as well as for automatically determining the class of a given expression. In this section, we thus assess the goodness of each devised measure (and hence the relevance of the corresponding linguistic property) in separating a set of diverse expressions into different classes.

First, we explain the methodological aspects of our evaluation in §4.1. In §4.2, we examine the scores assigned by each measure to a sample list of diverse expressions from the

four classes to determine how well each measure separates expressions of each class from other classes. Wherever appropriate, we use statistical significance tests to infer conclusions about the general properties of the semantic classes as linguistic constructions. Using a machine learning approach, §4.3 investigates the extent to which the measures can be used to predict the semantic class of a new expression (whose class is unknown), by learning from similar expressions with known class.

4.1 Methodology

For evaluation, we need a list of MWEs whose semantic class is known, as well as a corpus from which to collect frequency counts required by the statistical measures. We use the British National Corpus (BNC, Burnard, 2000), which we have automatically parsed using a parser developed by Collins (1999), and further processed with a parse tree extraction tool (TGrep2, Rohde, 2004). 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 highly frequent transitive verbs (which are known to commonly form MWEs in combination with their direct object argument). We use 12 such verbs ranked highly according to the number of different object nouns they appear with in the BNC. The verbs in alphabetical order are *bring, find, get, give, hold, keep, lose, make, put, see, set, take.* (Even though *have* was ranked at the very top, we exclude it because of its common use as an auxiliary verb.)

A native English speaker annotated the initial list of verb–noun pairs extracted from the BNC, and the quality of the annotations were confirmed by having three other annotators label the same expressions (details of the annotation process can be found in Fazly, 2007). From the annotated list, we randomly choose 102 pairs from each class (LIT, ABS, LVC, and IDM) as our final set of experimental expressions, which we then pseudo-randomly divide into 240 training (TRAIN), 168 test (TEST) pairs. We ensure that each of these three data sets has an equal number of pairs from each class, and that pairs with the same verb belonging to the same class are divided equally among the three sets.

The first part of our evaluation analyzes the scores assigned by each measure to the 240

verb–noun pairs in TRAIN. Specifically, we use each statistical measure to rank the TRAIN pairs, and then compare statistics over the ranks (e.g., mean of the ranks or sum of the ranks assigned to the members of a class) across all classes. For all measures, the rankings are in decreasing order, i.e., items with higher score are placed at the top of the ranked list. We only analyze the ten statistical measures that assign continuous scores which can be converted into ranks; $\text{Pattern}_{\text{dom}}$ is excluded because its measurement scale is nominal and hence cannot be used to rank items. We also use significance tests to infer conclusions about a *population* (all instances of a class, e.g., all English IDMs) from the observed *sample* (only the observed instances, e.g., the 60 IDM pairs in TRAIN). Results are presented in §4.2.

The second part of our evaluation investigates the extent to which the proposed measures can be used to predict, for a new MWE with unknown class, which semantic class it belongs to. We perform a number of experiments, in which we automatically classify the 168 pairs TEST into the four classes of LIT, ABS, LVC and IDM, using the 240 pairs in TRAIN for learning. We use the decision tree induction system C5.0 (`http://www.rulequest.com`) as our machine learning software, and the statistical measures as features. We exclude $\text{Sim}_{\text{dist}}(v+n, v)$ from the classification experiments because it was not found informative in our quantitative data analysis of §4.2. The machine learning (classification) results are given in §4.3.

4.2 Quantitative Data Analysis

For each measure, we first examine the mean ranks (i.e., the average of the ranks assigned to the members of each of the four classes in TRAIN) to see whether they differ significantly across classes. Using the Kruskal-Wallis analysis of variance by ranks (Hollander & Wolfe, 1999; R, 2004), we find that, for all the statistical measures under study, the observed differences in the mean ranks of the four classes are statistically significant ($p < .05$ for $\text{Sim}_{\text{dist}}(v+n, v)$, and $p < .01$ for the other 9 measures). To determine which pairs of classes have significantly different rankings assigned by a given measure, we perform a follow-up test, the multiple comparisons using rank sums (Dunn, 1964).[9] Table 3 gives the significant differences that are found among classes for each of the 10 statistical measures under study. The differences are given in

Table 3: Statistically significant differences found among the classes, as well as the observed trends in the ranks assigned to them by each of the 10 statistical measures.

Group	Measure Name	Significant Differences	Observed Trend in Ranking
INST	Freq	{ LIT }:{ LVC }	LVC ≈ IDM ≈ ABS < LIT
	PMI	{ LIT, ABS }:{ LVC, IDM }	IDM ≈ LVC ≪ ABS < LIT
FIXD	$\text{Fixedness}_{\text{lex}}$	{ LIT }:{ LVC, IDM }	IDM < LVC < ABS ≪ LIT
		{ IDM }:{ LIT, ABS }	
	$\text{Fixedness}_{\text{syn}}$	{ IDM }:{ LIT, ABS, LVC }	IDM ≪ LVC ≈ LIT ≈ ABS
	$\text{Fixedness}_{\text{overall}}$	{ IDM }:{ LIT, ABS, LVC }	IDM ≪ LVC < LIT ≈ ABS
	$\text{Fixedness}_{\text{adj}}$	{ IDM }:{ ABS, LVC }	IDM < LIT ≈ LVC ≈ ABS
	$\text{Tendency}_{\text{adj}}$	{ LVC }:{ LIT, IDM }	LVC < ABS ≪ LIT < IDM
		{ IDM }:{ ABS, LVC }	
COMP	$\text{Sim}_{\text{dist}}(v+n, v)$	—	
	$\text{Sim}_{\text{dist}}(v+n, n)$	{ IDM }:{ LIT, ABS, LVC }	LVC ≈ LIT ≈ ABS ≪ IDM
	$\text{Sim}_{\text{dist}}(v+n, rv)$	{ LVC }:{ LIT, ABS, IDM }	LVC ≪ ABS < IDM < LIT
		{ LIT }:{ ABS, LVC }	

the third column, in the form of "{ $X_1, \cdots, X_m$ }:{ $Y_1, \cdots, Y_n$ }", which should be interpreted as classes X_i ($1 \leq i \leq m$) being found significantly different from classes Y_j ($1 \leq j \leq n$).

The above tests show, for each statistical measure, whether each pair of classes are well separated from each other (e.g., as can be seen in Table 3, Freq separates LIT expressions from LVCs). In addition, we would like to know in which order a measure ranks the classes to see whether this order matches the corresponding linguistic predictions (e.g., whether LVCs are more frequent than LITs). For each measure, we thus calculate, for each target class $T \in$ { LIT, ABS, LVC, IDM }, the sum of the number of pairs from the other classes that are ranked before each pair from T, referring to the sum as U_T (Hollander & Wolfe, 1999). If a measure tends to rank members of class X before members of class Y, U_X will be smaller than U_Y. The values of U_T ($T \in$ { LIT, ABS, LVC, IDM }) for a given measure thus can be used to determine how members of different classes in the observed sample are ranked by that measure (e.g., we observe that for Freq, U_{LVC} is smaller than U_{LIT}, hence we conclude that in the observed data, most LVCs are more frequent than most LITs).

Note that since we do not perform any statistical significance tests on the U_T's, we cannot draw general conclusions about the populations. Instead, we discuss the "observed trends" in the order in which members of the four classes (in the sample) are ranked by a given measure. We thus need to use thresholds on the value of U to decide for a given pair of classes, whether one precedes the other in ranking (as determined by a measure). For the thresholds, we use multiples of the standard deviation of U (the spread of its values, which for our data is close to 1000).[10] For a pair of classes X and Y, if the difference between U_X and U_Y is less than twice the standard deviation, we assume the difference is not notable, and say $X \approx Y$. If the difference is more than twice but less than four times the standard deviation, then we assume the difference is notable, and say $X < Y$. If the difference is larger than four times the standard deviation, we decide that the difference is substantial, and say $X \ll Y$.[11] Observed trends in the ranking of members of the four classes, by each statistical measure, are given in the fourth column of Table 3. In the rest of this section, we first examine the discriminative power of the statistical measures, and then look into the separability of the classes.

Analysis by statistical measure: We now look at the significant differences and the observed trends in ranking, given in Table 3, to examine whether the behaviour of the statistical measures match the linguistic predictions about the corresponding property (as discussed in §3). We examine the measures by group. Looking at the first `INST` measure, Freq, we can see that the three classes of ABS, LVC, and IDM tend to have higher frequency than LITs (column four of Table 3). Nonetheless, only the two classes of LVC and LIT are found to be well separated from each other (column three of Table 3). The behaviour of the other `INST` measure, PMI, fits our prediction in some respects: it separates members of LVC and IDM from the other two classes. Moreover, as we expected, some positive correlation is found between the degree of semantic idiosyncrasy and the value of PMI, as shown by the observed trends in ranking. However, in contrast to our prediction, there is no significant difference between ABS and LIT or between LVC and IDM, with respect to the degree of institutionalization as measured by PMI.

Looking at the `FIXD` measures, we can see that all of them are good at identifying the class of IDM. Also, we can see that generally for $\text{Fixedness}_{\text{lex}}$ and $\text{Fixedness}_{\text{overall}}$, the more

semantically idiosyncratic an expression, the more fixed it is (i.e., these measures tend to rank IDMs before LVCs, and both of these before members of the other two classes). In contrast to our prediction, however, Fixedness$_{syn}$ only separates IDM from the other three classes, but does not distinguish among LVC, ABS and LIT. Another interesting observation is that, in accord with our prediction in §3, Fixedness$_{adj}$ tends to rank IDMs at the top (high degree of fixedness with respect to adjectival modification), but it does not distinguish among the other three classes. Also, as we hypothesized, Tendency$_{adj}$ ranks LVCs high, and also is effective in separating members of this class from IDM and LIT. Note that we find members of LVC to be generally ranked before ABS in the sample. However, according to the significance tests, we cannot conclude that the measure generally separates the two populations from each other.

An interesting observation is that $\text{Sim}_{\text{dist}}(v+n, v)$ does not seem to differentiate among the classes. This is not surprising given that the verbs under study are highly frequent (and highly polysemous), and hence the distributional context of such a verb may not correspond to any of its particular senses. The observed behaviour of $\text{Sim}_{\text{dist}}(v+n, n)$ fits our predictions in one respect: it separates members of IDM from the other three classes, ranking them at the very bottom (low degrees of compositionality). However, in contrast to what we expected, the measure does not seem to distinguish among the other three classes. $\text{Sim}_{\text{dist}}(v+n, rv)$ also has a behaviour that in some aspects matches our predictions: it seems to separate LVCs from the other classes, ranking them at the very top (high similarity between the meaning of an LVC and the verb related to its noun constituent). Unexpectedly, this measure also separates the two classes of LIT and ABS, which might be an accident of the data: we assign a zero similarity score to $\text{Sim}_{\text{dist}}(v+n, rv)$ if the noun constituent of the target $v+n$ does not have a morphologically related verb, and we expect this to be the case for many LIT expressions.

Analysis by class: Now, let us look into the separability of members of each class as determined by the three groups of statistical measures. The class of IDMs seems to be the most distinct class of the four: its members can be distinguished from the other three classes on the basis of two fixedness measures (Fixedness$_{\text{syn}}$, Fixedness$_{\text{overall}}$), and one compositionality measure ($\text{Sim}_{\text{dist}}(v+n, n)$). In addition, many other measures, including PMI and Fixedness$_{\text{lex}}$,

distinguish members of this class from one or two other classes. LVCs can also be distinguished from the other three classes using one of the compositionality measures: $\text{Sim}_{\text{dist}}(v+n, rv)$. In addition, members of this class can be separated from those of LIT and ABS using the two institutionalization measures, and from the class of IDMs using many of the fixedness measures. Overall, the class of LVCs is well-separated from the other classes by the statistical measures. The class of ABS is clearly distinguished from IDMs by all the fixedness measures. This class can also be separated from LVCs using PMI and one compositionality measure, $\text{Sim}_{\text{dist}}(v+n, rv)$. Nonetheless, most of the measures (all of the fixedness measures, as well as one institutionalization measure, PMI, and one compositionality measure, $\text{Sim}_{\text{dist}}(v+n, n)$) confuse ABS with either LIT or LVC. The class of LIT can only be reliably separated from LVC and IDM.[12]

All in all, the most confusing class according to the statistical tests is that of ABS. It is not clear, however, whether this is due to our choice of statistical measures, or an inherent difficulty in identifying abstract combinations in general. The latter reason is likely to be a cause, given that this class has also been the main source of disagreement among our human annotators. Note, however, that the disagreement in our human-annotated (gold-standard) data also has an adverse affect on the outcome of the statistical tests. Thus, further research is needed to provide more information about the linguistic properties of the ABS class, as well as its status in relation to the other classes.

4.3 Using Measures for Prediction

The analyses presented in §4.2 reveal that the statistical measures capturing some of the well-known properties of MWEs are in fact informative about the semantic class of these expressions. In many cases, we find that the behaviour of the measures generally matches the linguistic predictions, and that the measures assign significantly different scores to members of the different classes. Results of the classification experiments presented in this section can help understand whether the measures are also useful in practise, e.g., for predicting the semantic class of a new expression.

Table 4 presents the classification accuracy on the 168 pairs in TEST, for the three measure

Table 4: Classification accuracy (%*Acc*) on TEST, for individual and combined feature groups.

Baseline	INST	FIXD	COMP	INST +FIXD	INST +COMP	FIXD +COMP	INST +FIXD +COMP
25	34	45	40	50	45	45	52

groups, INST, FIXD, and COMP, as well as for combinations of these, e.g., INST+FIXD. Accuracy is measured as the proportion of items classified correctly (assigned to the correct class) to the total number of items classified. Since there are four classes, the chance accuracy (baseline) is 25%. We can see that all three feature groups, as well as their combinations, outperform the baseline, showing their usefulness in semantic class prediction for MWEs.

A close look at the results shows that FIXD is the most informative single group for classification, reinforcing that most MWEs exhibit fixedness of some kind, and that information about their fixedness can be used to acquire knowledge about their underlying semantic properties. On the other hand, INST is the least informative group for MWE classification, revealing that simply looking at the frequency of occurrence of expressions is not sufficient for determining their semantic class. In addition, we can see that the best accuracy (52%) is achieved by combining all three feature groups, reinforcing that collective evidence from various properties of MWEs is beneficial. Although this performance is well above the baseline, it is still substantially lower than human performance in the annotation task (with observed agreement ranging from 67% to 80%). This suggests that we need to look at other properties related to the semantic idiosyncrasy of MWEs to improve their classification (see, e.g., Fazly & Stevenson, 2007, for some additional properties).

We also look into the classification performance of the measure groups (and their combinations) on each of the four semantic classes. Table 5 reports the per-class classification performance using *F*-score (the equivalent of per-class accuracy);[13] the best performance for each class is shown in boldface. As can be seen, for two of the classes, ABS and IDM, the best performance is achieved by combining all features, whereas this is not the case for the other two classes, LIT and LVC.

Table 5: *F*-scores (%*F*) of classification for the individual classes, on TEST.

Class	INST	FIXD	COMP	INST +FIXD	INST +COMP	FIXD +COMP	INST +FIXD +COMP
LIT	42	42	39	**62**	50	55	55
ABS	30	28	14	33	16	21	**39**
LVC	32	54	**59**	49	50	44	51
IDM	32	51	35	58	53	61	**67**

For LIT, combining the two groups with the highest individual performance (INST and FIXD) produces the best results. We can see here that on LIT, the performance of COMP is not much worse than the performance of INST or FIXD. Moreover, our analysis in §4.2 showed the usefulness of COMP features in separating LIT from the other three classes. It is thus not clear to us why the combination of the three feature groups does not yield better results than the combination of INST and FIXD. It remains to be tested whether a different choice for our classification algorithm would produce different results. For LVC, the most relevant feature group is COMP, perhaps because (as seen in §4.2) $\text{Sim}_{\text{dist}}(v+n, rv)$ is very useful in separating members of LVC from those of the other classes.

An interesting observation is that IDM is the easiest class to identify (a reasonably high *F*-score of 67%). This is not surprising, given that, according to the analyses presented in §4.2, many of our statistical measures are very good at identifying members of this class. Moreover, the information provided by the different measure groups seem to be complementary in the case of IDMs: although the performance of the individual feature groups are not generally very high, a very good performance is achieved by combining these. The hardest class to distinguish is ABS, with a low *F*-score of 39%. Again, this is in line with our findings in §4.2, that this class is often confused with the other classes. Nonetheless, we achieve a higher-than-baseline performance by combining evidence from features that do not perform well on their own. Even though this class was also the hardest to annotate for our human annotators, there is still a notable gap between the performance of the classifier and that of humans, signifying the need for further research.

5 Conclusions

We have identified several semantic classes of multiword expressions (MWEs) on the basis of their degree of semantic idiosyncrasy, and have argued that they should be encoded differently in a lexicon. Subsequently, we have examined some of the linguistic properties of MWEs pertaining to their semantic idiosyncrasy, and have proposed statistical measures that capture each property by looking at evidence from the distributional behaviour of the expressions.

Our analysis of the statistical measures reveal that they are in general reasonably good at separating members of the different classes from each other. Nonetheless, some classes are better identified than others: IDMs seem to be more easily distinguished by our measures, confirming their special status as a distinct linguistic phenomenon. ABS combinations, on the other hand, are the hardest to separate, suggesting that they are more complex — perhaps due to their close ties with metaphorical language — and that more research is needed to understand their status in relation to well-established MWEs, such as LVCs and IDMs.

Overall, the trends we find in our data with respect to the role of each statistical measure in identifying members of each class match our hypothesis regarding the corresponding linguistic property. These findings confirm that incorporating linguistic knowledge into statistical distributional methods is beneficial for learning aspects of the semantic properties of MWEs. Nonetheless, the gap between the performance of automatic means and that of human annotators suggests a need for further research, e.g., to refine the definition of the classes, or to draw on other relevant linguistic properties of MWEs.

Acknowledgements

This article is an extended and updated version of a paper that appeared in the proceedings of the ACL 2007 Workshop on A Broader Perspective on Multiword Expressions; we thank the anonymous reviewers of that paper. We also thank Alessandro Lenci and the anonymous reviewer of this article for their insightful recommendations that helped improve the quality of the article. We are grateful to our human annotators for their help with the development of our experimental expressions. We thank Eric Joanis

and Saif Mohammad for providing us with the required software for various aspects of text processing. Our work is financially supported by the Natural Sciences and Engineering Research Council of Canada, the Ontario Graduate Scholarship program, and the University of Toronto.

Notes

[1]For example, the idiom *spill the beans* can appear in a passivized form, as in *The beans were spilt by Mary*, whereas *kick the bucket* usually cannot: Most speakers consider *The bucket was kicked by John* to have a literal interpretation only.

[2]Clearly, abstract combinations or collocations need not be stored in a lexicon the same way as idioms and LVCs are. However, as also pointed out by the anonymous reviewer of this article, such combinations should be treated differently from literal phrases: information about the preferred collocates of the "base" needs to be added to its lexical entry. For example, whereas one can *put a price (on something)*, one usually does not ?*place a price (on something)* — this is in contrast to the appropriateness of literal phrases such as *put/place a book (somewhere)*. Although similar restrictions can be found in literal combinations, we expect the semantic restrictions imposed by the constituents of an abstract combination to be more idiosyncratic.

[3]As mentioned before, we are aware that the borderline between the four identified classes of MWEs may not always be clear-cut, partly because idiomatization is a gradual process, and partly due to ambiguity in meaning. Nonetheless, the high agreement among our human annotators shows that it is reasonable to assume that many expressions can be reliably distinguished in terms of membership in one of these classes.

[4]Other terms, such as *paradigmatic modifiability* and *paradigmatic substitutability*, have been used in the linguistics literature to refer to the lexical flexibility of expressions.

[5]Other types of syntactic variation, such as relativization or the use of *wh*-questions, are also considered to be relevant to syntactic fixedness. Nonetheless, such patterns are expected to have a low frequency; moreover, their automatic extraction is often very hard and hence inaccurate. We thus do not include these in our initial set of patterns, but the fixedness measure itself can accommodate them if it is deemed desirable.

[6]α is a parameter whose value is set empirically based on performance over a held-out data set. Here, we set it to 0.6 as this was found to perform best when applied to the development data set of Fazly (2007).

[7]The proximity of two vectors in Euclidean space is often measured by the cosine of the angle between the two vectors.

[8]If no such verb exists, $\text{Sim}_{\text{dist}}(v+n, rv)$ 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.

[9]We need the multiple comparisons test because the Kruskal-Wallis test does not determine which pairs of

classes have significantly different rankings, only that some significant difference exists. Note, however, that we need to perform the Kruskal-Wallis test first, since the multiple comparisons test is appropriate only if the Kruskal-Wallis test finds some statistically significant difference among groups.

[10] U_T is a statistic whose standard deviation is known, and depends only on the total number of items (here, 240 TRAIN pairs) and the number of items in each sample (here, 60 pairs in each class) (Hollander & Wolfe, 1999). The standard deviation is thus the same for all measures, and hence it is appropriate to use multiples of this value as thresholds.

[11] Recall that since rankings are in decreasing order, X $<$ Y and X $\ll$ Y should be interpreted as members of X being generally ranked before members of Y.

[12] Although $\mathrm{Sim}_{\mathrm{dist}}(v+n, rv)$ seems to separate LIT from ABS, as discussed above, we believe this is mainly an accident of the data and hence not reliable.

[13] F-score is a measure of accuracy that balances precision (P) and recall (R) of assigning a class label X. Precision tells, out of all items that are assigned the label X, what proportion is truly of type X; recall tells, out of all things that are truly of type X, what proportion is given the label X. The balanced F-score is calculated as: $F = \frac{2 \cdot P \cdot R}{P+R}$.

References

BANNARD Colin, Timothy BALDWIN & Alex LASCARIDES 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan. 65–72.

BRINTON Laurel J. & Minoji AKIMOTO (eds.) 1999. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. Amsterdam & Philadelphia: John Benjamins.

Burnard 2000. *Reference Guide for the British National Corpus (World Edition)*, 2nd edition.

CACCIARI Cristina & Patrizia TABOSSI (eds.) 1993. *Idioms: Processing, Structure, and Interpretation*. Hillsdale, NJ: Lawrence Erlbaum Associates.

CHURCH Kenneth, William GALE, Patrick HANKS, & Donald HINDLE 1991. Using statistics in lexical analysis. In ZERNIK, Uri (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Hillsdale, NJ: Lawrence Erlbaum Associates. 115–164.

COLLINS Michael 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

COWIE Anthony P., Ronald MACKIN, & Isabel R. MCCAIG 1983. *Oxford Dictionary of Current Idiomatic English*, volume 2: Phrase, Clause and Sentence Idioms. Oxford: Oxford University Press.

DUNN Olive J. 1964. Multiple comparsions using rank sums. *Technometrics*, 6(3):241–252.

FAZLY Afsaneh 2007. *Automatic Acquisition of Lexical Knowledge about Multiword Predicates*. Ph.D. thesis, University of Toronto.

FAZLY Afsaneh & Suzanne STEVENSON 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)*, Trento, Italy. 337–344.

FAZLY Afsaneh & Suzanne STEVENSON 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the ACL'07 Workshop on A Broader Perspective on Multiword Expressions*, Prague, Czech Republic. 9–16.

FELLBAUM Christiane (ed.) 1998. *WordNet, An Electronic Lexical Database*. Cambridge, MA: MIT Press.

FELLBAUM Christiane 2006. Corpus-based studies of German idioms and light verbs. *International Journal of Lexicography*, 19(4):349–360.

FELLBAUM Christiane 2007. The ontological loneliness of idioms. In SCHALLEY, Andrea & Dietmar ZAEFFERER (eds.), *Ontolinguistics: How Ontological Status Shapes the Linguistic Coding of Concepts*. Berlin: Mouton de Gruyter. 419–434.

FIRTH John R. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis (special volume of the Philological Society)*, Oxford: Basil Blackwell. 1–32.

GIBBS Raymond W., Jr., Nandini P. NAYAK, J. BOLTON, & M. KEPPEL 1989. Speaker's assumptions about the lexical flexibility of idioms. *Memory and Cognition*, 17:58–68.

HARRIS Zellig S. 1954. Distributional structure. *Word*, 10(23):146–162.

HOLLANDER Myles & Douglas A. WOLFE 1999. *Non-parametric Statistical Methods*. New York: John Wiley and Sons, 2nd edition.

JACKENDOFF Ray 1997. *The Architecture of the Language Faculty*. Cambridge, MA: The MIT Press.

LAKOFF George & Mark JOHNSON 1980. *Metaphors We Live by*. Chicago: The University of Chicago Press.

LIN Dekang 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98)*, Montreal, Canada. 768–774.

LIN Dekang 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, Maryland. 317–324.

MCCARTHY Diana, Bill KELLER, & John CARROLL 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan. 73–80.

MCDONALD Scott & Michael RAMSCAR 2001. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, 611–616.

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

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

NEWMAN John & Sally RICE 2004. Patterns of usage for English SIT, STAND, and LIE: A cognitively inspired exploration in corpus linguistics. *Cognitive Linguistics*, 15(3):351–396.

NICKEL Gerhard 1968. Complex verbal structures in English. *International Review of Applied Linguistics*, 6:1–21.

NUNBERG Geoffrey, Ivan A. SAG, & Thomas WASOW 1994. Idioms. *Language*, 70(3):491–538.

PAUWELS Paul 2000. *Put, Set, Lay and Place: A Cognitive Linguistic Approach to Verbal Meaning*. Munich: Lincom Europa.

R 2004. *Notes on R: A Programming Environment for Data Analysis and Graphics*.

ROHDE Douglas L. T. 2004. TGrep2 User Manual.

SAG Ivan A., Timothy BALDWIN, Francis BOND, Ann COPESTAKE, & Dan FLICKINGER 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'02)*, Mexico City. 1–15.