Automatic Acquisition of Lexical Knowledge about Multiword Predicates

by

Afsaneh Fazly

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Abstract

Automatic Acquisition of Lexical Knowledge about Multiword Predicates

Afsaneh Fazly

Doctor of Philosophy

Graduate Department of Computer Science

University of Toronto

2007

A multiword predicate is the combination of a predicate (often a verb) with one or more of its arguments, that together form a single unit of predicative meaning. We focus on a broad class of multiword predicates, in which a verb combines with a noun in the direct object position (e.g., *give a groan* and *shoot the breeze*). The semantic interpretation of such multiword predicates involves a certain degree of idiosyncrasy; moreover, they are crosslinguistically frequent and appear in all text genres. Hence, they pose a great challenge to the current models of natural language processing. Most existing computational models treat multiword predicates as syntactically-dependent word sequences or collocations. Such a treatment ignores other important characteristics of these constructions, reflected in their distinct lexical and syntactic behaviour. Nonetheless, cues from the lexicosyntactic properties of multiword predicates have often been used in linguistic and psycholinguistic studies to explain their peculiar semantic behaviour. On the one hand, simple statistical approaches that only draw on the frequency of multiword predicates fail to account for much of the syntactic and semantic behaviour of these constructions. On the other hand, linguistic theories provide generalizations about the behaviour of multiword predicates that can be augmented with probabilistic knowledge about language in use. The main goal of the present study is to propose ways of combining the predictive power of linguistic theories with the coverage and robustness of statistical techniques to acquire linguistically-plausible and reliable corpus-drawn knowledge about multiword predicates.

Dedication

For Reza,

who started me on this path many years before.

Acknowledgements

I am very lucky to have been given the opportunity to work with a great group of people here at the University of Toronto. I am especially grateful to my supervisor, Suzanne Stevenson, for her tireless support and guidance, for all the encouragement she has given me, and for her many inspiring and insightful ideas and observations. All these years, she has been an amazing supervisor and a great friend.

I would like to express my gratitude to the members of my supervisory committee, Graeme Hirst, Anne-Marie Brousseau, and Gerald Penn, for their enlightening questions and comments. It has been an honour to have Diana McCarthy as my external examiner, and I would like to thank her for taking the trouble to come a long way to attend my PhD examination.

During the years I have been a graduate student, I had the pleasure of meeting many bright people and forming lasting friendships. I would like to thank my wonderful officemates, Afra Alishahi and Vivian Tsang, for their invaluable friendship, for listening to my complaints, for the nice chats during our daily tea breaks, and for the chocolate and cookies! I am indebted to my friends in the computational linguistics group, Chris Collins, Paul Cook, Mike Demko, Tim Fowler, Eric Joanis, Saif Mohammad, Ryan North, and Robert Swier, for the enjoyable discussions we shared. Without their help, the evaluation of my ideas would not have been possible. I wish to thank Faye Baron, for her kindness, and Yun Niu, Cosmin Munteanu, Diana Inkpen, and Melanie Baljko for being always ready to answer my questions. My special thanks goes to Jane Morris, for being such a nice and inspiring friend and for the delightful nights out!

I am grateful for the financial support that I have received from the Department of Computer Science, University of Toronto, and the Government of Ontario.

I would like to extend my warmest gratitude to my Iranian friends in Toronto, for all the happy moments we shared, and for their support through the hard times. They will forever remain in my heart and in my thoughts. I want to say thank you to my sweet sisters, Farzaneh and Azadeh, and my dear brother, Amir, for the beautiful memories we share. I could not have done it without their support. I wish to thank my beloved husband, Reza, for his endless patience, his untiring support, and his invaluable companionship over the years. He encouraged me to start this and has always stood by my side. I shall not conclude without thanking my loving parents, Ahmad Fazli and Eshrat-Banoo Eskandari, who have given me my dreams, have always wanted the best for me, and have put my happiness before theirs.

Contents

1 Multiword Expressions **1**

1.1 Challenges for Natural Language Processing 2

1.2 Collocations versus Multiword Lexical Units 5

1.3 Focus of the Study 7

2 Light Verb Constructions **13**

2.1 Linguistic Properties 14

2.2 Separating Literal and Figurative Expressions 16

2.2.1 Syntactic Flexibility 18

2.2.2 A Statistical Measure of Figurativeness 20

2.3 Evaluation 23

2.3.1 Corpus and Experimental Expressions 24

2.3.2 Baselines 25

2.3.3 Human Judgments of Figurativeness 26

2.3.4 Results 28

2.4 Related Work 30

2.5 Summary of Contributions 31

3 Idiomatic Combinations **33**

3.1 Idiomaticity, Semantic Analyzability, and Flexibility 35

3.1.1 Semantic Analyzability 35

3.1.2 Lexical and Syntactic Flexibility 36
3.2 Automatic Recognition of VNICs 39
3.2.1 Measuring Lexical Fixedness 39
3.2.2 Measuring Syntactic Fixedness 41
3.2.3 A Hybrid Measure of Fixedness 45
3.3 Experimental Setup 46
3.3.1 Corpus and Data Extraction 47
3.3.2 Experimental Expressions 47
3.3.3 Parameters 48
3.4 Results 49
3.4.1 Classification Performance 50
3.4.2 Retrieval Performance 52
3.4.3 Summary of Results 54
3.5 Determining the Canonical Forms of Idioms 55
3.6 Related Work 57
3.7 Summary of Contributions 60

4 Idioms, LVCs, and Compositional Combinations 62

4.1 The Four Classes of Verb+Noun Combinations 63
4.2 Properties of Figurative Language 66
4.3 Automatic Classification of Verb+Noun Combinations 67
4.3.1 Measuring Degree of Institutionalization 68
4.3.2 Measuring Degree of Fixedness 68
4.3.3 Measuring Degree of (Non-)Compositionality 70
4.3.4 Other Relevant Properties 72
4.4 Experimental Setup 72
4.4.1 Corpus and Experimental Expressions 72
4.4.2 Classification Strategy and Features 74

4.4.3 Human Judgments 74
4.5 Results 76
4.5.1 Overall Classification Performance 77
4.5.2 Performance on Individual Classes 78
4.6 Related Work 80
4.7 Summary of Contributions 82

5 Summary and Outlook **84**
5.1 Summary of Contributions 85
5.2 Future Directions 88
5.2.1 Short-term Extensions 88
5.2.2 Long-term Goals 90

Appendices **94**

A List of abbreviations **94**

B On the annotation task from Chapter 2 **95**

C Experimental expressions from Chapter 2 **97**

D Experimental pairs from Chapter 3 **101**

E Rankings over test verb–noun pairs **105**

F Canonical forms from Chapter 3 **108**

G Annotation guide **111**

H Experimental pairs from Chapter 4 **122**

I Precision and recall values for the classification task **124**

J Per-class inter-annotator agreements **125**

Bibliography **126**

List of Tables

1.1 The different structure of two seemingly similar English sentences 3

2.1 The different structure of sentences with literal and figurative usages of *give* . . 17
2.2 Pattern sets used in measuring syntactic rigidity 21
2.3 Questions for expressions containing *give* 26
2.4 Distribution of experimental expressions according to human ratings of figurativeness 27
2.5 Correlations between human figurativeness ratings and the statistical measures . 29

3.1 Patterns used in the syntactic fixedness measure 43
3.2 Performance of fixedness measures on test data 50
3.3 Performance of the hybrid measure on test data 51
3.4 Interpolated 3-point average precision values 54

4.1 Features used in classification of verb+noun combinations 75
4.2 Overall observed agreement and kappa score among annotators 76
4.3 Classification performance of the individual feature groups 78
4.4 Classification performance of all features but one feature group 78
4.5 Classification performance of individual feature groups on each class 80

B.1 Questions for expressions containing *take* 95
B.2 Interpretation of answers to the questions for expressions with *take* 96
B.3 Interpretation of answers to the questions for expressions with *give* 96

C.1 Development expressions and their human ratings 97
C.1 Development expressions and their human ratings 98
C.1 Development expressions and their human ratings 99
C.2 Test expressions and their human ratings 99
C.2 Test expressions and their human ratings 100

D.1 Test pairs and their frequencies in the BNC 101
D.1 Test pairs and their frequencies in the BNC 102
D.1 Test pairs and their frequencies in the BNC 103
D.1 Test pairs and their frequencies in the BNC 104

E.1 Test pairs ranked by PMI 106
E.2 Test pairs ranked by $\text{Fixedness}_{\text{overall}}$ 107

F.1 Individual precisions and recalls of automatically identifying canonical forms . 108
F.1 Individual precisions and recalls of automatically identifying canonical forms . 109
F.1 Individual precisions and recalls of automatically identifying canonical forms . 110

H.1 Test pairs and their class labels 122
H.1 Test pairs and their class labels 123

I.1 Individual recall and precision values on test pairs 124

J.1 Per-class observed agreement and kappa score among annotators 125

List of Figures

1.1 Classes of verb+noun combinations on the figurativeness continuum 9

2.1 A pictorial representation of the literal–figurative continuum for *give* 18

2.2 The minimum distance between two PMI values 23

3.1 Performance of $\text{Fixedness}_{\text{lex}}$ and $\text{Fixedness}_{\text{overall}}$ on development data 49

3.2 Performance of $\text{Fixedness}_{\text{overall}}$ on test data as a function of α 52

3.3 Precision–recall curves for PMI and for the fixedness measures 53

4.1 Classes of verb+noun combinations on the figurativeness continuum 63

G.1 A pictorial representation of classes and their properties 113

Chapter 1

Multiword Expressions

The term *multiword expression* (MWE) refers to a combination of two or more, not necessarily contiguous, words that together have a special meaning. Examples are *frying pan*, *in order to*, *give a call*, and *shoot the breeze* ("to chat"). The meaning of a multiword expression often involves idiosyncrasy to some extent, i.e., its meaning may not be fully described by the conventional compositional rules of semantics. Moreover, different types of MWEs may have distinct syntactic and semantic properties. Multiword expressions are cross-linguistically prominent and appear in all text genres; they are used frequently to communicate ideas that are difficult to express using a single word. In fact, the number of MWEs is sometimes argued to equal or even exceed the number of single words in a speaker's lexicon (Jackendoff, 1997; Pauwels, 2000; Sag et al., 2002).

Most multiword expressions stand somewhere between lexical items and syntactic structures. Semantically, they resemble lexical items; syntactically, however, they exhibit some of the behaviour attributed to units with internal structure. Because of this, MWEs are more or less productive, and hence cannot be exhaustively listed. MWEs thus pose a serious challenge for the development of large-scale, linguistically plausible natural language processing (NLP) systems. Of great significance is the well-known problem of determining whether and how different types of MWEs should be represented in a computational lexicon.

In this thesis, we address the problem of lexical acquisition for an important, though mostly-overlooked, class of MWEs. More specifically, we look into the distinct syntactic behaviour of members of this class, as well as how such behaviour relates to their underlying semantic properties. We develop statistical models that capture this relationship. Such models can be used both for acquiring lexical and syntactic information about these MWEs, and for predicting their semantic properties. We show that by combining the predictive power of linguistic theories with the coverage and robustness of statistical techniques, we can acquire linguistically plausible and reliable corpus-drawn knowledge about these MWEs.

In Section 1.1, we give an overview of different classes of MWEs and their properties, as well as the problems they pose to current models of language processing. Section 1.2 points out some of the salient but largely unaddressed issues related to the handling of MWEs. We elaborate the main focus of our study in Section 1.3, also presenting the organization of the thesis and summarizing major contributions of this work.

1.1 Challenges for Natural Language Processing

MWEs include a wide range of linguistic phenomena, such as nominal compounds (e.g., *frying pan*, *car park*), complex prepositions (e.g., *in order to*, *in conformity with*), phrasal verbs (e.g., *give a call*, *eat up*), and idioms (e.g., *shoot the breeze*, *trip the light fantastic*). It is hard to provide a unique set of characteristics that covers all types of MWEs. They vary from completely frozen expressions (e.g., *by and large*) to more flexible ones (e.g., *take a walk*, *take a long walk*). Some are not productive (e.g., *spill the beans*/?*peas*) while others are partially productive (e.g., *take a walk/hop/stroll*). Their meaning can be largely compositional (e.g., *frying pan*), somewhat compositional (e.g., *give a call*), or completely idiosyncratic (e.g., *shoot the breeze*).

Given the rich variety and the cross-linguistic prominence of MWEs, it is evident that NLP applications need to identify and treat them appropriately. This is especially important for

Table 1.1: The different structure of two seemingly similar English sentences.

Sentence in English	Intermediate representation	Translation in French
Jill and Tim shot the bird.	(e1/*shoot*	*Jill et Tim ont abattu l'oiseau.*
	:agent (a1/"*Jill* ∧ *Tim*")	*Jill and Tim shot down the bird.*
	:theme (p1/"*oiseau*")	
Jill and Tim shot the breeze.	(e2/*shoot-the-breeze*	*Jill et Tim ont bavardé.*
	:agent (a1/"*Jill* ∧ *Tim*"))	*Jill and Tim chatted.*

applications that require some degree of semantic interpretation, such as automatic thesaurus extraction, machine translation, and text summarization.

One problem posed by MWEs is their identification. For example, the two sentences shown in Table 1.1 look similar on the surface; however, their syntactic and semantic structures are completely different. The differences are reflected in their representation as well as in their translation into another language, here French. An idiom like *shoot the breeze* should be translated as a single unit of meaning, while a verb phrase such as *shoot the bird* typically has a more direct translation, often a word-for-word one.

The same is true for sentences in 1(a–b) and 2(a–b)—i.e., the underlying syntactic and semantic structure of these sentences cannot be predicted merely by looking at their surface structure. Sentences in 1(a) and 2(a) involve a simple verb, whereas those in 1(b) and 2(b) contain a complex verb (CV).

1. (a) Sam [looked]$_{V}$ [up the street]$_{PP_{arg}}$.
   (b) Sam [**looked up**]$_{CV_{v+prt}}$ the answer.
   (c) Sam [**looked**]$_{CV_{v}}$ the answer [**up**]$_{CV_{prt}}$.
2. (a) Azin [made]$_{V}$ [a cake]$_{NP_{arg}}$ for the party.
   (b) Azin [**made an offer**]$_{CV_{v+np}}$ to the group.
   (c) Azin [**made a** great **offer**]$_{CV_{v+np}}$ to the group.
   (d) [**An offer** was **made**]$_{CV_{np+v}}$ to the group.

Idioms such as *shoot the breeze* and complex verbs such as *look up* and *make an offer* introduce a greater challenge due to their morphosyntactic flexibility. For example, the verb in *shoot the breeze* can appear in all inflectional forms. Also, as sentences in 1(b–c) and 2(b–d) show, the constituent parts of a complex verb can be separated by intervening words, e.g., an adjective or even an argument of the verb.

Because of their particular syntactic behaviour (reflected in the examples above), flexible MWEs require specific treatment in a computational lexicon. Flexible MWEs and fixed MWEs cannot be treated uniformly within a lexicon (as also noted by Sag et al., 2002). Completely fixed expressions, such as *by and large* and *ad hoc*, can be simply stored as "words with spaces". It is clear that such an approach would not suffice for more flexible MWEs: either it does not account for their morphosyntactic variability, or it results in lexical proliferation. For example, capturing the morphological variability of an idiom such as *shoot the breeze* would come with the expense of having to list all the inflectional forms.[1] Also, it is inappropriate (and perhaps infeasible in many cases) to list all the possible syntactic variations of an MWE such as *make an offer*. On the other hand, flexible MWEs cannot be treated as fully compositional syntactic structures either. Such an approach may lead to overgeneration, e.g., generating an improper syntactic form such as ?*the breeze was shot*.[2] This is because many MWEs have restricted syntactic variability. The compositional approach is also not capable of predicting whether a given expression is an MWE (involving some degree of semantic idiosyncrasy), or a compositional phrase, as shown in the sentences in Table 1.1 and Examples 1 and 2 above.

Clearly, it is essential to develop accurate methods both for the appropriate handling of flexible MWEs, and for the acquisition of syntactic and semantic knowledge about them. The development of such methods requires a careful examination of the distinct syntactic and se-

[1]The lexical proliferation problem is more serious when dealing with productive MWEs. These often come in families (e.g., *take a walk/hop/stroll*, and *wipe/clean/wash up*), hence listing them separately results in a rapid growth of the lexicon as well as the loss of generality. Dealing with the productivity issue is outside the scope of this thesis; the interested reader is referred to (Fazly et al., 2006).

[2]In the case of *semi*-productive MWEs such an approach may result in the generation of unacceptable lexical variations such as ?*take a limp*—in parallel to *take a walk/hop/stroll* (see previous footnote).

mantic properties of these expressions. The following section discusses some of the salient and differentiating characteristics of different types of MWEs, focusing mainly on issues not addressed by previous studies in the field.

1.2 Collocations versus Multiword Lexical Units

Multiword expressions can be broadly classified into institutionalized phrases and lexicalized phrases (Sag et al., 2002). Institutionalized phrases or **collocations** are known to be syntactically and semantically regular to a large extent, but statistically idiosyncratic. In other words, collocations are conventional associations of words whose cooccurrence happens more often than by chance. A classic example of a collocation is *strong tea*, which is fully compositional, but occurs with greater frequency than any other lexical instantiation of the same concept, such as ?*powerful tea*. In other aspects, the collocation *strong tea* behaves like any other adjective–noun combination.

By contrast, lexicalized MWEs or **multiword lexical units** (MWUs) involve some degree of lexical, syntactic, and/or semantic idiosyncrasy, but may or may not be observed with higher than expected frequency in a given context. In other words, a multiword lexical unit is a combination of two or more words, not necessarily contiguous, that together form a single unit of meaning. MWUs are semantically idiosyncratic to some extent, i.e., the unitary meaning of the expression cannot be determined merely by combining the meanings of the parts. They are also syntactically peculiar, i.e., they often behave differently from similar-on-the-surface combinations that are syntactic structures rather than lexical units. For example, *give a present* is a literal verb phrase and hence can appear in a variety of syntactic constructions, such as the passive, whereas *give a try* is an MWU and syntactically more restricted.

Although statistical idiosyncrasy is not a necessary condition for being an MWU, some MWUs may occur more frequently than expected by chance. Thus, some of the existing techniques for identifying collocations may also work for certain MWUs. It is nonetheless

important to note that most MWUs cannot be identified by merely relying on the statistical significance of their cooccurrence frequency. For example, the idiom *take heart* ("to start to feel more hopeful") appears in the British National Corpus (BNC, 2000) with a frequency lower than expected by chance, given the individual frequencies of its constituents.

Evidently, it is important to distinguish the two phenomena, i.e., institutionalized versus lexicalized combinations. This is especially needed since the two types of MWEs have different properties and hence require different acquisition strategies, as well as different treatments within a computational system. Although some researchers have recognized the importance of this distinction (see, e.g., Sag et al., 2002; Bannard, 2005), a great deal of confusion still exists, both in the terminology used for the two types of expressions,[3] and in the empirical approaches addressing them.

Many existing systems for identifying multiword expressions use raw frequency or a standard measure of association strength, such as pointwise mutual information or log-likelihood ratio, to extract statistically significant cooccurrences of words (Church et al., 1991; Seretan et al., 2003). Some existing systems, even when focusing on a particular class of MWUs, treat them mainly as collocations (Dras and Johnson, 1996; Krenn and Evert, 2001). Others attempt to improve the existing measures of collocation extraction, aiming to address some of their drawbacks, such as their sensitivity to frequency, or their limitation in terms of the length of the extracted word sequences (Deane, 2005; Wermter and Hahn, 2005).

Smadja (1993) goes one step further by also looking into the rigidity of a word sequence with respect to the relative position of words therein, in order to identify multiword expressions. His simple frequency-based statistics cover a broad range of expressions, from simple collocations (e.g., *sales fell*), to phrasal verbs (e.g., *make a decision*), to more idiomatic combinations (e.g., *make sense*). Smadja's approach is interesting because it is capable of identifying collocations as well as MWUs. Nonetheless, its main drawback is that it does not distinguish

[3]Many different terms have been used to refer to MWEs, such as "collocations", "multiword expressions", "multiword phrases", "multiword terms", "non-compositional compounds", and "non-compositional phrases".

between the two, hence ignoring the fact that these should be treated differently in a computational system.

As mentioned previously, compared to collocations, MWUs are harder to identify since one cannot identify them simply by examining their frequency of occurrence. Many previous studies thus look into the linguistic properties of particular classes of MWUs in order to learn about their syntax and semantics (see, e.g., Grefenstette and Teufel, 1995; Melamed, 1997a; Lin, 1999; McCarthy et al., 2003; Bannard et al., 2003; Baldwin et al., 2003). These and other similar studies are what we build our work on. We attempt to address some of the issues overlooked by these studies, and to extend them in various directions. Here, we pay special attention to the properties that distinguish MWUs from collocations and other compositional phrases. We thus concentrate on a largely overlooked class of MWUs, in order to provide a deep analysis of the distinct syntactic and semantic properties of its members. By drawing on such properties, we are able to develop accurate statistical techniques for the identification of these MWUs, as well as for the acquisition of syntactic and semantic knowledge about them. We also compare our proposed statistical measures with a widely used measure of collocation extraction. The next section further elaborates on the focus of our study.

1.3 Focus of the Study

In the study presented in this thesis, we focus on a class of flexible MWUs, i.e., those that combine a verb and a noun in the direct object position to form a new (complex) predicate, such as *make an offer*, *make a killing*, and *shoot the breeze*. Throughout the thesis, we use the general term *multiword predicate* to refer to any such complex predicate. Multiword predicates are semantically idiosyncratic to a large extent, i.e., the meaning of a multiword predicate cannot be derived from a simple composition of the meanings of its parts. Rather, most multiword predicates involve a figurative (metaphorical and/or idiomatic) use of one or both of their constituents, as in the examples above.

Figurative language is a powerful mechanism, enabling creative expression of unfamiliar, abstract notions in terms of concrete, easily visualizable things and situations (Lakoff and Johnson, 1980; Johnson, 1987; Nunberg et al., 1994). Indeed, figurative language is such a central part of linguistic competence that many terms, especially multiword expressions, that are currently accepted as "regular" language have their origin in figurative uses (Newman, 1996; Brinton and Akimoto, 1999). Some of these expressions are viewed as meaning extensions of their component words, which at least partly contribute their semantics or a figurative version of their semantics. Others have become idioms with idiosyncratic semantics whose relation to their component words is not obvious (except possibly historically).

In particular, it is common across languages for multiword predicates to form around certain high frequency verbs that easily undergo metaphorization (Pauwels, 2000; Newman and Rice, 2004). In their literal uses, these so-called **basic** verbs typically refer to states or acts that are central to human experience (e.g., *cut*, *make*, *give*, *keep*).[4] Verb+noun combinations containing a basic verb involve a range of meaning extensions of the verb, from abstract to more metaphorical and idiomatic uses, as shown in 3(a–d):

3. (a) cut taxes, cut corners, cut the mustard
   (b) make a case, make a bow, make a killing
   (c) give confidence, give a groan, give a whirl
   (d) keep the peace, keep a check, keep one's cool

Many linguists have acknowledged that figurativeness is a matter of degree. Nonetheless, one can distinguish literal expressions, such as *give a present*, more abstract combinations such as *give confidence*, metaphorical combinations such as *give a groan*, and idioms such as *give a whirl*, as coherent classes falling on the figurativeness continuum, as depicted in Figure 1.1 below.[5]

[4]Cacciari (1993) refers to such words as "idiom-prone lexemes"; Claridge (2000) calls them basic verbs that are very common and multifunctional.

[5]We are aware that representing figurativeness as a linear continuum is a simplification of the phenomenon. However, we believe that such a simplification is necessary to tackle an inherently difficult problem. In Chapter 4, we look at different dimensions of figurativenss to partially overcome this limitation.

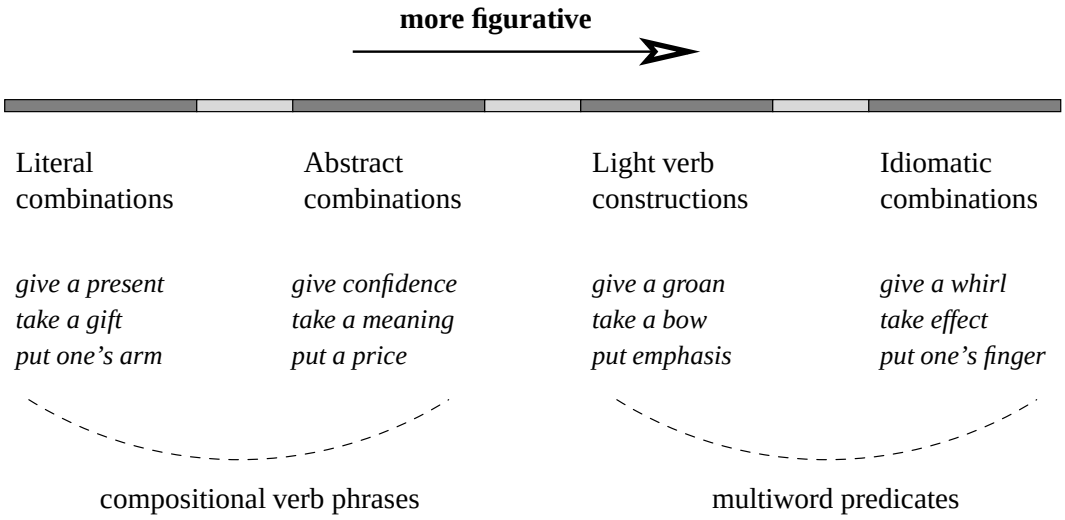


Figure 1.1: Classes of verb+noun combinations on the figurativeness continuum.

Note that all these different types of verb+noun combinations conform to the grammar rules of verb phrase formation; thus, on the surface, they are indistinguishable. A computational system, however, must be capable of telling them apart, hence treating them differently. *Give a present* is a compositional combination of a verb and a noun, both contributing their literal semantics. *Give confidence* is also largely compositional, but it involves a different sense of the verb *give* (an abstract transfer rather than a physical one). *Give a groan* also contains a figurative sense of the verb, but differs from a compositional verb phrase because the noun *groan* contributes the primary predicative meaning of the expression—that is, *give a groan* can be roughly paraphrased by the verb *groan*. Such semi-compositional combinations belong to a general class of multiword predicates known as **light verb constructions** or LVCs.[6] In contrast, *give a whirl* is an **idiom** whose meaning has nothing to do with *give* or *whirl* (*give* something *a whirl* means "to try something to see whether one likes it"). In addition, a higher level of distinction is required between LVCs and idioms as multiword predicates, and literal

[6]LVCs are also known as "support verb constructions" (Nicolas, 1995), "complex verbs" (Brinton and Akimoto, 1999, Chapter 1), "complex/composite predicates" (Cattell, 1984), "verbal phrases" (Hiltunen, 1999), "verbo-nominal structures" (Akimoto, 1999), "verbo-nominal combinations" and "verb-adjective combinations" (Claridge, 2000), and "complex verbal structures" (Nickel, 1968).

and abstract combinations as largely compositional verb phrases.

The distinction among these different types of verb+noun combinations is not always as clear-cut as one would hope. For example, expressions such as *take care* and *make haste* are regarded as LVCs by some researchers (e.g., Nicolas, 1995), but as semi-idioms or idioms by others (e.g., Cowie et al., 1983; Nunberg et al., 1994). The disagreement on the borderline between idioms and LVCs is partly due to the fact that LVCs exhibit much of the behaviour often attributed to idioms, e.g., semantic opacity and morphosyntactic fixedness. However, as the examples in the previous paragraph show, the distinction between the these classes is essential for an NLP system. For example, recall that the nature and the degree of figurativeness is different in idioms and LVCs: in LVCs, the verb is known to have a metaphorical meaning, and hence can form new LVCs (semi-)productively, whereas idioms are mainly considered as "fossilized" or "dead" metaphors (Hobbs 1979; Seaton and Macaulay 2002; see Gibbs (1993) for an alternative view). Idioms and LVCs also differ with respect to their lexical and morphosyntactic flexibility, and this needs to be addressed in their lexical representation.

Despite the dispute on the boundary of idioms and LVCs in some cases, there is strong agreement on their key role in language, and hence on the importance of studying them. Idioms are known to be extremely difficult to characterize; at the same time, their use in language has been shown to be widespread (Cowie et al., 1983; Marantz, 1984; d'Arcais, 1993; Seaton and Macaulay, 2002). LVCs are widely used in many diverse languages, including, but not limited to, English (Kearns, 2002), French (Desbiens and Simon, 2003), Spanish (Alba-Salas, 2002), Persian (Karimi, 1997), Urdu, Hindi (Butt, 2003), Chinese (Lin, 2001), and Japanese (Miyamoto, 2000). In some languages, such as Persian, LVCs are known to outnumber single-word verbs (Khanlari, 1973).

It is evident that LVCs and idioms are a clear challenge to the current computational models of language (Fellbaum, 2005). Nonetheless, unlike other types of MWUs, such as nominal compounds and verb-particle constructions (e.g., *give up* and *figure out*), these multiword predicates have been granted relatively little attention within the computational linguistics com-

munity (though see Grefenstette and Teufel, 1995; Dras and Johnson, 1996; Fellbaum, 2002; Villada Moirón, 2004; Venkatapathy and Joshi, 2005a). We thus choose to focus on these multiword predicates, i.e., LVCs and idioms, as important classes of multiword lexical units that require special treatment in a computational lexicon. We first analyze the linguistic properties of these constructions in order to better understand the differentiating characteristics of each class. We then devise statistical measures that model each of these properties. The predictions of the models are tested to evaluate the usefulness of the statistical measures as well as that of the underlying linguistic properties in distinguishing members of these two classes from each other, and from similar-on-the-surface abstract and literal combinations.

In order to tackle this classification problem, we first look at the two classes of multiword predicates (LVCs and idioms) separately. In Chapter 2, we focus on the acquisition of semantic knowledge about LVCs. More specifically, we examine the relation between semantic idiosyncrasy and syntactic behaviour of LVCs, as discussed in the linguistics literature. We propose statistical, corpus-based measures that use this relationship to place a potential LVC (i.e., a given combination of a light verb and a noun) on a continuum from literal to more figurative. The idea is that the higher the degree of figurativeness, the more likely it is that the given combination is an LVC. Earlier versions of this work are published in Stevenson et al. (2004) and Fazly et al. (2005). The work has also been accepted for publication in the *Journal of Lexical Resources and Evaluation* (Fazly et al., 2006).

In Chapter 3, we concentrate on idioms and try to distinguish them from similar-on-the-surface literal combinations. More specifically, we address two important issues regarding the acquisition and representation of knowledge about idioms: their degree of flexibility and their degree of idiomaticity. We first tackle the issue of flexibility by proposing techniques for automatically determining the degree of lexical and syntactic flexibility of a given verb+noun combination. We then deal with idiomaticity, drawing on its relation to flexibility. We show that our approach has promising results in separating idiomatic from literal verb+noun combinations. A slightly different version of this work is published in Fazly and Stevenson (2006).

Finally, in Chapter 4, we combine these pieces of work in order to identify LVCs and idioms, and to separate them from literal and abstract verb+noun combinations. The measures proposed in Chapters 2 and 3 look into the lexical and syntactic behaviour of LVCs and idioms, respectively, to draw inferences about their semantic properties. In other words, the main focus of these measures is on the lexicosyntactic fixedness of LVCs and idioms. In Chapter 4, we look at other aspects of figurative language, such as institutionalization and semantic non-compositionality. We employ existing measures to model these properties, and use them in combination with the fixedness measures in order to classify a set of given verb+noun combinations into idioms, LVCs, abstract combinations and literal phrases.

Chapter 5 summarizes our contributions and presents possible directions for future extension.

Appendix A presents a list of abbreviations used in this thesis. Appendix B contains information about the human judgments used as the gold standard in evaluating the work presented in Chapter 2. Appendix C and Appendix D list the experimental expressions used for evaluation in Chapter 2 and Chapter 3, respectively. Appendix E and Appendix F provide more detail on some of the experiments reported in Chapter 3. Appendix G contains the annotation guide given to our human judges for the annotation task described in Chapter 4. Appendix H presents the list of unseen test items used for evaluation, and Appendix I provides further detail on the results of experiments from Chapter 4. Appendix J contains more information regarding inter-annotator agreements reported in Chapter 4.

Chapter 2

Light Verb Constructions

As explained in Chapter 1, a wide variety of multiword predicates result from basic verbs in combination with noun complements, lying along a continuum of less to more figurative usage, culminating in idiomatic expressions. In this chapter, we focus on light verb constructions, which have a figurative use of the verb but are not fully idiomatic. Examples of English LVCs with varying degrees of figurativeness are given in 4(a–d).

4. (a) give a speech, give a groan
   (b) make an offer, make haste
   (c) take a walk, take care
   (d) put pressure, put emphasis

As previously explained, LVCs are semantically transparent for the most part. Nonetheless, they vary in the degree to which the meaning of the basic verb (referred to as a light verb when used in an LVC) differs from its literal semantics. LVCs also exhibit varying levels of tolerance for lexical and syntactic variations. Consequently, an appropriate syntactic and semantic treatment of these expressions cannot be uniform. We address these issues through a careful linguistic analysis of the properties of LVCs and their relation to statistical behaviour, to support the automatic acquisition of semantic and syntactic knowledge about them.

Section 2.1 presents a brief description of the important general characteristics of LVCs, especially those that motivate their widespread use across different languages. In Section 2.2, we put forward our proposal for determining the degree of figurativeness of a potential LVC, i.e., a given combination of a basic verb and a noun complement. Section 2.3 presents an evaluation of our proposed technique. Results of the evaluation show that our proposed measure highly correlates with human judgments on the same property, providing support for the appropriateness of such an approach. Section 2.4 provides a concise survey on related studies. Section 2.5 summarizes our contributions.

2.1 Linguistic Properties

An LVC is formed around a highly polysemous verb, such as *do*, *give*, *have*, *make*, or *take* in English (Quirk et al., 1985).[1] The verb constituent of an LVC is called a light verb because it is assumed to have lost its literal semantics to some degree (Butt, 2003), contributing a figurative meaning that is a metaphorical extension of its literal semantics. The complement of the light verb in an LVC—that can be a noun, as well as a verb, an adverb, an adjective, or a prepositional phrase—usually contributes compositionally to the overall meaning of the expression.

In many languages, the light verb constituent of an LVC has been observed to contribute to the aspectual properties of the multiword predicate. Telicity, durativity, and perfectivity are among aspectual properties that are determined, or at least affected, by the choice of the light verb in many LVCs (Wierzbicka, 1982; Butt, 1997; Folli et al., 2003). One of the most central aspectual functions of a light verb is considered to be adding an intended end point to the open-ended action that is expressed by the complement (Tanabe, 1999). Thus, the LVC formation process converts an activity to an accomplishment or achievement, without the need for an explicit goal, as in *move* vs. *make a move*, and *bite* vs. *take a bite* (Brinton and Akimoto,

[1]Light verbs with similar meanings to these have been documented in many languages, including those that are genetically unrelated (Butt, 1997), e.g., *faire* (“to do/make”) in French, *kardan* (“to do/make”) in Persian, and *suru* (“to do/make”) in Japanese.

1999, Chapter 1).

Another important role of LVCs is that they provide lexical alternatives to syntactic structures (Claridge, 2000). More specifically, by choosing different light verbs, speakers can generate the same effect as they would by using different syntactic structures. For example, in some languages, different light verbs in combination with the same complement account for an argument structure alternation. The Persian light verbs *kardan* (“make/do”) and *šodan* (“become”) are responsible for the causative/inchoative alternation, when used with the same complement (Vahedi-Langrudi, 1996), as shown in 5(a–b):

5. (a) âsemân sâf **šod**.
sky clear **become**-past.
The sky cleared.

(b) bâd âsemân râ sâf **kard**.
wind sky Obj-marker clear **make**-past.
The wind cleared the sky.

In fact, the choice of the light verb may affect the causativity of the event in general (see Folli et al., 2003; Butt, 2003, among others). For example, certain light verbs in Persian, e.g., *âvardan* (“to bring”), *dâdan* (“to give”), and *kardan* (“to do/make”), often give a causative meaning to the LVCs they appear in. Some English light verbs also have a similar effect, e.g., compare *bring to light* and *come to light*. In some languages, such as English, using a light verb construction in place of the corresponding single-word verb increases the volitionality of the agent. For example, note the contrast between the two sentences in 6(a) and 6(b):

6. (a) Sam *walked* along the street.

(b) Sam *took a walk* along the street.

Yet another salient property of LVCs, and perhaps one of the main motivations for their frequent use, is the flexibility of verb modification that they allow. For example, in English, the adjectival modifiers of LVCs appear to be easier to use than the adverbial modifiers of simple

verbs, e.g., *take a quick/ brief/ long look* versus *look quickly/ briefly/ for a long time* (Nickel, 1968; Brinton and Akimoto, 1999).

Overall, the semantic spreading that occurs in LVCs (as the result of the splitting of the verbal content over multiple words) gives them an immense expressive power. Such high descriptive and communicative power motivates their frequent use in spoken as well as in written language. With no doubt, LVCs can be said to be a core component of many languages today, including English (Hiltunen, 1999).

2.2 Separating Literal and Figurative Expressions

As mentioned previously, we focus on the broadly-documented class of LVCs in which a basic verb combines with a noun in its direct object position. The nominal component of such an LVC is often an indefinite, non-referential predicative noun—i.e., a noun that has an argument structure. The predicative noun (PN) is the primary source of semantic predication (Wierzbicka, 1982); and is (i) morphologically related to a verb, as in *make a decision* (*decide*) and *give a groan* (*groan*), (ii) etymologically related to a verb, as in *have a thought* (*think*) and *give a speech* (*speak*), or (iii) not related to a verb, as in *give an overview* and *make an effort*[2]. The predicative noun constituent of such an LVC (in its canonical form) appears as a bare noun, or with an indefinite article (Brinton and Akimoto, 1999), as shown in 7(a–d):

7. (a) Parissa *took a walk* along the beach.
   (b) Azin *took revenge* after 19 years.
   (c) Dana *gave* her some *help*.
   (d) Sam *made a joke* to his friends.

We refer to this class of light verb constructions as “LV+PN” constructions, although we will also continue to use the broader term LVC to refer to these expressions.

[2]Cattell (1984) states that according to the *Oxford English Dictionary*, there is an obsolete verb *effort*.

Table 2.1: The different structure of sentences with literal and figurative usages of *give*.

Sentence in English	Intermediate representation	Translation in French
Azin gave Sam a present.	(e1/*give* :agent (a1/"*Azin*") :theme (p1/"*present*") :recipient (s1/"*Sam*"))	*Azin a donné un cadeau à Sam.* *Azin gave a present to Sam.*
Azin gave the lasagna a try.	(e2/*give-a-try*≈*try* :agent (a1/"*Azin*") :theme (l1/"*lasagna*"))	*Azin a essayé le lasagne.* *Azin tried the lasagna.*
Azin gave a groan.	(e2/*give-a-groan*≈*groan* :agent (a1/"*Azin*"))	*Azin a gémi.* *Azin groaned.*

As a first step in the creation of a lexicon of LVCs, we propose methods for the acquisition of syntactic and semantic knowledge about them. More specifically, we propose automatic means for distinguishing expressions that have figurative uses of a basic verb in an LVC (i.e., a light verb) from those that have literal uses of the verb in a fully compositional verb phrase. When used literally, basic verbs (like any other verb) compositionally contribute their literal meaning to the phrase they appear in. For example, in *give a present*, *give* refers to the "transfer of possession" of a THEME (*present*) to a RECIPIENT. When used in an LVC, the meaning of a light verb is often a metaphorical (figurative) extension of the basic semantics. For example, *give permission* means that an abstract entity (*permission*) is "transferred" to someone, but no "possession" is involved. In *give a groan*, the notions of "transfer" and "possession" are even further diminished.

While figurative uses of a verb are indistinguishable on the surface from the literal uses (*give a present* vs. *give a groan*), this distinction is essential to an NLP application. As an example, Table 2.1 illustrates the importance of such a distinction for a machine translation system: the figurative expressions should be translated as a single unit of meaning, while the literal usage typically has a more direct translation (often word-for-word). Determining automatic mechanisms for distinguishing literal and figurative uses of a basic verb thus proves

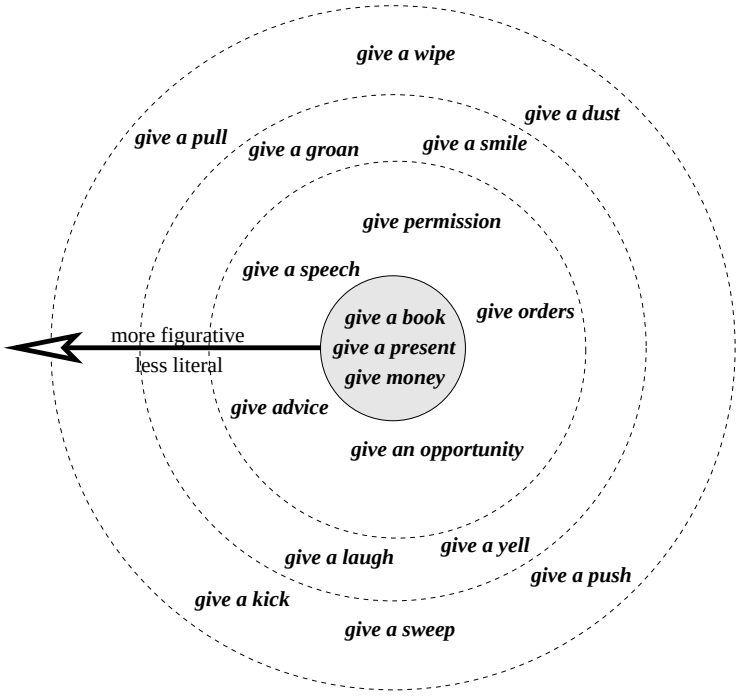


Figure 2.1: A pictorial representation of the literal–figurative continuum for *give*.

to be an indispensable task, especially when some degree of semantic interpretation is required.

Determining the degree of figurativeness of the verb constituent of a potential LVC is the main focus of this section. Section 2.2.1 presents a discussion of the particular syntactic properties of "LV+PN" constructions that relate to the degree of figurativeness of the light verb constituent. We then propose, in Section 2.2.2, a statistical measure which incorporates these properties to place light verb usages on a continuum of meaning from literal to figurative, as depicted in Figure 2.1.

2.2.1 Syntactic Flexibility

There is linguistic evidence that the semantic properties of a lexical item largely determine its syntactic behaviour. In particular, the degree of compositionality of a multiword expression is known to affect its syntactic flexibility, and therefore its appearance in certain syntactic constructions (Gibbs, 1993; Glucksberg, 1993; Nunberg et al., 1994). The same is true for

English "LV+PN" constructions, which enforce restrictions on the syntactic freedom of their noun constituents (Brinton and Akimoto, 1999; Kearns, 2002).

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

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

In others, the noun may be introduced by a definite article, pluralized, passivized, relativized, or even *wh*-questioned, as in 9(b–f). Note, however, that the dative use, as in 9(g), is still questionable.

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

The degree to which an LVC has restricted syntactic freedom, as in these examples, is related to the degree to which the light verb has lost its literal semantics. *Give* in an expression such as *give a groan* (cf. 8) is presumed to be a more abstract usage than *give* in expressions such as *give a speech* (cf. 9); here, we see that it is also more syntactically restricted. By contrast, a literal phrase, such as *give a present*, which is a compositional combination of the (literal)

verb and a noun, exhibits complete syntactic freedom, allowing all the constructions in these examples.

The linguistic explanation for this spectrum of behaviour relies on properties of the noun constituent. If the noun has an independent semantic identity, as in a literal phrase, then it exhibits full syntactic freedom (Gibbs, 1993). As the sentences in 9 above show, LVCs whose noun constituent can be treated, possibly metaphorically, as the direct object of the light verb also show syntactic flexibility to a large extent. However, in more abstract LVCs, the flexibility of the noun is much more restricted, as in 8.

To summarize, the less flexible the noun constituent of an "LV+PN" construction, the less literal (more figurative) the meaning of the light verb. We use this insight to devise a statistical measure of figurativeness, which uses evidence of syntactic rigidity of a potential LVC to situate it on a scale of literal to figurative usage of the light verb. This measure can be used to determine whether a particular combination of a light verb and a noun contains a literal usage of the verb (as in *give a book*) or a figurative usage in an LVC (as in *give a groan*), and the degree of figurativeness of the latter.

2.2.2 A Statistical Measure of Figurativeness

We propose a statistical measure that quantifies the degree of figurativeness of the light verb (lv) in an expression with a noun n, represented as a pair $\prec lv, n \succ$. The measure assigns a score to the target pair $\prec lv, n \succ$ by examining its frequency of occurrence in any of a set of relevant syntactic patterns, such as those in examples 8 and 9 above. The measure is defined as:

$$\textsc{Figness}_{\textsc{lv}}(lv,n) \doteq \textsc{Assoc}(lv,n) + \textsc{Diff}(\textsc{Assoc}_{pos}, \textsc{Assoc}_{neg}) \tag{2.1}$$

whose components are explained in turn in the following paragraphs.

The first component, $\textsc{Assoc}(lv,n)$, measures the strength of the association between the light verb and the complement noun. This is expected to reflect the degree to which these

Table 2.2: Pattern sets used in measuring syntactic rigidity.

$\mathcal{PS}_{pos}$	$= \{$ "lv_{active} det_{nondef} n_{sg}" $\}$
$\mathcal{PS}_{neg}$	$= \{$ "lv_{active} det_{nondef} n_{pl}",
	"lv_{active} det_{def} $n_{sg,pl}$",
	"$\text{det}_{def,nondef}$ $n_{sg,pl}$ $lv_{passive}$ " $\}$

two components are bound together within a single unit of meaning. Church et al. (1991) propose an objective measure for estimating word association norms. The measure is based on the information theoretic notion of mutual information, and is referred to as pointwise mutual information (PMI). We thus choose to calculate $\text{ASSOC}(lv,n)$ using PMI, as in:

$$\begin{aligned} \text{ASSOC}(lv,n) &\doteq \text{PMI}(lv;n) \\ &\doteq \log \frac{P(lv,n)}{P(lv)\,P(n)} \\ &\approx \log \frac{N \times f(lv,n)}{f(lv,*)\,f(*,n)} \end{aligned} \tag{2.2}$$

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

Recall from Section 2.2.1 that the more figurative the light verb of an LVC, the more rigid the LVC. The second component, DIFF, estimates the degree of syntactic rigidity of the expression formed from *lv* and *n*, by examining their association within different syntactic patterns. ASSOC_{pos} measures the strength of association between the expression and $\mathcal{PS}_{pos}$, the pattern set that includes syntactic patterns preferred by more figurative LVCs. Similarly, ASSOC_{neg} measures the strength of association between the expression and $\mathcal{PS}_{neg}$, representing patterns that are less preferred by LVCs. Thus, the greater the difference between ASSOC_{pos} and ASSOC_{neg}, the more syntactically rigid is the expression $\prec lv,n \succ$.

In our current formulation, the two sets $\mathcal{PS}_{pos}$ and $\mathcal{PS}_{neg}$ contain syntactic patterns encod-

ing the following attributes: the voice of the extracted expression (active or passive); the type of the determiner introducing the noun constituent, *n* (definite or non-definite, the latter including the indefinite determiner *a/an* as well as no determiner); and the number of *n* (singular or plural). As shown in Table 2.2, $\mathcal{PS}_{pos}$ consists of a single pattern with values active, non-definite, and singular, for these attributes, corresponding to the use of $\prec lv, n \succ$ in the prototypical LVC pattern (e.g., "*Azin gave a groan before falling asleep.*"). $\mathcal{PS}_{neg}$ has all the patterns with at least one of these attributes having the alternative value. We choose these attributes based on evidence from linguistic studies on English LVCs (e.g., Wierzbicka, 1982; Brinton and Akimoto, 1999). Note, however, that this formulation is flexible and could be expanded to incorporate more attributes if necessary.

To measure the strength of association of an expression with a set of patterns, e.g., $\mathcal{PS}_{neg}$, we use the PMI between the expression and the set, as shown in Eqn. (2.3) below. (ASSOC_{pos} is calculated similarly, by replacing $\mathcal{PS}_{neg}$ with $\mathcal{PS}_{pos}$.)

$$\begin{aligned}
\text{ASSOC}_{neg} &\doteq \text{PMI}(lv, n; \mathcal{PS}_{neg}) \\
&\doteq \log \frac{P(lv, n, \mathcal{PS}_{neg})}{P(lv, n)\, P(\mathcal{PS}_{neg})} \\
&\approx \log \frac{n \times f(lv, n, \mathcal{PS}_{neg})}{f(lv, n, *)\, f(*, *, \mathcal{PS}_{neg})} \\
&= \log \frac{N \sum_{pt_j \in \mathcal{PS}_{neg}} f(lv, n, pt_j)}{f(lv, n, *) \sum_{pt_j \in \mathcal{PS}_{neg}} f(*, *, pt_j)}
\end{aligned} \tag{2.3}$$

Our calculations of the PMI values in the estimation of ASSOC_{pos} and ASSOC_{neg} use maximum likelihood estimates of the true probabilities. This results in PMI values with different levels of confidence (since different syntactic patterns have different frequencies of occurrence in text). Thus, directly comparing the two association strengths, ASSOC_{pos} and ASSOC_{neg}, is subject to a certain degree of error. Following Lin (1999), we estimate the difference more accurately, by comparing the two confidence intervals surrounding the calculated association strength values, at a confidence level of 95%. Like Lin (1999) and Dunning (1993), we assume the estimates of the true probabilities are normally distributed. We form confidence intervals

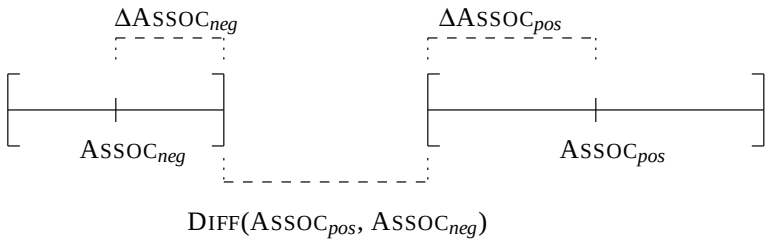


Figure 2.2: Approximating the difference between two PMI values as the minimum distance between the two corresponding confidence intervals.

representing the ranges of the estimates of $P(lv,n,\mathcal{PS}_{pos})$ and $P(lv,n,\mathcal{PS}_{neg})$ in Eqn. (2.3) above. We use these ranges to form confidence intervals for the PMI values. We then take the minimum distance between the two intervals as a conservative estimate of the true difference, as depicted in Figure 2.2, and shown in the following formula:

$$\begin{aligned}&\textsc{Diff}(\textsc{Assoc}_{pos}, \textsc{Assoc}_{neg}) \doteq \\ &\quad (\textsc{Assoc}_{pos} - \Delta\textsc{Assoc}_{pos}) - (\textsc{Assoc}_{neg} + \Delta\textsc{Assoc}_{neg})\end{aligned} \tag{2.4}$$

We expect that estimating the difference in this way lessens the effect of differences that are not statistically significant.

To summarize, the stronger the association between *lv* and *n*, and the greater the rigidity of their use together (as measured by the difference between their association with positive and negative syntactic patterns), the more figurative the meaning of the light verb, and the higher the score given by $\textsc{Figness}_{\textsc{lv}}(lv,n)$.

2.3 Evaluation

To determine how well our proposed measure, $\textsc{Figness}_{\textsc{lv}}$, captures the degree of figurativeness of a light verb usage, we compare the scores it assigns to a list of test expressions with the ratings assigned by human judges. Section 2.3.1 describes the selection of experimental

expressions, and the corpus we use to estimate frequency counts required by the measure. Section 2.3.2 presents two baselines with which we compare the performance of our proposed figurativeness measure. In Section 2.3.3, we elaborate on our approach in collecting consensus human ratings of figurativeness for the experimental expressions. Finally, Section 2.3.4 presents the evaluation results.

2.3.1 Corpus and Experimental Expressions

Common light verbs in English include *do*, *have*, *give*, *make*, and *take*, among others (Quirk et al., 1985; Brinton and Akimoto, 1999). In the evaluation of our figurativeness measure, we focus on two of these, *give* and *take*, that are frequently and productively used in light verb constructions (Claridge, 2000). Moreover, these verbs are highly polysemous: The number of different WordNet senses for *give* and *take* are 44 and 42, respectively (Fellbaum, 1998). This is an important consideration for us since we need expressions that cover a wide range of possible meaning extensions of a particular verb.

We use the British National Corpus (BNC, 2000), both as a source for extracting experimental expressions, and as a corpus for estimating the frequency counts required by the figurativeness measure. We automatically parse the BNC using the Collins parser (Collins, 1999), and further process it using TGrep2 (Rohde, 2004) and NP-head extraction software based on heuristics from Collins (1999).[3]

We first extract, from the BNC, all verb–noun pairs in which the noun is the head of the noun phrase (NP) in the direct object position of the verb. We then randomly select our experimental expressions from among those verb–noun pairs whose verb is either *give* or *take*. These are divided into development and test expressions, referred to as DEV and TEST, respectively. In total, we have 150 development and 70 test expressions, of which 114 involve the verb *give* and 106 involve *take*. Originally, we extracted equal number of expressions for *give* and *take*. However, we had to remove a few expressions of each type that were considered ambiguous or

[3]The NP-head extraction software has been generously provided to us by Eric Joanis.

unacceptable by a majority of our human judges.[4]

Each verb–noun instance extracted from the corpus is augmented with information about the syntactic pattern it appears in: whether the sentence they appear in is active or passive, whether the noun is singular or plural, whether the determiner introducing the noun is definite or non-definite. This information is used in the maximum-likelihood estimation of the probabilities needed to estimate $\text{FIGNESS}_{\text{LV}}(lv, n)$ for any given $\prec lv, n \succ$ (see equations in Section 2.2.2).

2.3.2 Baselines

In the evaluation of our proposed figurativeness measure, we use two baselines: (i) a standard measure of collocation, PMI; (ii) a linguistically-informed baseline to which we refer as PMI_{LVC}. PMI simply measures the strength of collocation of the two constituents of a target expression. $\text{PMI}(lv; n)$ is calculated as in Eqn. (2.2) (page 21). PMI_{LVC} measures the strength of association between the two constituents, considering only those occurrences in the syntactic patterns preferred by LVCs. PMI_{LVC} is defined as the conditional pointwise mutual information between *lv* and *n* given they cooccur in the pattern in $\mathcal{PS}_{pos}$ (see Cover and Thomas, 1991, for details on conditional mutual information):

$$\begin{aligned} \text{PMI}_{\text{LVC}} &\doteq \text{PMI}(lv; n|\,\mathcal{PS}_{pos}) \\ &\doteq \log \frac{P(lv, n|\,\mathcal{PS}_{pos})}{P(lv|\,\mathcal{PS}_{pos})\,P(n|\,\mathcal{PS}_{pos})} \\ &\approx \log \frac{N \times f(lv, n, \mathcal{PS}_{pos})}{f(lv, *, \mathcal{PS}_{pos})\,f(*, n, \mathcal{PS}_{pos})} \end{aligned} \tag{2.5}$$

Both PMI and PMI_{LVC} are appropriate baselines, since the degree of collocation of an expression is expected to also reflect its degree of figurativeness to some extent. This is because the noun constituent of a more figurative expression is often an abstract noun. It is thus ex-

[4] In total, 11 expressions were removed; these are *give effect*, *give a nose*, *give drinking*, *take a meaning*, *take a team*, *take a load*, *take a tack*, *take a line*, *take a hand*, and *take round*.

Table 2.3: Questions for expressions containing *give*.

	Question	Possible answers
As a result of the *event* expressed by the given *expression*:		
I.	Does "SUBJ **transfer** a physical object to AP[a]"?	yes, no, maybe, ?
II.	Does "SUBJ **transfer** something (non-physical) to AP"?	yes, no, maybe, ?
III.	Does "SUBJ **emit** something (non-physical)"?	yes, no, maybe, ?

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

pected to have strong preferences for cooccurring with a light verb (i.e., in the context of an LVC), as opposed to appearing as the direct object of a full verb. Moreover, in such cases the noun is not referential and tends to appear as a bare noun or with an indefinite determiner. PMI_{LVC} is thus an informed baseline because it draws on linguistic properties of LVCs such as their preferred syntactic forms.

2.3.3 Human Judgments of Figurativeness

To provide human judgments on figurativeness, three native speakers of English with sufficient linguistic knowledge answered several yes/no questions for each of the development and test expressions. The questions are devised so that they indirectly capture the degree to which aspects of the literal meaning of the light verb constituent are retained in the meaning of an expression. There are two sets of questions, one for each light verb under study. Table 2.3 presents the questions for expressions involving *give*; questions for expressions involving *take* are presented in Appendix B.

Each possible combination of answers to these questions is transformed into a numerical rating, ranging from 0 (largely figurative) to 4 (largely literal). For example, the combination (**yes**, **no**, **no**) implies the involvement of some kind of physical transfer, and hence translates to a figurativeness rating of 4. the combination (**no**, **no**, **no**) implies that *give* has lost almost

Table 2.4: Distribution of development and test expressions according to human ratings of figurativeness, along with sample expressions.

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

all aspects of its literal semantics, and hence translates to a rating of 0 (e.g., *give a go*). The complete list of all possible combinations of answers to these questions, and the rating each combination translates to are given in Appendix B.

We also requested the judges to provide short paraphrases for each expression; in the final experiments, we only include those expressions for which the majority of the judges expressed the same sense. On the final set of experimental expressions (including both development and test expressions), the three sets of human ratings yield linearly weighted kappa values of .34 and .70 for *give* and *take*, respectively. (We use linearly weighted kappa, since our ratings are ordered.)

In order to perform a plausible evaluation, we need development and test data sets that cover a wide range of figurative and literal usages of the two light verbs under study. However, the original lists of rated expressions were biased toward figurative usages of *give* and *take*. This was due to our selection of the candidate expressions.[5] To achieve a spectrum of literal to figurative usages, we augmented the lists with literal expressions associated an average rating of

[5]In most LVCs of the form V+N, the noun constituent is morphologically related to a verb. Hence, in our selection of experimental expressions, we only selected those that have a related verb according to WordNet. Originally, we added this extra constraint to ensure our candidate lists include LVCs. However, this resulted in the exclusion of most literal combinations.

5 (completely literal).[6] Table 2.4 shows the distribution of the augmented lists of experimental expressions (divided into development and test portions) across three intervals of figurativeness level: 'high' (human ratings ≤ 1), 'medium' ($1 <$ ratings < 3), and 'low' (ratings ≥ 3). The table also contains sample expressions for each figurativeness level. (Note that we do not perform any evaluation on these "bucketized" data sets. This is only to give the reader a feel of the distribution of the experimental expressions with respect to their figurativeness level.)

To form a consensus set to be used for final evaluation, the human ratings are averaged. Note that since we average the values, the consensus rating for an expression may be a non-integer value. The individual and consensus human ratings for all the development and test expressions containing *give* and *take* are given in Appendix C.

2.3.4 Results

We use the Spearman rank correlation coefficient (Siegel and Castellan, 1988), r_s, to compare the ratings assigned by our figurativeness measure to the consensus human ratings. We also compare the "goodness" of $\text{FIGNESS}_{\text{LV}}$ (as determined by the correlation tests) with that of our baselines, PMI and PMI_{LVC}. Table 2.5 displays the correlation scores between the human figurativeness ratings and those assigned by each statistical measure: PMI, PMI_{LVC} and $\text{FIGNESS}_{\text{LV}}$. Scores for the measure with the highest correlations are shown in boldface. In all cases the correlations are statistically significant ($p \ll .01$); we thus omit p values from the table.[7] We report correlation scores not only on our test set (TEST), but also on development and test data combined (DEV+TEST).

As previously noted, there are two different types of expressions in our experimental sets: those that typically include the indefinite determiner *a/an* (e.g., *give a kick*), and those that typically appear without a determiner (e.g., *give guidance*). Despite shared properties, the two

[6] For this purpose, we selected expressions that we were confident to be literal (i.e., to involve some kind of physical transfer); examples include *give a book* and *take a bowl*.

[7] We use R (2004) to calculate r_s and the p values; p values are calculated using a t-test.

Table 2.5: Correlations between human figurativeness ratings and the statistical measures.

Light verb	Data set	(*n*)	PMI	PMI_{LVC}	$FIGNESS_{LV}$
	TEST	(36)	.59	.62	**.66**
give	DEV+TEST	(114)	.60	.68	**.70**
	DEV+TEST/a	(79)	.62	.68	**.77**
	TEST	(34)	.47	.51	**.57**
take	DEV+TEST	(106)	.47	.52	**.56**
	DEV+TEST/a	(68)	.55	.63	**.68**

types of expressions may differ with respect to syntactic flexibility, due to differing semantic properties of the noun complements in the two cases.[8] We thus calculate separate correlation scores for the two types of expressions. Fortunately, the majority of our experimental expressions include a determiner; hence we have sufficient number of data points to get a reliable correlation score and make a comparison. Again, we use expressions from both development and test sets (DEV+TEST/a).

As can be seen in Table 2.5, both the informed baseline, PMI_{LVC}, and our proposed figurativeness measure, $FIGNESS_{LV}$, outperform the simple PMI baseline. These results demonstrate that simply treating LVCs as collocations, without considering their particular linguistic properties, is not sufficient. In fact, we achieve notable improvements in performance by incorporating more and more linguistic information about the syntactic behaviour of LVCs. The highest correlation with the human ratings occurs with the most informed measure, i.e., $FIGNESS_{LV}$. These results reinforce our initial hypothesis, i.e., that the degree of syntactic flexibility of an LV+N combination correlates with the the degree of figurativeness of the light verb constituent.

[8]One of the most important shared properties of the two types of expressions is that in both the noun complement is a predicative nominal. The use of an indefinite determiner or no determiner in an LVC relates to both the semantic characteristics of the predicative noun—e.g., aspectual properties of the state or event expressed by it (Wierzbicka, 1982; Tanabe, 1999), and to the diachronic aspects of LVC formation in English (Hiltunen, 1999). Another motivation for separating the two groups of LVCs is the observation of Hiltunen (1999) that a correlation exists between the choice of the determiner and the type of the noun: zero-derived nouns tend to appear more frequently with the indefinite determiner, while suffixally derived ones often appear without a determiner. Other researchers also have considered these as two subtypes of LVCs (see, e.g., Cattell, 1984; Claridge, 2000). A more detailed discussion of their differences, however, is outside the scope of this study.

Such a measure can thus be used for separating LVCs (metaphorical verb+noun combinations) from compositional verb phrases (literal uses of a verb with a direct object noun).

Table 2.5 also shows that $\text{FIGNESS}_{\text{LV}}$ has higher correlation scores (with large improvements over the baseline) when tested on expressions that typically appear with an indefinite determiner, (i.e., expressions in DEV+TEST/a). These results support our hypothesis on the difference between the syntactic and semantic behaviour of the two types of expressions. Note that the correlation scores are highly significant—very small p values—on both data sets, DEV+TEST and DEV+TEST/a.

2.4 Related Work

A handful of early studies recognized the importance of appropriately handling LVCs as a separate class of MWEs with specific linguistic properties. The main goal in these studies is to find the best choice of light (support) verb for a given predicative noun. Like us, Grefenstette and Teufel (1995) also note that the noun constituent of an LVC is a predicative noun, and hence should have arguments/adjuncts similar to those of the morphologically-related verb. They use this knowledge to extract only those occurrences of the target noun that are more likely to be abstract (hence predicative) usages. We achieve the same goal by restricting the local syntactic patterns that predicative nominals tend to appear in (e.g., predicative nominals are more likely to appear with an indefinite determiner). Dras and Johnson's (1996) approach uses the prior knowledge that certain verbs tend to appear as light verbs more often than others. More specifically, they assume that verbs with higher relative frequency of occurrence in verb–object relations are more productive and hence more likely to form LVCs. It is nonetheless important to note that productivity of a verb is determined by the diversity in the semantic class of the nouns that cooccur with it, and not just by its token frequency. Both Grefenstette and Teufel (1995) and Dras and Johnson (1996) lack a comprehensive evaluation, providing only subjective assessment of their results.

The work most related to ours is that of Villada Moirón (2004), which also draws on the syntactic behaviour of LVCs for their automatic identification. The goal of this study is to determine to what extent a semi-automatic application of linguistic diagnostic tests can help to distinguish (German) LVCs from compositional occurrences of verb and prepositional phrase. The diagnostic tests all relate to the syntactic fixedness of LVCs, i.e., that they tend to appear in certain syntactic constructions, and not in some others. Using human judgments as the gold standard, Villada Moirón manually identifies the linguistic tests capable of correctly distinguishing the majority of a small set of experimental expressions. She then uses these tests to remove noise from a larger set of automatically extracted LVCs. A main disadvantage of this work is that the linguistic tests are considered to be definitive. That is, if a given expression appears in a syntactic pattern prohibited by one of the tests, the expression is considered to be not an LVC. In our work, we address this issue by examining the overall distribution of potential LVCs in preferred and less preferred syntactic patterns, hence making our decision based on probabilistic evidence.

Mason (2004) also looks at metaphorical uses of verbs, although focusing on specific domains. Mason incorporates automatically-induced knowledge about the domain of use of a verb to help identify different metaphorical meanings. It is important to note that highly polysemous verbs that we focus on in this work cannot be easily associated with particular domains. Hence such an approach overlooks the great potential of basic verbs in forming metaphorical multiword expressions.

2.5 Summary of Contributions

The work presented in this chapter differs from related studies, not only in focusing on a mainly unaddressed class of MWUs, but also in presenting a different view on the semantic idiosyncrasy of LVCs. We propose statistical measures that examine the degree to which a light verb usage is “similar” to the prototypical LVC, as an inverse indicator of the degree to which

the light verb retains aspects of its literal semantics. Specifically, our measures assume that the less syntactically flexible a target expression, the more figurative (and less compositional) the use of the verb. The measures identify a continuum of literal to figurative usages of a light verb, that correlates well with the literal–figurative spectrum represented in human judgments, supporting such an approach.

Both our linguistically-motivated measures achieve higher correlations than a simple measure of collocation extraction. Moreover, the best correlation scores belong to the measure that incorporates the most linguistic information about LVCs. These results point out that, as we hypothesized, statistical measures benefit from linguistic knowledge. Our focus thus far has been on a particular coherent class of MWUs with specific linguistic properties. In the following chapter, we further verify this hypothesis by broadening the scope of our study. We cover a more heterogeneous class of MWUs, and propose more general statistical measures that draw on properties such as syntactic fixedness in a more systematic way.

Chapter 3

Idiomatic Combinations

The term *idiom* has been applied to a fuzzy category with prototypical examples such as *by and large, kick the bucket* and *let the cat out of the bag*. It is extremely difficult to provide a definitive answer for what idioms are, and how they are learned and understood (Glucksberg, 1993; Cacciari, 1993; Nunberg et al., 1994). Nonetheless, they are broadly defined as phrases or sentences that involve some degree of lexical, syntactic and/or semantic idiosyncrasy. Idiomatic expressions, as a part of the vast family of figurative language, are widely used both in colloquial speech and in written language. Moreover, a phrase develops its idiomaticity over time (Cacciari, 1993); consequently, new idioms come into existence on a daily basis (Cowie et al., 1983; Seaton and Macaulay, 2002). It is therefore necessary to devise mechanisms for the appropriate handling of idioms within a computational system.

The treatment of completely frozen idioms, such as *by and large*, is straightforward: they can be simply listed in a lexicon as "words with spaces" (Sag et al., 2002). Nonetheless, such idioms are limited in number. The majority of idioms are in the form of syntactically well-formed phrases (phrasal idioms) or sentences (sentential idioms). In particular, many phrasal idioms are formed from the combination of a basic verb with one or more of its arguments (Cowie et al., 1983; Gibbs and Nayak, 1989; d'Arcais, 1993; Nunberg et al., 1994; Fellbaum,

2005).[1] We focus on idioms that involve the combination of a verb and a noun in its direct object position. Examples include *shoot the breeze*, *spill the beans*, *pull strings*, and *push one's luck*. We refer to these as verb+noun idiomatic combinations or VNICs.

VNICs pose a serious challenge, both for the creation of wide-coverage computational lexicons, and for the development of large-scale, linguistically plausible NLP systems (Sag et al., 2002). One problem is due to the range of syntactic idiosyncrasy of these expressions. Some VNICs exhibit limited morphosyntactic flexibility, while others are more versatile in form. Clearly, a words-with-spaces approach does not capture the full range of behaviour of such idiomatic expressions. Another barrier to the appropriate handling of VNICs in a computational system is their semantic idiosyncrasy. These expressions are indistinguishable on the surface from compositional (non-idiomatic) phrases, but a computational system must be capable of distinguishing the two. For example, a machine translation system should consider the idiom *shoot the breeze* as a single unit of meaning ("to chat idly") when translating it to another language, whereas this is not the case for the literal phrase *shoot the bird*.

In this chapter, we look into three closely related problems confronting the appropriate treatment of VNICs: (i) the problem of determining their degree of flexibility; (ii) the problem of determining their level of idiomaticity; and (iii) the problem of determining their canonical forms. Section 3.1 elaborates on the lexicosyntactic flexibility of VNICs, and how this relates to their idiomaticity. In Section 3.2, we propose two linguistically-motivated statistical measures for quantifying the degree of lexical and syntactic inflexibility (or fixedness) of verb+noun combinations. Section 3.3 describes the experimental setup in which we evaluate the proposed measures. Discussion of the results is presented in Section 3.4. In Section 3.5, we put forward a technique for determining the syntactic variations that a VNIC can undergo, hence identifying their canonical forms which are needed for their lexical representation. Section 3.6 provides a survey on related studies, and Section 3.7 concludes the chapter by summarizing

[1] Recall from Chapter 1 that a basic verb refers to a state or act that is central to human experience, and hence is highly frequent.

our contributions to the field.

3.1 Idiomaticity, Semantic Analyzability, and Flexibility

Phrasal idioms (including VNICs), although syntactically well-formed, involve a certain degree of semantic idiosyncrasy. This means that idioms are to some extent opaque or non-transparent, i.e., that their meaning is often hard to guess without special context or previous exposure. There is much evidence in the linguistic literature that the idiosyncrasy of idiomatic combinations is not limited to their semantics, but is also reflected in their lexical and syntactic behaviour. In what follows, we first focus on semantic idiosyncrasy, defining semantic analyzability and its relation to idiomaticity. We then expound on the lexical and syntactic behaviour of VNICs, drawing the attention of the reader to a suggestive relation between the semantic properties of VNICs and their lexicosyntactic behaviour.

3.1.1 Semantic Analyzability

Idioms have been traditionally believed to be completely non-compositional (e.g., Fraser, 1970; Katz, 1973). This means that unlike compositional combinations, the meaning of an idiom cannot be solely predicted from the meaning of its parts. Nonetheless, many linguists and psycholinguists argue against such a view, providing evidence from idioms that exhibit lexical and syntactic behaviour that can only be attributed to phrases with internal semantic structure (e.g., Nunberg et al., 1994; Gibbs, 1995). Researchers who take this alternative view suggest that many idioms in fact do have internal semantic structure and speakers of a language make assumptions about such structure to understand idioms. This new definition is encapsulated in the new terms used by these researchers in place of compositionality, i.e., *semantic decomposability* and/or *semantic analyzability*.

To say that an idiom is semantically analyzable to some extent means that the constituents contribute some sort of independent meaning—not necessarily their literal semantics—to the

overall idiomatic interpretation. Generally, the more semantically analyzable an idiom is, the easier it is to map the idiom constituents onto their corresponding idiomatic referents. In other words, the more semantically analyzable an idiom is, the easier it is to make predictions about the idiomatic meaning from the meaning of the idiom parts. Semantic analyzability is thus inversely related to idiomaticity or semantic opacity. For example, the meaning of *shoot the breeze* ("to chat idly"), a highly idiomatic expression, has nothing to do with either *shoot* or *breeze*. A less idiomatic expression, such as *spill the beans* ("to reveal a secret"), may be analyzed as *spill* metaphorically corresponding to "reveal" and *beans* referring to "secret(s)". An idiom such as *pop the question* is even less idiomatic since the relations between the idiom parts and their idiomatic referents are more directly established, i.e., *pop* corresponds to "suddenly ask" and *question* refers to "marriage proposal".

Many linguists and psycholinguists conclude that idioms clearly form a heterogeneous class, not all of them being truly non-compositional or unanalyzable (see, Abeillé, 1995; Moon, 1998; Grant, 2005, among others). Rather, semantic analyzability in idioms is a matter of degree.

3.1.2 Lexical and Syntactic Flexibility

Most idioms are known to be lexically fixed, meaning that the substitution of a near synonym (or a closely-related word) for a constituent part does not preserve the idiomatic meaning of the expression. For example, neither *shoot the wind* nor *hit the breeze* are valid variations of the idiom *shoot the breeze*. Similarly, *spill the beans* has an idiomatic meaning, while *spill the peas* and *spread the beans* have only literal interpretations. There are, however, idiomatic expressions that have one (or more) lexical variants. For example, *blow one's own trumpet* and *toot one's own horn* have the same idiomatic interpretation (Cowie et al., 1983); also *keep one's cool* and *lose one's cool* have closely related meanings (Nunberg et al., 1994). Nonetheless, it is not the norm for idioms to have lexical variants; when they do, there are usually unpredictable restrictions on the substitutions they allow.

Idiomatic combinations are also syntactically distinct from compositional combinations. Many VNICs cannot undergo syntactic variations and at the same time retain their idiomatic interpretations. It is important, however, to note that VNICs differ with respect to the extent to which they can tolerate syntactic operations, i.e., the degree of syntactic flexibility they exhibit. Some are syntactically inflexible for the most part, while others are more versatile; as illustrated in the sentences in 10 and 11:

10. (a) Sam and Azin shot the breeze.
    (b) ?? Sam and Azin shot a breeze.
    (c) ?? Sam and Azin shot the breezes.
    (d) ?? Sam and Azin shot the casual breeze.
    (e) ?? The breeze was shot by Sam and Azin.
    (f) ?? The breeze that Sam and Azin shot was quite refreshing.
    (g) ?? Which breeze did Sam and Azin shoot?

11. (a) Azin spilled the beans.
    (b) ? Azin spilled some beans.
    (c) ?? Azin spilled the bean.
    (d) Azin spilled the official beans.
    (e) The beans were spilled by Azin.
    (f) The beans that Azin spilled caused Sam a lot of trouble.
    (g) Which beans did Azin spill?

Linguists have often explained the lexical and syntactic flexibility of idiomatic combinations in terms of their semantic analyzability (e.g., Gibbs, 1993; Glucksberg, 1993; Fellbaum, 1993; Nunberg et al., 1994). The common belief is that because the constituents of a semantically analyzable idiom can be mapped onto their corresponding referents in the idiomatic interpretation, analyzable (less idiomatic) expressions are often more open to lexical substitution and syntactic variation. Psycholinguistic studies also support this hypothesis: Gibbs

and Nayak (1989) and Gibbs et al. (1989), through a series of psychological experiments, demonstrate that there is variation in the degree of lexicosyntactic flexibility of idiomatic combinations. (Both studies narrow their focus to verb phrase idiomatic combinations, mainly of the form verb+noun.) Moreover, their findings provide evidence that the lexical and syntactic flexibility of VNICs are not arbitrary phenomena, but rather are influenced by the speakers' assumptions about the semantic analyzability of these idioms.

Corpus-based studies such as those by Moon (1998) and Grant (2005) conclude that idioms are neither lexically nor syntactically fixed. Nonetheless, their claim is often based on observing certain idiomatic combinations in a form other than their so-called canonical forms. For example, Moon mentions that she has observed both *kick the pail* and *kick the can* as variations of *kick the bucket*. Also, Grant finds evidence of variations such as *eat one's heart (out)* and *eat one's hearts (out)* in the BNC. It is important to note that most such observed variations are constrained, often with unpredictable restrictions. Moreover, our understanding from such claims is that idiomatic combinations are not inherently frozen and that it is possible for them to appear in forms other than their agreed-upon canonical forms.

We are well aware that semantic analyzability is neither a necessary nor a sufficient condition for an idiomatic combination to be lexically or syntactically flexible. Other factors, such as communicative intentions and pragmatic constraints, can motivate a speaker to use a variant in place of a canonical form (Glucksberg, 1993). For example, journalism is well-known for manipulating idiomatic expressions for humour or cleverness (Grant, 2005). The age and the degree of familiarity of an idiom have also been shown to be important factors that affect its flexibility (Gibbs and Nayak, 1989).

Nonetheless, lexicosyntactic behaviour of a VNIC, although affected by historical and pragmatic factors, can be at least partially explained in terms of semantic analyzability or idiomaticity. Many linguists use observations about lexical and syntactic flexibility of VNICs in order to make judgments about their degree of idiomaticity (e.g., Tanabe, 1999; Kytö, 1999).

3.2 Automatic Recognition of VNICs

We use the observed connection between idiomaticity and (in)flexibility to devise statistical measures for automatically distinguishing idiomatic from literal verb+noun combinations. While VNICs vary in their degree of flexibility (cf. 1 and 2 above), on the whole they contrast with compositional phrases, which are more lexically productive and appear in a wider range of syntactic forms. We thus propose to use the degree of lexical and syntactic flexibility of a given verb+noun combination to determine the level of idiomaticity of the expression.

Note that our assumption here is in line with corpus-linguistic studies on idioms: we do not claim that it is inherently impossible for VNICs to undergo lexical substitution or syntactic variation. In fact, for each given idiomatic combination, it may well be possible to find a specific situation in which a lexical or a syntactic variant of the canonical form is perfectly plausible. However, the main point of the assumption here is that VNICs are more likely to appear in fixed forms (known as their canonical forms), more so than compositional phrases. Therefore, the overall distribution of a VNIC in different lexical and syntactic forms is expected to be notably different from the corresponding distribution of a typical verb+noun combination.

The following subsections describe our proposed statistical measures for idiomaticity, which quantify the degree of lexical, syntactic, and overall fixedness of a given verb+noun combination, represented as a verb–noun pair.

3.2.1 Measuring Lexical Fixedness

A VNIC is lexically fixed if the replacement of any of its constituents by a semantically (and syntactically) similar word does not generally result in another VNIC, but in an invalid or a literal expression. One way of measuring lexical fixedness of a given verb+noun combination is thus to examine the idiomaticity of its variants, i.e., expressions generated by replacing one of the constituents by a similar word. This approach has two main challenges: (i) it requires prior knowledge about the idiomaticity of expressions (which is what we are developing our

measure to determine); (ii) it needs information on "similarity" among words.

Inspired by Lin (1999), we examine the strength of association between the verb and the noun constituent of a combination (the target combination or its variants) as an indirect cue to its idiomaticity. We use the automatically-built thesaurus of Lin (1998) to find words similar to each constituent, in order to automatically generate variants.[2] Variants are generated by replacing either the noun or the verb constituent of a pair with a semantically (and syntactically) similar word.[3] Examples of automatically generated variants for the pair $\prec$*spill, bean*$\succ$ are $\prec$*pour, bean*$\succ$, $\prec$*stream, bean*$\succ$, $\prec$*spill, corn*$\succ$, and $\prec$*spill, rice*$\succ$.

Let $\mathcal{S}_{sim}(v) = \{v_i \,|\, 1 \leq i \leq K_v\}$ be the set of the K_v most similar verbs to the verb v of the target pair $\prec v, n \succ$, and $\mathcal{S}_{sim}(n) = \{n_j \,|\, 1 \leq j \leq K_n\}$ be the set of the K_n most similar nouns to the noun n (according to Lin's thesaurus). The set of variants for the target pair is thus:

$$\mathcal{S}_{sim}(v,n) = \{\prec v_i, n \succ, \prec v, n_j \succ \;|\, 1 \leq i \leq K_v \wedge 1 \leq j \leq K_n\}$$

We calculate the association strength for the target pair and for each of its variants using pointwise mutual information (PMI) (Church et al., 1991):

$$\begin{aligned} \mathrm{PMI}(v_k, n_t) &= \log \frac{P(v_k, n_t)}{P(v_k)\,P(n_t)} \\ &= \log \frac{f(*,*)\, f(v_k, n_t)}{f(v_k, *)\, f(*, n_t)} \end{aligned} \tag{3.1}$$

where $\prec v_k, n_t \succ \in \{\prec v, n \succ\} \cup \mathcal{S}_{sim(v,n)}$; $f(v_k, n_t)$ is the frequency of v_k and n_t cooccurring as a verb–object pair; $f(v_k, *)$ is the total frequency of the target verb with any noun; $f(*, n_t)$ is the total frequency of the noun n_t in the direct object position of any verb; and $f(*,*)$ is the total number of verb–object pairs in the corpus.

[2] We also replicated our experiments with an automatically-built thesaurus created from the BNC in a similar fashion, and kindly provided to us by Diana McCarthy. Results were more or less similar, hence we do not report them here.

[3] In an early version of this work (Fazly and Stevenson, 2006), only the noun constituent was varied since we expected replacing the verb constituent with a related verb to be more likely to yield another VNIC, as in *keep/lose one's cool*, *give/get the bird*, *crack/break the ice* (according to Nunberg et al., 1994; Grant, 2005). Later experiments on the development data showed that variants generated by replacing both constituents, one at a time, produce better results.

In his work, Lin (1999) assumes that a target expression is non-compositional if and only if its PMI value is significantly different from that of any of the variants. Instead, we propose a novel technique that brings together the association strengths (PMI values) of the target and the variant expressions into a single measure reflecting the degree of lexical fixedness for the target pair. We assume that the target pair is lexically fixed to the extent that its PMI deviates from the average PMI of its variants. By our measure, the target pair is considered lexically fixed only if the difference between its PMI value and that of most of its variants (not necessarily all) is high. Our measure calculates this deviation, normalized using the sample's standard deviation:

$$\text{Fixedness}_{\text{lex}}(v, n) \doteq \frac{\text{PMI}(v, n) - \overline{\text{PMI}}}{s} \tag{3.2}$$

where $\overline{\text{PMI}}$ is the mean and s the standard deviation of the sample:

$$\left\{\text{PMI}(v_k, n_t) \mid \prec v_k, n_t \succ \in \{\prec v, n \succ\} \cup \mathcal{S}_{sim(v,n)}\right\}$$

PMI can be negative, zero, or positive; thus $\text{Fixedness}_{\text{lex}}(v, n) \in [-\infty, +\infty]$.

3.2.2 Measuring Syntactic Fixedness

Compared to compositional verb+noun combinations, VNICs are expected to appear in more restricted syntactic forms. To quantify the syntactic fixedness of a target verb–noun pair, we thus need to: (i) identify relevant syntactic patterns, i.e., those that help distinguish VNICs from literal verb+noun combinations; (ii) translate the frequency distribution of the target pair in the identified patterns into a measure of syntactic fixedness.

3.2.2.1 Identifying Relevant Patterns

Determining a unique set of syntactic patterns appropriate for the recognition of all idiomatic combinations is difficult indeed: exactly which forms an idiomatic combination can occur in is not entirely predictable (Sag et al., 2002). Nonetheless, there are hypotheses about the

difference in behaviour of VNICs and literal verb+noun combinations with respect to particular syntactic variations (Nunberg et al., 1994). Linguists note that semantic analyzability of VNICs is related to the referential status of the noun constituent, which is in turn related to participation in certain morphosyntactic forms. In what follows, we describe three types of variation that are assumed to be tolerated by literal combinations, but are prohibited by many VNICs.

Passivization: There is much evidence in the linguistic literature that VNICs often do not undergo passivization.[4] Linguists mainly attribute this to the fact that only a referential noun can appear as the surface subject of a passive construction (e.g. Gibbs and Nayak, 1989). Due to the non-referential status of the noun constituent in most VNICs, we expect that they do not undergo passivization as often as compositional verb+noun combinations do. Another explanation for this assumption is that passives are mainly used to put focus on the object of a clause or sentence. For most VNICs, no such communicative purpose can be served by topicalizing the noun constituent through passivization. The passive construction is thus considered as one of the syntactic patterns relevant to measuring syntactic flexibility.

Determiner type: A strong correlation has been observed between the flexibility of the determiner preceding the noun in a verb+noun combination and the overall flexibility of the phrase (Fellbaum, 1993; Kearns, 2002; Desbiens and Simon, 2003). It is however important to note that the nature of the determiner is also affected by other factors, such as the semantic properties of the noun. For this reason, determiner flexibility is sometimes argued not to be a good predictor of the overall syntactic flexibility of an expression. Nonetheless, many researchers consider it as an important part in the process of idiomatization of a verb+noun combination (Akimoto, 1999; Kytö, 1999; Tanabe, 1999).

[4]There are idiomatic combinations that are used only in a passivized form; we do not consider such cases in our study.

Table 3.1: Patterns used in the syntactic fixedness measure, along with examples for each.

1	v_{act}	det:NULL	n_{sg}	(*give money*)
2	v_{act}	det:*a/an*	n_{sg}	(*give a book*)
3	v_{act}	det:*the*	n_{sg}	(*give the book*)
4	v_{act}	det:DEM	n_{sg}	(*give this book*)
5	v_{act}	det:POSS	n_{sg}	(*give my book*)
6	v_{act}	det:NULL	n_{pl}	(*give books*)
7	v_{act}	det:*the*	n_{pl}	(*give the books*)
8	v_{act}	det:DEM	n_{pl}	(*give those books*)
9	v_{act}	det:POSS	n_{pl}	(*give my books*)
10	v_{act}	det:OTHER	$n_{sg,pl}$	(*give many books*)
11	det:ANY	$n_{sg,pl}$	v_{pass}	(*a book/books was/were given*)

Pluralization: While the verb constituent of a VNIC is morphologically flexible, the morphological flexibility of the noun relates to its referential status (Grant, 2005). Again, one should note that the use of a singular or plural noun in a VNIC may also be affected by the semantic properties of the noun. Nonetheless, the process of idiomatization of a verb+noun combination is believed to be accompanied by a change from concreteness to abstractness for the noun. In this process, the noun constituent loses some of its nominal features, including number (Akimoto, 1999). The non-referential noun constituent of a VNIC is thus expected to mainly appear in just one of the singular or plural forms.

Merging the three types of variation results in a pattern set, $\mathcal{PS}$, of 11 distinct syntactic patterns that are displayed in Table 3.1.[5] The table also contains examples for each pattern, given in parentheses. Note that we merge some of the individual patterns into one, e.g., we include only one passive pattern independently of the choice of the determiner or the number of the noun. The motivation here is to merge low frequency patterns in order to acquire more reliable evidence on the distribution of a particular verb–noun pair over the resulting pattern set.

[5]Our choice of patterns is consistent with the idiom typology developed by Nicolas (1995).

In principle, however, the set can be expanded to include more patterns; it can also be modified to contain different patterns for a different class of idiomatic combinations. Nonetheless, the choice of the individual patterns may affect the results in unpredictable ways.

3.2.2.2 Devising a Statistical Measure

The second step is to devise a statistical measure that quantifies the degree of syntactic fixedness of a verb–noun pair, with respect to the selected set of patterns, $\mathcal{PS}$. We propose a measure that compares the "syntactic behaviour" of the target pair with that of a "typical" verb–noun pair. Syntactic behaviour of a typical pair is defined as the prior probability distribution over the patterns in $\mathcal{PS}$. The prior probability of an individual pattern $pt \in \mathcal{PS}$ is estimated as:

$$\begin{aligned} P(pt) &= \frac{\sum_{v_i \in \mathcal{V}} \sum_{n_j \in \mathcal{N}} f(v_i, n_j, pt)}{\sum_{v_i \in \mathcal{V}} \sum_{n_j \in \mathcal{N}} \sum_{pt_k \in \mathcal{PS}} f(v_i, n_j, pt_k)} \\ &= \frac{f(*, *, pt)}{f(*, *, *)} \end{aligned}$$

where V and N are the sets of all verbs and all nouns under consideration, respectively.

The syntactic behaviour of the target verb–noun pair $\prec v, n \succ$ is defined as the posterior probability distribution over the patterns, given the particular pair. The posterior probability of an individual pattern pt is estimated as:

$$\begin{aligned} P(pt|v, n) &= \frac{P(v, n, pt)}{P(v, n)} \\ &= \frac{f(v, n, pt)}{\sum_{pt_k \in \mathcal{PS}} f(v, n, pt_k)} \\ &= \frac{f(v, n, pt)}{f(v, n, *)} \end{aligned}$$

The degree of syntactic fixedness of the target verb–noun pair is estimated as the divergence of its syntactic behaviour (the posterior distribution over the patterns), from the typical syntac-

tic behaviour (the prior distribution). The divergence of the two probability distributions is calculated using a standard information-theoretic measure, the Kullback Leibler (KL-)divergence:

$$\begin{aligned} \text{Fixedness}_{\text{syn}}(v,n) &\doteq D(P(pt|v,n)\,||\,P(pt)) \\ &= \sum_{pt_k \in \mathcal{PS}} P(pt_k|\,v,n) \log \frac{P(pt_k|\,v,n)}{P(pt_k)} \end{aligned} \tag{3.3}$$

KL-divergence is always non-negative and is zero if and only if the two distributions are exactly the same. Thus, $\text{Fixedness}_{\text{syn}}(v,n) \in [0,+\infty]$.

KL-divergence is argued to be problematic, partly because it is not a symmetric measure. Nonetheless, it has proven useful in many NLP applications (Resnik, 1999; Dagan et al., 1994). Moreover, here we are concerned with the relative distance of several posterior distributions from the same prior distribution, and hence the asymmetry is not an issue.

3.2.3 A Hybrid Measure of Fixedness

VNICs are hypothesized to be, in most cases, both lexically and syntactically more fixed than literal verb+noun combinations (see Section 3.1). We thus propose a new measure of idiomaticity to be a measure of the overall fixedness of a given pair. We define $\text{Fixedness}_{\text{overall}}(v,n)$ as a weighted combination of $\text{Fixedness}_{\text{lex}}$ and $\text{Fixedness}_{\text{syn}}$:

$$\text{Fixedness}_{\text{overall}}(v,n) \doteq \alpha\,\text{Fixedness}_{\text{syn}}(v,n) + (1-\alpha)\,\text{Fixedness}_{\text{lex}}(v,n) \tag{3.4}$$

where α weights the relative contribution of the measures in predicting idiomaticity.

Recall that $\text{Fixedness}_{\text{lex}}(v,n) \in [-\infty,+\infty]$, and $\text{Fixedness}_{\text{syn}}(v,n) \in [0,+\infty]$. To combine them in the overall fixedness measure, we normalize them, so that they fall in the range $[0,1]$. Thus, $\text{Fixedness}_{\text{overall}}(v,n) \in [0,1]$.

3.3 Experimental Setup

To evaluate our proposed fixedness measures, we determine their appropriateness as indicators of idiomaticity. This is done through applying each measure to each of two different tasks: (i) a classification task in which idiomatic verb–noun pairs are distinguished from literal ones; (ii) a retrieval task in which idiomatic verb–noun pairs are retrieved from a mixed list of idiomatic and non-idiomatic pairs. We first use each measure to assign scores to the experimental pairs (see Section 3.3.2 below). In the classification task, we use the assigned scores to label each pair as idiomatic or literal. This is done by setting a threshold, here the median score, where all pairs with scores higher than the threshold are labeled as idiomatic and the rest as literal. We then determine the accuracy of each measure in separating idioms from non-idioms. In the retrieval task, we use the assigned scores to rank the pairs with respect to their idiomaticity. We then examine the precision–recall curves of each measure to determine its effectiveness in ranking idiomatic pairs before non-idiomatic ones.

We assess the overall goodness of a measure by looking at its performance at both tasks described above. For the classification task, we report accuracy (*Acc*) and the relative reduction in error rate (*RER*). Accuracy of a measure is defined as the percentage of cases that the measure labels correctly. The *RER* of a measure reflects the improvement in its accuracy relative to another measure, often a baseline. We consider two baselines: (i) a random baseline, Rand, that randomly assigns a label (literal or idiomatic) to each verb–noun pair; (ii) a more informed baseline, PMI, an information-theoretic measure widely used for extracting statistically significant collocations.[6]

For the retrieval task, we present the precision–recall curves. We also report the interpolated 3-point average precision (*IAP*), a standard performance measure used in the evaluation of information retrieval systems (Manning and Schütze, 1999). The output of each idiomaticity measure is a ranked list of verb–noun pairs. For a good measure, we expect idiomatic pairs to

[6] As in Eqn. (3.1), our calculation of PMI here restricts the counts of the verb–noun pair to the direct object relation.

be frequent near the top of the list, and become less frequent towards the bottom. The average precision is thus more informative than the precision at a particular cut-off. We compare the performance of the fixedness measures with that of our informed baseline, PMI.

The rest of this section elaborates on the methodological aspects of our experiments. Results are presented in the following section.

3.3.1 Corpus and Data Extraction

We use the British National Corpus (BNC, 2000), to extract verb–noun pairs, along with information on the syntactic patterns they appear in. We automatically parse the BNC using the Collins parser (Collins, 1999). We further process the corpus, using TGrep2 (Rohde, 2004), in order to extract syntactic dependencies. For each instance of a transitive verb, we use heuristics to extract the noun phrase (NP) in either the direct object position (if the sentence is active), or the subject position (if the sentence is passive). We then use NP-head extraction software[7] to get the head noun of the extracted NP, its number (singular or plural), and the determiner introducing it.

3.3.2 Experimental Expressions

We select our development and test expressions from verb–noun pairs that involve a member of a predefined list of basic (transitive) verbs. Recall that basic verbs, in their literal use, refer to states or acts that are central to human experience. They are thus frequent, highly polysemous, and tend to combine with other words to form idiomatic combinations. An initial list of such verbs was selected from several linguistic and psycholinguistic studies on basic vocabulary (Ogden, 1968; Clark, 1978; Nunberg et al., 1994; Goldberg, 1995; Pauwels, 2000; Claridge, 2000; Newman and Rice, 2004). We further augmented this initial list with verbs that are semantically related to another verb already in the list; e.g., *lose* is added in analogy with *find*.

[7] We use a modified version of the software provided by Eric Joanis based on heuristics from (Collins, 1999).

Here is the final list of the 28 verbs in alphabetical order:

> *blow, bring, catch, cut, find, get, give, have, hear, hit, hold, keep, kick, lay, lose, make, move, place, pull, push, put, see, set, shoot, smell, take, throw, touch*

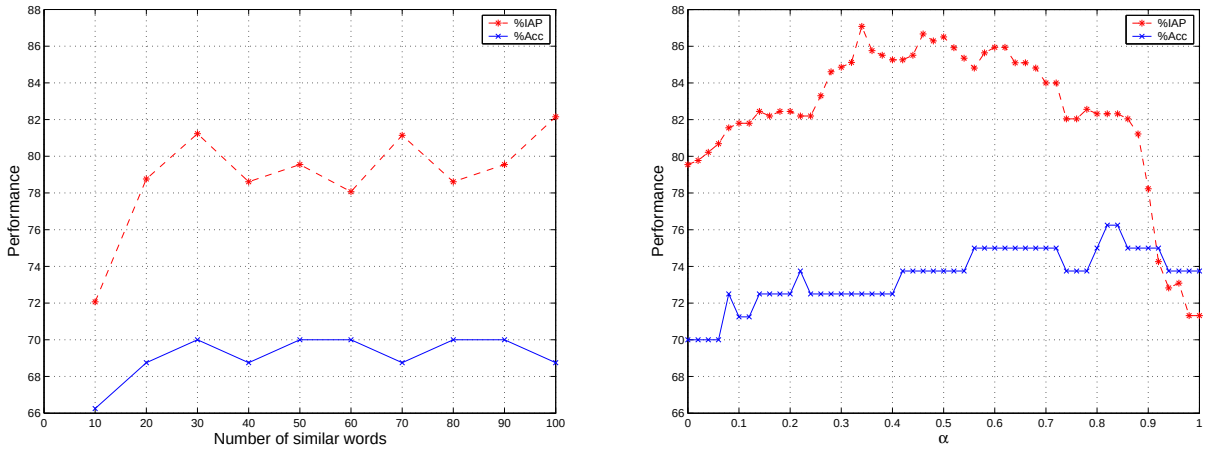
From the corpus, we extract all the verb–noun pairs with minimum frequency of 10 that contain any of the above-listed basic verbs. From these, we semi-randomly select a subset that are idiomatic, and another subset that are literal. A particular verb–noun pair is considered idiomatic if it appears in an idiom listed in a credible dictionary such as the Oxford Dictionary of Current Idiomatic English (ODCIE) (Cowie et al., 1983), or the Collins COBUILD Idioms Dictionary (CCID) (Seaton and Macaulay, 2002). The pair is considered literal if it involves a physical act (i.e., the basic semantics of the verb) and does not appear in any of the above-mentioned dictionaries as an idiom. From the set of idiomatic pairs, we then randomly pull out 80 development pairs, and 100 test pairs, ensuring that we have items of both low and high frequency. We then double the size of each data set (development and test) by adding equal number of literal pairs, with similar frequency distributions. Some of the idioms corresponding to the experimental idiomatic pairs are: *kick the habit*, *move mountains*, *lose face*, and *keep one's word*.

Development expressions are used in devising the fixedness measures, as well as in determining the values of the parameters α in Eqn. (3.4), and K_v and K_n needed for measuring lexical fixedness as in Eqn. (3.2). Test expressions are saved as unseen data for the final evaluation; they are given in Appendix D.

3.3.3 Parameters

Our lexical fixedness measure, as defined in Eqn. (3.2), includes two parameters, K_v and K_n, where $K_v + K_n$ is the total number of variants considered in measuring the lexical fixedness of a given verb–noun pair.[8] In our experiments, we assume $K_v = K_n$, hence we only need to

[8]Note that K_v and K_n are not explicit in Eqn. (3.2), but they are part of the definition of $\mathcal{S}_{sim}(v,n)$.



(a) Performance of $\text{Fixedness}_{\text{lex}}$ as a function of M. (b) Performance of $\text{Fixedness}_{\text{overall}}$ as a function of α.

Figure 3.1: %*IAP* and %*Acc* of $\text{Fixedness}_{\text{lex}}$ and $\text{Fixedness}_{\text{overall}}$ over development data.

know the total number of variants, referred to as M. The value of M is determined empirically, by performing experiments over the development data, in which M ranges from 10 to 100 by steps of 10. Figure 3.1(a) shows the change in performance of $\text{Fixedness}_{\text{lex}}$ as M changes. According to these results, there is not much variation in the performance of the measure for $M \in [30, 90]$. We thus choose an intermediate value for M that yields the highest accuracy and a reasonably high precision, i.e., we set M to 50.

The overall fixedness measure defined in Eqn. (3.4) also uses a parameter, α. To determine the best value for this parameter, we experiment with different values of α ranging from 0 to 1 by steps of .02. As shown in Figure 3.1(b), $\text{Fixedness}_{\text{overall}}$ reaches its best performance (as determined by both *Acc* and *IAP*) when α is set to .6, giving slightly more weight to syntactic fixedness.

3.4 Results

We report the results of evaluating our measures on unseen test expressions, with parameters set to the values determined in Section 3.3.3 above. (Results on development data have similar trends to those on test data.) For analytical purposes, we further divide the set of all test expres-

Table 3.2: Accuracy and relative error reduction for the two fixedness and the two baseline measures over all test pairs (TEST_{all}), and test pairs divided by frequency ($\text{TEST}_{f_{\text{low}}}$ and $\text{TEST}_{f_{\text{high}}}$).

	TEST_{all}		$\text{TEST}_{f_{\text{low}}}$		$\text{TEST}_{f_{\text{high}}}$	
Measure	%*Acc*	(%*RER*)	%*Acc*	(%*RER*)	%*Acc*	(%*RER*)
Rand	50	-	50	-	50	-
PMI	63	(26)	56	(12)	70	(40)
$\text{Fixedness}_{\text{lex}}$	68	(36)	70	(40)	70	(40)
$\text{Fixedness}_{\text{syn}}$	**71**	(**42**)	**72**	(**44**)	**82**	(**64**)

sions, TEST_{all}, into two sets corresponding to two frequency bands: $\text{TEST}_{f_{\text{low}}}$ contains 50 idiomatic and 50 literal pairs, each with total frequency between 10 and 40 ($10 \leq \mathit{freq}(v, n, *) < 40$); $\text{TEST}_{f_{\text{high}}}$ consists of 50 idiomatic and 50 literal pairs, each with total frequency of 40 or greater ($\mathit{freq}(v, n, *) \geq 40$). We first present the results on the classification task, and then discuss the effectiveness of the measures in the retrieval task.

3.4.1 Classification Performance

We first look into the performance of the individual fixedness measures, $\text{Fixedness}_{\text{lex}}$ and $\text{Fixedness}_{\text{syn}}$, as well as that of the two baselines, Rand and PMI. Results for the overall fixedness measure are presented later in this section. As can be seen in the first two columns of Table 3.2, the informed baseline, PMI, shows a large improvement over the random baseline (26% error reduction) on all test pairs. This shows that many VNICs have turned into institutionalized (i.e., statistically significant) cooccurrences. Hence, one can get relatively good performance by treating verb+noun idiomatic combinations as collocations.

$\text{Fixedness}_{\text{lex}}$ performs considerably better than the informed baseline (36% vs. 26% error reduction on all test pairs). $\text{Fixedness}_{\text{syn}}$ has the best performance (shown in boldface), with 42% error reduction over the random baseline, and 21.6% error reduction over the informed baseline. These results demonstrate that lexical and syntactic fixedness are good indicators of idiomaticity, better than a simple measure of collocation (PMI). The results further suggest

Table 3.3: Performance of the hybrid measure over TEST_{all}.

	TEST_{all}	
Measure	%*Acc*	(%*RER*)
$\text{Fixedness}_{\text{lex}}$	68	(36)
$\text{Fixedness}_{\text{syn}}$	71	(42)
$\text{Fixedness}_{\text{overall}}$	**74**	(**48**)

that looking into deep linguistic properties of VNICs is both necessary and beneficial for the appropriate treatment of these expressions.

PMI is known to perform poorly on low frequency items. To examine the effect of frequency on the measures, we analyze their performance on the two divisions of the test data, corresponding to the two frequency bands, $\text{TEST}\,f_{\text{low}}$ and $\text{TEST}\,f_{\text{high}}$. Results are given in the four rightmost columns of Table 3.2, with the best performance shown in boldface.

As expected, the performance of PMI drops substantially for low frequency items. Interestingly, although it is a PMI-based measure, $\text{Fixedness}_{\text{lex}}$ performs slightly better when the data is separated based on frequency. The performance of $\text{Fixedness}_{\text{syn}}$ improves quite a bit when it is applied to high frequency items, while it improves only slightly on the low frequency items. The results show that both fixedness measures perform better on homogeneous data, while retaining comparably good performance on heterogeneous data, suggesting that these measures are not as sensitive to frequency as PMI. Hence they can be used with a higher degree of confidence, especially when applied to data that is heterogeneous with regard to frequency. This is important because while some VNICs are very common, others have very low frequency (see Grant, 2005, for a detailed look at the frequency of idioms in the BNC).

We now look at the performance of the hybrid fixedness measure. Table 3.3 presents the performance of $\text{Fixedness}_{\text{overall}}$, repeating that of $\text{Fixedness}_{\text{lex}}$ and $\text{Fixedness}_{\text{syn}}$ for comparison. Here again the error reductions are relative to the random baseline. $\text{Fixedness}_{\text{overall}}$ outperforms both lexical and syntactic fixedness measures, with notable improvements over the

individual fixedness measures (18.8% error reduction relative to $\text{Fixedness}_{\text{lex}}$, and 10% error reduction relative to $\text{Fixedness}_{\text{syn}}$). According to the classification results, each of the lexical and syntactic fixedness measures are good at separating idiomatic from literal combinations, with syntactic fixedness performing better. Here we demonstrate that combining them into a single measure of fixedness, while giving more weight to the better measure, results in a more effective classifier. The overall behaviour of this measure as a function of α is displayed in Figure 3.2.

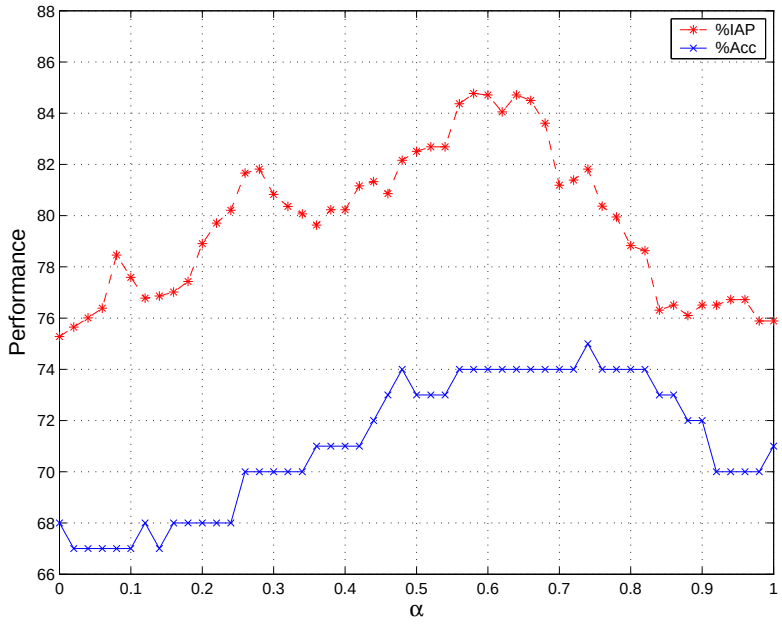


Figure 3.2: Performance of $\text{Fixedness}_{\text{overall}}$ on test data as a function of α.

3.4.2 Retrieval Performance

The classification results suggest that the fixedness measures are better than a simple measure of collocation at separating idiomatic pairs from literal ones. Nonetheless, a long-term goal is to use these measures for the more difficult task of extracting VNICs. In this section, we evaluate the goodness of our fixedness measures in ranking verb+noun combinations according to their degree of idiomaticity. The fixedness measures are devised to reflect the degree of

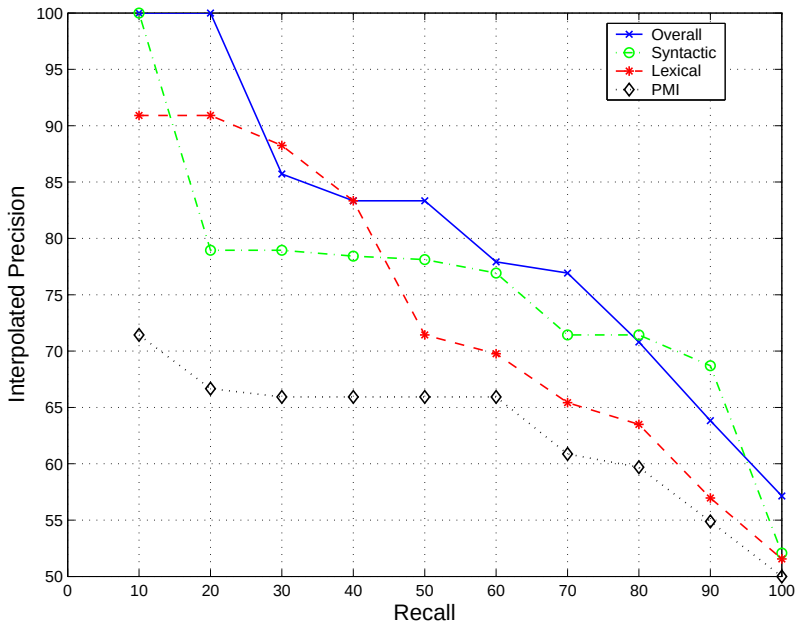


Figure 3.3: Precision–recall curves for PMI and for the fixedness measures.

fixedness and hence the degree of idiomaticity of a target verb–noun pair. Thus, the result of applying each measure to a list of mixed pairs is a list, ranked in the order of idiomaticity. For a good measure, we expect idiomatic pairs to be very frequent near the top of the ranked list, and become less frequent towards the bottom. Precision–recall curves are well indicative of this trend.

The interpolated precision–recall curves for PMI and for the lexical, syntactic, and overall fixedness measures are depicted in Figure 3.3. Note that the minimum interpolated precision is 50% due to the equal number of idiomatic and literal pairs in the test data. For the ease of comparison, we also report, in Table 3.4, the interpolated 3-point average precision values (*IAP*) for all measures. *IAP* is the average of the interpolated precisions at the recall levels of 20%, 50% and 80%.

The precision–recall curve of PMI is nearly flat, with an *IAP* of only 63.5%, showing that the distribution of idiomatic pairs in the list ranked by this measure is nearly random. In comparison, *IAP* of $\text{Fixedness}_{\text{lex}}$ is substantially higher, 75.3%. A closer look at the precision–

Table 3.4: Interpolated 3-point average precision for PMI and the fixedness measures over TEST_{all}.

Measure	%*IAP*
PMI	63.5
$\text{Fixedness}_{\text{lex}}$	75.3
$\text{Fixedness}_{\text{syn}}$	75.9
$\text{Fixedness}_{\text{overall}}$	**84.7**

recall curve of $\text{Fixedness}_{\text{lex}}$ reveals that up to the recall level of 50%, the precision of this measure is substantially higher than that of PMI. This means that, compared to PMI, $\text{Fixedness}_{\text{lex}}$ places more idiomatic pairs at the very top of the list. Nonetheless, the comparable and low precision of the two measures at the higher recall levels (50% and higher) suggests that both measures also put many literal pairs before some of the idiomatic ones.

As can be seen in Table 3.4, both $\text{Fixedness}_{\text{syn}}$ and $\text{Fixedness}_{\text{overall}}$ have high *IAP*: 75.9% and 84.7%, respectively. The precision–recall curve of $\text{Fixedness}_{\text{syn}}$ shows an interesting behaviour of this measure, i.e, that $\text{Fixedness}_{\text{syn}}$ maintains high precision at very high levels of recall (e.g., its precision is close to 70% at the recall level of 90%). Up to the recall level of 50%, $\text{Fixedness}_{\text{overall}}$ has substantially higher precision than $\text{Fixedness}_{\text{syn}}$. At the higher recall levels, their precision remains comparable and reasonably high, suggesting that the two measures are good at retrieving the desired items (idiomatic pairs) from a mixed set. For comparison purposes, the top and bottom 30 pairs in the lists ranked by PMI and $\text{Fixedness}_{\text{overall}}$ are given in Appendix E.

3.4.3 Summary of Results

Overall, the worst performance belongs to PMI, both in classifying test pairs as idiomatic or literal, and in ranking the pairs according to their degree of idiomaticity. This suggests that although some VNICs are institutionalized, many do not appear with markedly high frequency, and hence only looking at their frequency is not sufficient for their recognition. $\text{Fixedness}_{\text{overall}}$

is the best performer of all, supporting the hypothesis that many VNICs are both lexically and syntactically fixed, more so than compositional verb+noun combinations. In addition, these results demonstrate that incorporating such linguistic properties into statistical measures is beneficial for the recognition of VNICs.

Although we focus on experimental expressions with frequency higher than 10, PMI still shows great sensitivity to frequency differences, performing especially poorly on items with frequency between 10 and 40. In contrast, none of the fixedness measures are as sensitive to such frequency differences. Especially interesting is the consistent performance of $\text{Fixedness}_{\text{lex}}$, which is a PMI-based measure, on low and high frequency items. These observations put further emphasis on the importance of devising new alternative methods for extracting multiword expressions with particular syntactic and semantic properties, such as VNICs.

3.5 Determining the Canonical Forms of Idioms

Our evaluation of the fixedness measures demonstrates their usefulness for the automatic recognition of idiomatic verb–noun pairs. To represent such pairs in a lexicon, however, we must turn them into full expressions, i.e., we must find their canonical form(s), henceforth referred to as Cforms. For example, the lexical representation of $\prec$*shoot*, *breeze*$\succ$ should include *shoot the breeze* as a Cform.

Since VNICs are syntactically fixed, they are mostly expected to have a single Cform. Nonetheless, there are idioms with two or more acceptable forms. For example, *hold fire* and *hold one's fire* are both listed in CCID as variations of the same idiom. Our approach should thus be capable of predicting all allowable forms for a given idiomatic verb–noun pair.

We expect a VNIC to occur in its Cform(s) more frequently than it occurs in any other syntactic patterns. To discover the Cform(s) for a given idiomatic verb–noun pair, we thus examine its frequency of occurrence in each syntactic pattern in $\mathcal{PS}$. Since it is possible for an idiom to have more than one Cform, we cannot simply take the most dominant pattern as

the canonical one. Instead, we calculate a z-score for the target pair $\prec v, n \succ$ and each pattern $pt_k \in \mathcal{PS}$:

$$z_k(v,n) = \frac{f(v,n,pt_k) - \overline{f}}{s} \tag{3.5}$$

in which $\overline{f}$ is the mean and s the standard deviation over the sample $\{f(v,n,pt_k) \,|\, pt_k \in \mathcal{PS}\}$:

$$\overline{f} = \frac{\sum_{pt_k \in \mathcal{PS}} f(v,n,pt_k)}{|\mathcal{PS}|}$$

$$s = \frac{\sum_{pt_k \in \mathcal{PS}} (f(v,n,pt_k) - \overline{f})^2}{|\mathcal{PS}|}$$

The statistic $z_k(v,n)$ indicates how far and in which direction the frequency of occurrence of the target pair $\prec v, n \succ$ in a particular pattern pt_k deviates from the sample's mean, expressed in units of the sample's standard deviation. To decide whether pt_k is a canonical pattern for the target pair, we check whether $z_k(v,n) > T_z$, where T_z is a threshold. For evaluation, we set T_z to 1, based on the distribution of z and through examining the development data.

We evaluate the appropriateness of this approach in determining the Cform(s) of idiomatic verb–noun pairs by verifying its predicted forms against ODCIE and CCID. Specifically, for each of the 100 idiomatic pairs in TEST_{all}, we calculate the precision and recall of its predicted Cforms (those whose z-scores are above T_z), compared to the Cforms listed in the two dictionaries. The average precision across the 100 test pairs is 81.2%, and the average recall is 88% (with 68 of the pairs having 100% precision and 100% recall). Moreover, we find that for the overwhelming majority of the pairs, 86%, the predicted Cform with the highest z-score appears in the dictionary entry of the pair. Thus, our method of detecting Cforms performs quite well. The individual precision and recall values for the test pairs are given in Appendix F.

3.6 Related Work

The significance of the role idioms play in language has long been recognized; however, due to their peculiar behaviour, they have been mostly overlooked by researchers in computational linguistics. Recently, there has been growing awareness of the importance of identifying non-compositional multiword expressions. However, most research on the topic has focused on compound nouns and verb particle constructions. Earlier work on idioms has recognized their importance, but failed to explicitly propose mechanisms for appropriately handling them in a computational system. In this work, we focus on a broadly documented and crosslinguistically frequent class of idioms: those that involve the combination of a verb and the noun in its direct object position (VNICs).

Earlier research on the lexical encoding of idioms mainly relied on the existence of human annotations, especially for detecting which syntactic variations (e.g., passivization) an idiom can undergo (Villavicencio et al., 2004). We propose techniques for the automatic acquisition and encoding of knowledge about the lexicosyntactic behaviour of idiomatic combinations. We put forward a means for automatically discovering the set of syntactic variations that are tolerated by a VNIC and that should be included in its lexical representation. Moreover, we incorporate such information into statistical measures that effectively predict the level of idiomaticity of an expression. In this regard, our work relates to previous studies on determining the compositionality (inverse of idiomaticity) of multiword expressions other than idioms.

Most previous work on compositionality of MWEs either treat them as collocations (Smadja, 1993), or examine the distributional similarity between the expression and its constituents (McCarthy et al., 2003; Baldwin et al., 2003; Bannard et al., 2003). Lin (1999) and Wermter and Hahn (2005) go one step further and look into a linguistic property of non-compositional compounds—their lexical fixedness—to identify them. Venkatapathy and Joshi (2005a) combine aspects of the above-mentioned work, by incorporating lexical fixedness, collocation-based, and distributional similarity measures into a set of features which are used to rank verb+noun combinations according to their compositionality.

Our work differs from such studies in that it carefully examines several linguistic properties of VNICs that distinguish them from literal (compositional) combinations. Moreover, we suggest novel techniques for translating such characteristics into measures that predict the level of idiomaticity of verb+noun combinations. More specifically, we propose statistical measures that quantify the degree of lexical, syntactic, and overall fixedness of such combinations. We demonstrate that these measures can be successfully applied to the task of automatically distinguishing idiomatic combinations (VNICs) from non-idiomatic ones. We also show that our syntactic and overall fixedness measures substantially outperform a widely used measure of association, PMI, even when the latter takes syntactic relations into account.

Like us, Evert et al. (2004) and Ritz and Heid (2006) also note that idiomatic word combinations tend to have strong morphosyntactic preferences, and hence propose methods for determining such preferences. The approaches presented in these studies treat individual morphosyntactic markers (e.g., the number of the noun in a verb+noun combination) as independent features. They rely mainly on the relative frequency of each possible value for a feature (e.g., plural for number) as an indicator of a preference for that value. If the relative frequency of a particular value of a feature for a given word combination (or the lower bound of the confidence interval, in the case of Evert et al.'s approach) is higher than a certain threshold, then the expression is said to have a preference for that value. These studies both recognize that morphosyntactic preferences can be employed as clues to the identification of idiomatic combinations; however, none proposes a systematic approach for such a task. Moreover, only subjective evaluations of the proposed methods are presented.

Others have also drawn on the notion of syntactic fixedness for the detection of idioms and other MWEs. Widdows and Dorow (2005), for example, look into the fixedness of a highly constrained type of idiom, i.e., those of the form "X **conj** X" where X is a noun or an adjective, and **conj** is a conjunction such as *and*, *or*, *but*. Smadja (1993) also notes the importance of syntactic fixedness in identifying strongly associated multiword sequences, including collocations and idioms. Nonetheless, in both these studies, the notion of syntactic fixedness is limited to

the relative position of words within the sequence. Such a general notion of fixedness has the drawback that it treats all significant combinations uniformly. A uniform treatment, although may work reasonably well for a lexicographic tool (that involves further human intervention), is not appropriate for many other NLP applications (as also noted by Smadja et al. 1996). Our syntactic fixedness measure looks into a more general set of patterns associated with a more coherent, though large, class of idiomatic expressions.

There is also work on the more difficult task of distinguishing literal and non-literal *tokens* (particular instances) as opposed to *types*. Birke and Sarkar (2006) propose a semi-supervised algorithm for distinguishing between literal and non-literal usages of verbs in context (i.e., token-based classification). Their algorithm uses seed sets of literal and non-literal usages that are automatically extracted from online resources such as WordNet. The similarity between the context of a target token and that of each seed set determines the class of the token. The approach is general in that it uses a slightly modified version of an existing word sense disambiguation algorithm. This is both an advantage and a drawback: the algorithm can be easily extended to other parts of speech and other languages; however, such a general method ignores the specific properties of non-literal (metaphorical and/or idiomatic) language. It remains to be tested whether our measures can be used as linguistically-informed priors for the task of identifying multiword tokens in context.

There is also work that uses evidence from another language to identify MWEs. Melamed (1997a), for example, assumes that non-compositional compounds (NCCs) are usually not translated word-for-word to another language. He thus proposes to discover NCCs by maximizing the information-theoretic predictive value of a translation model between two languages. The sample extracted NCCs reveal an important drawback of the proposed method: it relies on a translation model only, without taking into account any prior linguistic knowledge about possible NCCs within a language. Nonetheless, such a technique is capable of identifying many NCCs that are relevant for a translation task.

Another work that uses information from a second language is that of Villada Moirón and

Tiedemann (2006). They propose measures for distinguishing idiomatic expressions from literal ones (in Dutch), by examining their automatically generated translations into a second language, such as English or Spanish. Their approach is based on the assumptions that idiomatic expressions tend to have less predictable and less compositional meanings, compared to the literal ones. The meaning unpredictability of an expression is measured as the diversity in the translations for the expression, estimated using an entropy-based measure proposed by Melamed (1997b). The non-compositionality of an expression is measured as the overlap between the meaning of an expression (i.e., its translations) and those of its component words. Such approaches have the advantage of being general, hence applicable to different domains and languages. Our measures are more specific, but capable of acquiring more detailed knowledge about a class of MWEs.

3.7 Summary of Contributions

In this chapter, we have proposed statistical, corpus-based measures that incorporate some of the well-known linguistic properties of VNICs in order to determine their degree of idiomaticity. More specifically, our measures quantify the degree of lexical, syntactic, and overall fixedness of a given verb+noun combination. Our contributions are two-fold: we demonstrate that statistical methods both benefit from and add to the existing linguistic knowledge about idioms. On the one hand, generalizations and predictions provided by linguistic theories can be incorporated into statistical, corpus-based measures. On the other hand, such predictions can be automatically and empirically tested using corpus data: information drawn from corpora can be used to enrich the linguistic knowledge that has mainly emerged through introspection.

An important aspect of this work is in how it has interpreted the underlying linguistic theories to be appropriately incorporated into statistical methods. In the development of our measures, we do not take the linguistic predictions as definitive; rather, we allow for some degree of flexibility by looking at the overall probabilistic evidence gathered over large samples

of text. For example, in the development of our syntactic fixedness measure for VNICs, we do not predefine their expected behaviour with respect to their appearance in certain syntactic constructions. Instead, we look for a significant difference between the behaviour of a VNIC and that of a "typical" verb+noun combination—i.e., we use a more relaxed prediction than the one usually taken by the linguistic theories.

Chapter 4

Idioms, LVCs, and Compositional Combinations

In previous chapters, we looked at different parts of the figurativeness continuum, as depicted in Figure 1.1, page 9, repeated here for ease of reference as Figure 4.1. In Chapter 2, we examined the linguistic properties of light verb constructions (LVCs), and proposed methods for placing a candidate LVC on a continuum of less to more figurative uses of the light verb. In Chapter 3, we focused on idiomatic combinations (VNICs) and developed methods for separating them from literal combinations. Here, we look at the full range of figurativeness. We examine various properties of figurative language, and propose techniques for separating LVCs and VNICs from each other, and from similar-on-the-surface literal and abstract combinations.

So far, we have mainly looked at lexicosyntactic fixedness to determine the degree of figurativenss (metaphoricity and/or idiomaticity) of verb+noun combinations. Here, we take a new approach that also deals with other properties of figurative language, such as conventionalization and non-compositionality. This new approach combines evidence from these different sources to determine the extent to which a given verb+noun combination is figurative, thus determining where on the continuum it resides. We thus first define the four classes of verb+noun combinations in Section 4.1. We then expound on the various properties pertaining to figura-

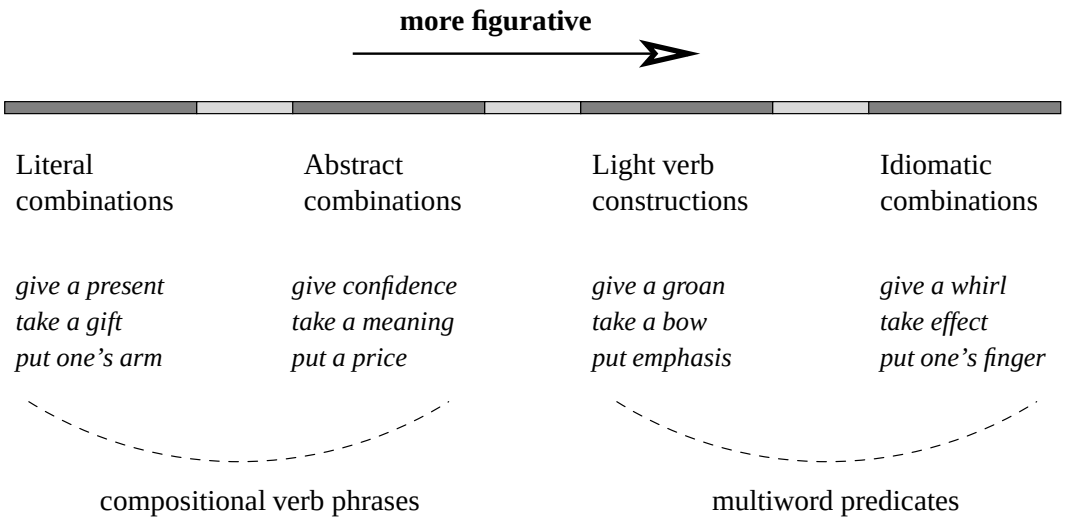


Figure 4.1: Classes of verb+noun combinations on the figurativeness continuum.

tiveness in Section 4.2. Section 4.3 presents our proposed approach for classifying verb+noun combinations. The experimental setup and results are explained in Section 4.4 and Section 4.5, respectively. Section 4.6 provides a survey of related studies, and Section 4.7 concludes the chapter by summarizing the most important contributions of this work.

4.1 The Four Classes of Verb+Noun Combinations

In previous chapters, we focused on the automatic acquisition of knowledge about light verb constructions and verb+noun idiomatic combinations (VNICs). In Chapter 2, we presented a statistical measure that uses evidence from the syntactic behaviour of LVCs (as provided in the linguistics literature) to position combinations of a light verb and a noun on a continuum from less to more figurative. In Chapter 3, we provided measures that incorporate linguistic knowledge about the lexicosyntactic behaviour of VNICs to determine their degree of idiomaticity. In both cases, we evaluated our measures against a standard measure widely used for extracting collocations, and showed that it is both necessary and beneficial to treat LVCs and VNICs as more than collocations. Nevertheless, each piece of work focuses on one of these two classes,

aiming at drawing lines between combinations that belong to the class (i.e., are metaphorical or idiomatic complex lexical units) and those that do not (i.e., are compositional verb phrases with literal semantics). Here, we would like to also distinguish LVCs and VNICs from each other.

LVCs and VNICs share some properties: both have undergone a process of idiomatization,[1] and as a result exhibit some degree of syntactic fixedness. The two classes also have distinct properties that differentiate them as two linguistic constructions. For example, they may differ with respect to the degree and type of fixedness they exhibit. Recall from previous chapters that the nature and the degree of figurativeness is different in VNICs and LVCs. LVCs are semi-compositional since the noun constituent determines the primary meaning of the unit. VNICs are largely non-compositional since the meaning of the expression often diverges substantially from the compositional combination of the meanings of the constituents. In addition, unlike VNICs which are lexically fixed for the most part, LVCs are (semi-)productive—that is semantically similar LVCs can be formed from combining a particular light verb with semantically similar nouns (see, e.g., Fazly et al., 2005).

So far, we have only talked about the more idiomatic combinations, such as LVCs and VNICs, and the fully literal phrases. There are also verb+noun combinations in which the verb is used in an abstract sense, which is a metaphorical extension of its literal semantics. These do not belong to the classes of LVCs or VNICs, nor are they literal verb phrases. They are abstract combinations, such as *cut taxes* and *give confidence*. Abstract combinations often are collocational since they appear with greater frequency than syntactically and semantically similar combinations. For example, whereas *cut taxes* is a completely acceptable combination, neither ?*slash taxes* nor ?*rip taxes* are. Similarly, although *grant confidence* is an acceptable combination, *give confidence* is used much more frequently.[2] Like collocations, the syntactic

[1]Idiomatization is the name often used to refer to the process through which a lexical unit receives a non-compositional semantic interpretation. However, it does not imply that the interpretation is fully idiomatic. In the case of LVCs—which are known to have gone through this process—the interpretation is often metaphorical (see Brinton and Akimoto, 1999, for more details).

[2]Searching on Google, we found a much higher frequency for the latter expression.

behaviour of abstract combinations resembles that of literal verb phrases.

In this chapter, we provide a framework for identifying members of each of these coherent, linguistically plausible classes of verb+noun combinations. The classes are: (i) literal phrases, (ii) abstract combinations, (iii) light verb constructions, and (iv) idiomatic combinations. Here is a brief description of each class, along with some examples. Each class is also given an abbreviated name, shown in boldface. For a more comprehensive explanation of the classes and further examples see the annotation guide in Appendix G.

- Literal combinations (**lit**): We define a literal combination to be one in which the verb contributes its basic meaning that involves a physical action, as in *make a cake*, *find a pen*, and *cut the bread*. We are well aware that this might be a rather conservative definition of the term *literal*. Here, we adopt this view for the sake of clarity and consistency.

- Abstract combinations (**abs**): We define a verb+noun collocation to be one in which the verb contributes a metaphorical extension of its basic literal semantics, as in *find happiness*, and *bring awareness*.

- Light verb constructions (**lvc**): LVCs, such as *make a suggestion* and *take a walk*, are combinations where the noun is the main source of semantic predication. The verb contributes a highly abstract or very little meaning.

- Idiomatic combinations (**idm**): In idiomatic combinations or VNICs, the holistic meaning of the combination diverges substantially from the compositional combination of the individual meanings of the constituents. Examples are *make a killing* (“to earn a lot of money very easily”), and *cut the mustard* (“to succeed”).

Specifically, we bring together ideas and observations from our previous work on metaphoricity and idiomaticity (as presented in the previous chapters), to determine, for a given verb+noun combination, which of the above four classes it belongs to. Such distinctions are important because, as explained previously, members of each class have distinct syntactic and semantic

properties that call for specific treatments within a computational system. For example, although LVCs and VNICs are multiword lexical units, abstract and literal combinations are not. In addition, abstract combinations often are collocational with distinct semantic properties, e.g., they are not fully compositional since the verb contributes a special meaning that differs from its literal semantics.

4.2 Properties of Figurative Language

This section looks into some of the important and widely-recognized characteristics of figurative language, taken from linguistic and lexicographical studies on idioms and other so-called fixed expressions (see Bauer, 1983; Moon, 1998; Brinton and Akimoto, 1999; Cowie, 1992, among others).

Institutionalization is the process through which a combination of words becomes recognized and accepted as a semantic unit (e.g., a collocation, or a lexical unit). Institutionalization is also known as the process through which a particular instantiation of a concept is favored relative to the others. For example, *strong tea* is highly favored over ?*powerful tea*. Institutionalization is, in principle, a necessary but not sufficient condition for a word combination to be considered an MWU. Collocations, for example, are institutionalized, but they are not lexical units.

Lexicosyntactic fixedness refers to some degree of lexical and syntactic defectiveness in a (multiword) lexical unit, appearing as a result of the expression having gone through an idiomatization process. Fixedness is complex, since it can be influenced by factors other than figurativeness (see Chapter 3 for details); it also varies from one lexical unit to another. Lexicosyntactic fixedness is thus neither a necessary nor a sufficient condition for a combination to be classifiable as an MWU.

Non-compositionality as a semantic criterion refers to the situation where the holistic meaning

of a word combination deviates from the meaning emerging from word-by-word interpretation of it. Non-compositionality is a problematic notion on which there has been much disagreement. There are strong arguments in the linguistics literature against the analysis of idioms (and other MWUs) as fully non-compositional units (Gibbs and Nayak, 1989; Gibbs, 1995; Moon, 1998). A rather non-problematic interpretation of this property is to see it as indicating that the constituents of an MWU have special meanings within the context of the unit that are rarely found outside this context. Non-compositionality is also neither a necessary nor a sufficient condition for being an MWU.

None of the above-mentioned properties are sufficient criteria by themselves for determining the degree of figurativeness of a given verb+noun combination. Moreover, like figurativeness itself, each property is also a matter of degree. Nonetheless, if we look at evidence from all these different sources, we expect members of the same class to be reasonably similar, and members of different classes to be notably different. In other words, we hypothesize that combining the different criteria would give us better predictability power for identifying members of each class. The next section elaborates on suggested techniques for quantifying each of these properties of figurative language, that will then be used in a system for classifying verb+noun combinations.

4.3 Automatic Classification of Verb+Noun Combinations

We use the connection between figurativeness and each of the above-mentioned properties (institutionalization, lexicosyntactic fixedness, and non-compositionality) on the one hand, and the relation between the degree of figurativeness and each class (literal, abstract, LVC, or VNIC) on the other, to determine for a given verb+noun combination which class it belongs to. The rest of this section describes statistical, corpus-based measures that are used to quantify the different properties explained in Section 4.2, as well as some other properties that are expected to be relevant to the classification task. At the same time, we draw the connection be-

tween these properties and the membership in any of the above predefined classes of verb+noun combinations.

4.3.1 Measuring Degree of Institutionalization

Like figurativeness, institutionalization is also a matter of degree: corpus-based approaches often assess it by the frequency with which a word combination occurs. Nonetheless, raw frequencies drawn from a corpus are not fully reliable on their own; hence association measures such as pointwise mutual information (PMI) are used in many NLP applications instead. Association measures have been widely and successfully used in the computational linguistics community to extract highly-associated word combinations, including collocations. Moreover, our results in Chapters 2 and 3 confirm the usefulness of such measures in identifying LVCs and VNICs. We expect frequency of a verb+noun combination, together with the strength of association between the two constituents (measured by PMI) to mostly help in separating literal phrases from members of the other classes.

4.3.2 Measuring Degree of Fixedness

In Chapter 3, we have developed measures of lexical, syntactic, and overall fixedness for a verb+noun combination, i.e., $\text{Fixedness}_{\text{lex}}$, $\text{Fixedness}_{\text{syn}}$, and $\text{Fixedness}_{\text{overall}}$. These measures have proven useful in separating VNICs from literal phrases. Our results in Section 3.4.2 suggest that the fixedness measures are also useful in determining the degree of idiomaticity of given verb+noun combinations. We thus expect them to be effective in separating VNICs from other classes, such as LVCs and abstract combinations. VNICs are known to be both lexically and syntactically fixed to a large extent. LVCs, on the other hand, are syntactically fixed, but lexically more flexible (productive), compared to VNICs. Here, we use these fixedness measures, with the same parameter settings we found in Section 3.3.3.

VNICs and LVCs also exhibit other types of fixedness, such as fixedness with respect to

adjectival modification. In fact, one of the main motivations for the use of LVCs in place of their corresponding single-word verbs is the flexibility and ease of modification in LVCs (see Chapter 2 for more details). LVCs are thus expected to often appear with an adjectival modifier, as in *take a relaxing walk*, *give a loud groan*, and *make a great offer*. We thus develop a new measure, $\text{Fixedness}_{\text{mod}}$, that quantifies the degree of fixedness of a given combination with respect to modification. This new measure is very similar to the syntactic fixedness measure described in Section 3.2.2, and is estimated as in Eqn. (4.1):

$$\text{Fixedness}_{\text{mod}}(v, n) \doteq D(P(mod|v,n)\,||\,P(mod)) \tag{4.1}$$

in which $P(mod|v,n)$ is the posterior probability distribution of modification, estimated as in Eqn. (4.2) below; and $P(mod)$ is the prior probability distribution of modification, over all verb–noun pairs, and is estimated as in Eqn. (4.3) below. The random variable *mod* has two possible values: *true* if the pair $\prec v,n \succ$ appears with an adjectival modifier, and *false* otherwise.

$$\begin{aligned} P(mod|v,n) &= \frac{P(v,n,mod)}{P(v,n)} \\ &= \frac{f(v,n,mod)}{\sum_{mod_t \in \{true,false\}} f(v,n,mod_t)} \end{aligned} \tag{4.2}$$

$$P(mod) = \frac{\sum_{v_i \in \mathcal{V}} \sum_{n_j \in \mathcal{N}} f(v_i, n_j, mod)}{\sum_{v_i \in \mathcal{V}} \sum_{n_j \in \mathcal{N}} \sum_{mod_t \in \{true,false\}} f(v_i, n_j, mod_t)} \tag{4.3}$$

$\text{Fixedness}_{\text{mod}}(v, n)$ determines the extent to which $\prec v,n \succ$ is fixed with respect to modification. However, it does not tell which pattern of modification (i.e., *true* or *false*) it prefers most. We thus augment this measure with another one that captures the latter property. This measure is defined as the odds of modification, and is estimated as:

$$\text{Odds}_{\text{mod}}(v,n) \doteq \frac{P(mod = true \mid v,n)}{P(mod = false \mid v,n)} \tag{4.4}$$

Recall from previous chapters that VNICs and LVCs are known to have strong preferences for the syntactic patterns they appear in—i.e., they are largely fixed with respect to their dominant pattern of occurrence. The noun constituent of an LVC is known to be an indefinite, non-referential predicative nominal. Hence, LVCs preferably appear in certain syntactic patterns in which the noun is introduced by an indefinite determiner, e.g., *give a groan*. In contrast, many VNICs tend to prefer syntactic patterns where the noun is introduced by a definite determiner, as in *shoot the breeze*, and *keep one's cool*. In Chapter 3, we introduced a set of patterns, $\mathcal{PS}$ (see Table 3.1), as well as a technique for determining the canonical forms—i.e., dominant patterns—for a given VNIC (see Section 3.5). Here, we use only the most dominant pattern of a given verb+noun combination, determined as in Eqn. (4.5), and take it as another clue to the overall fixedness of the target combination:

$$\text{Pattern}_{\text{dom}}(v,n) \doteq \underset{pt_k \in \mathcal{PS}}{\text{argmax}}\ f(v,n,pt_k) \tag{4.5}$$

4.3.3 Measuring Degree of (Non-)Compositionality

Recall from Section 4.2 that we take non-compositionality of an expression to mean that its constituents have special meanings within the context of the expression. Given this view on non-compositionality, its closest computational approximation is to compare the "context" of an expression with those of its constituents. The more "similar" the context of the expression to those of the constituents, the more compositional the expression.

Often, the context of a target word (or word sequence) is defined as its cooccurring words. These can be any words appearing within a fixed distance of the target, or words with specific

syntactic relations to the target. The context of the target is represented as a vector of words and their frequency of cooccurrence (or their strength of association) with the target. A measure is then required to quantify the semantic "similarity" between any pairs of context vectors. One group of such measures are measures of distributional similarity, such as cosine or Jensen-Shannon divergence (see Mohammad and Hirst, 2005, for a complete survey on distributional similarity measures).

Let t be the target verb+noun combination whose degree of compositionality we want to measure, and let v and n be the two constituents of the target combination. We define the context of a word or word combination to be a vector representation of the cooccurrence frequency of words cooccurring with it within a fixed distance. We use $\vec{t}$, $\vec{v}$, and $\vec{n}$ to refer to the context vectors of the target and its constituents, respectively. The distributional similarity between the target combination and each of its constituents is then measured, using the cosine of the angle between the two vectors. Eqn. (4.6) gives the formula used to estimate the similarity between t and the verb constituent v:

$$\begin{aligned} \mathrm{Sim}_{\mathrm{dist}}(t, v) &\doteq \mathrm{cosine}(\vec{t}, \vec{v}) \\ &\doteq \frac{\vec{t}.\vec{v}}{|\vec{t}| \times |\vec{v}|} \end{aligned} \tag{4.6}$$

$\mathrm{Sim}_{\mathrm{dist}}(t, n)$ is estimated using a similar formula.[3]

Recall from Chapter 2 that the noun constituent of an LVC is a predicative noun, and in most cases either morphologically or etymologically related to a verb. Because of this, an LVC can be roughly paraphrased by the related verb, e.g., *to make a decision* nearly means *to decide*, and *to give a groan* almost means *to groan*. For each verb+noun combination, we thus

[3]Context vectors and similarity values are calculated using CooccurrenceMatrix and DistributionalDistance packages, generously provided to us by Saif Mohammad. We choose cosine, since it proved to be the best measure in the initial experiments we performed on the development and test data sets used in Chapter 3. Based on these initial experiments, we take the context of each word (or word combination) to be a vector of nouns cooccurring with it within a window of ± 5 words.

automatically extract, from WordNet, the verb morphologically related to the noun constituent, *rv*. We then estimate the similarity between *t* and *rv* using a similar formula to Eqn. (4.6) above, and use it as another clue to the degree of compositionality of the target combination. $\text{Sim}_{\text{dist}}(t, rv)$ is expected to mainly help distinguish LVCs from members of the other classes.

4.3.4 Other Relevant Properties

We foresee that the semantic properties of the verb and the noun constituents of a given verb+noun combination are relevant to the classification task. We thus add two simple features that are intended to (indirectly) capture these properties:

- The verb constituent itself, *v*. Basic verbs are known to have preferences for appearing in items from a particular class. For example, *give*, *take* and *make* are recognized by linguists as verbs commonly participating in light verb constructions.

- The semantic class of the noun according to WordNet 2.1, $\text{SemClass}(n)$. We only look at the children of ENTITY in the noun hierarchy for determining the semantic class of a noun. (And for each noun, we only consider its first—predominant—sense to extract this information.) The classes are PHYSICAL ENTITY, ABSTRACT ENTITY, and a third concept that is not relevant to our task. Based on the description of these semantic classes, we decided to group nouns under PROCESS (a child of PHYSICAL ENTITY) as an ABSTRACT noun.

4.4 Experimental Setup

4.4.1 Corpus and Experimental Expressions

As in Chapter 3, we use the automatically parsed British National Corpus to extract our experimental verb–noun pairs, along with further information on their modification status and the syntactic patterns they appear in (see Section 3.3.1 for further detail on the extraction process).

Also, as in Chapter 3, we select our experimental expressions from verb–noun pairs that involve a member of a predefined list of basic (transitive) verbs. Here, however, we use only a subset of the 28 verbs used therein. We first rank the 28 verbs according to their type frequency in the BNC—i.e., the number of different verb–noun pairs containing each of these verbs. We then select 12 verbs (out of the original set of 28 verbs) ranked at the top.[4] Here is the final list of the 12 selected verbs in alphabetical order:

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

From the corpus, we extract all the verb–noun pairs with minimum frequency of 25 that contain any of the above-listed basic verbs. To ensure that the final set of expressions contains pairs from all four classes, we pseudo-randomly select our final set of expressions. We consult the two idioms dictionaries, ODCIE and CCID, and include expressions that appear both in the BNC (with frequency greater than 25), and in any of the two dictionaries. We also select pairs in which the noun has a morphologically related verb according to WordNet, as well as pairs in which the noun is not morphologically related to any verb.

This selection process resulted in a set of 632 pairs, reduced to 563 after the pairs were annotated by our primary annotator. More detail on the annotation process is given in Section 4.4.3. Out of 563 annotated pairs, 148 are literal, 196 are abstract, 102 are LVCs, and 117 are idiomatic. We randomly choose 102 pairs from each class to include in the final experimental set. We then pseudo-randomly divide these into training (TRAIN), development (DEV) and test (TEST) data sets. All three sets have equal number of pairs in 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.

We use DEV to perform initial experiments in order to find features most relevant to the classification task. TEST is kept unseen for the final evaluation. Results on the development data reveal that removing $\mathrm{Sim}_{\mathrm{dist}}(t, v)$ results in a notable improvement in performance. This

[4]We decided to remove *have* from this list because of its common use as an auxiliary verb, even though it is ranked at the very top.

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 believe that adding such inaccurate information is thus misleading, hence hurting the classification performance. (This is also confirmed by looking at the decision tree classifier build using this feature only.) We thus remove this feature in further experiments on TEST.

4.4.2 Classification Strategy and Features

We adopt a supervised strategy to classify our development and test verb–noun pairs. We use the decision tree induction system C5.0 (http://www.rulequest.com), the successor of C4.5 (Quinlan, 1993), as our machine learning software. We use the measures defined in Section 4.3 as features in the classification task. Table 4.1 presents the full list of features used in classification, along with a brief description of each. Features are grouped together according to the property they capture; each group is also given an abbreviated name, shown in boldface. In our experiments, we try to determine the relevance of each feature group in the classification task. We also look into the effectiveness of each group in identifying members of each class.

4.4.3 Human Judgments

To acquire judgments on the class of each experimental verb–noun pair, we ask four native speakers of English with sufficient linguistic background to annotate them. The annotation task was expected to be time-consuming, hence it was not feasible to ask all the judges to annotate all the expressions. Instead, we asked one judge to be our primary annotator, to whom we refer as PA henceforth. We then asked PA to annotate all the expressions before the other annotators. What comes next is a description of the annotation procedure.

First, PA annotated all the 632 pairs selected as described in Section 4.4.1. PA removed 69 of the pairs that could be potential sources of disagreement for various reasons. For example, if an expression was likely to be British, or to be used in a different way in British (and therefore

Table 4.1: Features used in the classification of verb+noun combinations, grouped according to the property they are intended to capture. (*t* refers to the target verb+noun combination being classified, *v* and *n* refer to its constituents, and *rv* refers to the verb related to *n*.)

Group# (Relevant Property)	Feature	Description
INST (Institutionalization)	$\text{Freq}(v, n)$	total frequency of cooccurrence of *v* and *n*
	$\text{PMI}(v, n)$	strength of association between *v* and *n* cooccurring as verb–object
FIXD (Fixedness)	$\text{Fixedness}_{\text{lex}}(v, n)$	lexical fixedness of $\prec v, n \succ$
	$\text{Fixedness}_{\text{syn}}(v, n)$	syntactic fixedness of $\prec v, n \succ$
	$\text{Fixedness}_{\text{ove}}(v, n)$	overall fixedness of $\prec v, n \succ$
	$\text{Fixedness}_{\text{mod}}(v, n)$	fixedness of $\prec v, n \succ$ w.r.t. to modification
	$\text{Odds}_{\text{mod}}(v, n)$	odds of $\prec v, n \succ$ being modified
	$\text{Pattern}_{\text{dom}}(v, n)$	dominant pattern of $\prec v, n \succ$
COMP (Non-Compositionality)	$\text{Sim}_{\text{dist}}(t, n)$	semantic contribution of *n* to *t*
	$\text{Sim}_{\text{dist}}(t, rv)$	semantic contribution of *rv* to *t*
VERB (Verb)	v	the verb itself
NSEM (Noun Semantics)	$\text{SemClass}(n)$	semantic class of *n* taken from WordNet

in the BNC) than in North American English, it was removed from the list.[5] Some other reasons for removing an expression are: if the expression was likely to be part of a larger phrase (e.g., verb+noun+particle); if PA was not familiar with the expression; if the expression was vulgar.

Next, we divided the remaining 563 pairs into three equal-sized sets, and gave each set to one judge to annotate. The annotators were given a comprehensive guide for the task, the full text of which can be found in Appendix G.

In our final evaluation, we use the annotations of PA as the gold standard. We use the annotations of the other three annotators to estimate the degree of difficulty of the task, i.e., to get inter-annotator agreement. Table 4.2 lists the observed agreement (p_o) as well as the kappa score (κ) between PA and each of the other three annotators; N_i is the number of items

[5]This is because the verb–noun pairs are extracted from the BNC, which is mainly British English, whereas the main dialect of all our annotators is American or Canadian English.

Table 4.2: Overall observed agreement and kappa score between PA and each of the three annotators.

	ANNOTATOR$_1$			ANNOTATOR$_2$			ANNOTATOR$_3$		
	N$_1$	p_o	κ	N$_2$	p_o	κ	N$_3$	p_o	κ
PA	188	79.8 %	.72	188	67 %	.56	187	72.2 %	.62

annotated by ANNOTATOR$_i$. (Inter-annotator agreements on the individual classes are given in Appendix J.) As can be seen, both p_o and κ are reasonably high for all three annotators. This is a confirmation that the classes are coherent and linguistically plausible. The observed agreements (between 67% and 80%) can also be seen as an upper bound on the task.

4.5 Results

We perform experiments where we use each of the five groups of features, separately, to classify DEV and TEST pairs. To further examine the usefulness of each feature group in distinguishing different classes, we perform experiments in which we combine all features, but the feature group under study. We also perform an experiment with all the feature groups combined to see how they perform together. We report the average accuracy over all classes to measure the goodness of each feature group (as well as all groups combined). To evaluate the performance of feature groups in identifying members of the individual classes, we report F-measure, which combines precision (P) and recall (R) into a single measure of overall performance, as in:

$$F = \frac{2PR}{(P+R)}$$

where precision and recall are equally weighted. Results presented here are on TEST set; those on DEV set have similar trends. We first look at the overall performance of classification in Section 4.5.1. Section 4.5.2 presents the results of classification for the individual classes.

4.5.1 Overall Classification Performance

Table 4.3 presents the results of classification for the individual feature groups, as well as for all groups combined. Note that in all cases, the baseline accuracy (random baseline) is 25% since we have four equal-sized classes in TEST. As can be seen, institutionalization features (those in INST) 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 figurativeness. Among the individual groups, fixedness features (those in in FIXD) achieve the highest accuracy, 50%, with a relative error reduction of 33%. Once again, these results confirm our hypothesis that fixedness is a salient characteristic of figurative language, and that it could be used effectively for the automatic acquisition of MWUs. Compositionality features (those in COMP) achieve reasonably good accuracy, around 40%, though still notably lower than the accuracy of fixedness features. This is especially interesting, because much previous research has focused solely on the non-compositionality of MWUs to identify them (e.g., McCarthy et al., 2003; Baldwin et al., 2003; Bannard et al., 2003; Katz and Giesbrecht, 2006). 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 (those in VERB and NSEM) overall perform comparably to the compositionality features. A closer look at the performance of these features on the individual classes reveals that, unlike compositionality features, they are only good at identifying items from certain classes (see below for further discussion on this). As hypothesized, we achieve the highest performance—an accuracy of 58% and a relative error reduction of 44%—when we combine all the feature groups.

Table 4.4 displays the accuracy of classification, when we use all the feature groups except one. These results are more or less consistent with those in Table 4.3 above, except some differences which we discuss below. Removing features relevant to fixedness results in a drastic decrease in performance (10.7%), while the removal of the institutionalization or compositionality features cause much smaller drops in performance (4.7% and 2.3%, respectively). Here

Table 4.3: Accuracy (%*Acc*) and relative error reduction (%*RER*) for the individual feature groups, as well as all features combined, over TEST pairs.

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

Table 4.4: Accuracy (%*Acc*) and relative error reduction (%*RER*) over TEST pairs, removing one feature group at a time.

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

again, we can see that features related to the semantics of the verb and the noun are salient features. Removing any of these results in a substantial decrease in performance—9.5% and 11.9%, respectively—which is comparable to the decrease resulting from removing the fixedness features. This is an interesting observation, since VERB and NSEM feature, on their own, do not perform nearly as well as FIXD features. Later, when we look at the performance on the individual classes, we see that VERB and NSEM features are especially relevant for identifying some classes, but do very poorly on the other classes. This is why their removal hurts the performance so much, despite the fact that they do not do well on their own.

4.5.2 Performance on Individual Classes

We now look at the performance of the feature groups, separately, and combined, on the individual classes. The *F*-measures of classification, for each combination of class and feature group, are given in Table 4.5. (The observed agreement and kappa scores among the annotators are given in Appendix J for comparison purposes.) The two highest *F*-measures for each class are shown in boldface. These results show that the combination of all feature groups yields the

best or the second-best performance on all four classes. (In fact, performance of ALL features is notably smaller than the best performance achieved by a single feature group for one class only.)

Looking at the performance of ALL features, we can see that we get reasonably high F-measure for all classes, except for **abs**. This is expected because members of this class share properties with all other classes. Their syntactic behaviour is likely to be similar to those in **lit**, while they might show some degree of lexical fixedness like expressions in **idm**. In terms of compositionality, this class can have a range of members, from fully compositional to somewhat conventionalized (and hence to some extent non-compositional). This is also reflected in the extremely poor performance of the compositionality features on this class. Fixedness features also yield their worst performance on this class.

According to the F-measures, the most relevant feature group for identifying members of **lit** and **abs** classes is NSEM, that is the semantic class of the noun. This is not surprising since this feature basically determines whether the noun is a PHYSICAL ENTITY or an ABSTRACT ENTITY. By definition, most items in **lit** class are expected to involve a physical noun, whereas those in **abs** are expected to have an abstract noun. Nonetheless, this is an interesting result, given that this is a very simple feature. It is important, however, to note that this feature as-is is completely irrelevant in the recognition of items from the other two classes, **lvc** and **idm**. In the future, we need to expand this feature to also include semantic classes other than PHYSICAL ENTITY and ABSTRACT ENTITY.

Interestingly, the most relevant feature group for **lvc** and **idm** is FIXD—i.e., features relevant to fixedness. Moreover, for the class **idm**, the performance of this feature group is notably higher than that of all feature groups combined. The lower performance of the combined feature set on **idm** is particularly due to the low performance of other feature groups on this class. Fixedness features thus once again prove to be salient in the recognition of idioms and other MWUs such as LVCs.

Institutionalization features, on their own, do not yield particularly good results for the in-

Table 4.5: *F*-measures on TEST pairs, for individual feature groups and all features combined.

	Only the features in group					
	INST	FIXD	COMP	VERB	NSEM	ALL
Class	%*F*	%*F*	%*F*	%*F*	%*F*	%*F*
lit	47.6	41.9	51.2	54.2	**57.1**	**60**
abs	40	31.6	16.7	26.6	**48.6**	**45.7**
lvc	21	**58.4**	47	55.1	0	**68.2**
idm	33.3	**66.7**	42.1	0	0	**56.4**

dividual classes. Their poor performance for **lvc** and **idm** shows that these MWUs may not necessarily appear with significantly high frequency of occurrence in a given corpus. Compositionality features have reasonable performance (*F*-measure ranging from 42% to 51%) on all classes but **abs**. Nonetheless, these feature groups, although not sufficient by themselves, help boost the classification performance when combined with other feature groups.[6]

4.6 Related Work

There are a number of studies on the classification of multiword verbs into predefined semantic groups. One group of work focuses on classifying verb-particle constructions (VPCs), mostly concerned with separating VPCs from compositional verb–preposition combinations. Baldwin and Villavicencio (2002) mainly focus on the automatic extraction of VPCs based on surface syntactic cues available in a PoS-tagged or chunked corpus. Instead of making a binary decision as to whether a given verb+particle combination is a VPC or a compositional combination, McCarthy et al. (2003) determine a continuum of compositionality in these constructions. They assess the compositionality of a VPC by combining evidence from the distributional similarity between the VPC and each of its parts, the verb and the particle. They thus do not distinguish

[6]In our experiments on DEV, we found that removing the two groups (INST and COMP) simultaneously drastically hurts the performance.

the contribution of the individual components, rather determine the degree of compositionality of each expression as a single unit. Bannard (2005), on the other hand, attempts to separately determine the contribution of the verb and the particle to the semantics of the VPC they compose together. Like McCarthy et al., Bannard also examines the distributional similarity between the VPC and each of the constituents for the purpose. Both studies compare the compositionality ratings of their proposed measures with those of humans. The results, although promising, show that there is still much space for improvement. Cook and Stevenson (2006) move a step further and look into deeper linguistic properties of a subclass of VPCs in order to classify them according to a finer-grained semantic contribution of the particle.

Another group of studies that are more relevant to ours attempt to classify LVCs or verb–noun combinations in general, according to some predetermined semantic criteria. The supervised classifier of Wanner (2004) divides verb–noun combinations into semantic groups, each corresponding to a particular semantic relation between the two constituents. His choice of semantic classes, although linguistically justified, is perhaps too fine-grained and too vague to be directly useful for NLP applications. Uchiyama et al. (2005) put forward a statistical approach for classifying Japanese LVCs (of the form verb–verb). Their proposed classes are very broad, identified based on high-level semantic contributions of the light verb: spatial, aspectual, or adverbial. It is not clear to us whether and how these classes can be directly useful for NLP applications.

Krenn and Evert (2001) also attempt to distinguish LVCs (which they refer to as support verb constructions) and idioms (which they call figurative expressions), both from each other and from literal phrases. Nonetheless, they treat these MWUs as collocations, 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 also look at several aspects of figurativeness other than institutionalization (that we quantify using an association measure).

The work most similar to ours is that of Venkatapathy and Joshi (2005a). Venkatapathy

and Joshi propose a minimally-supervised classification schema that incorporates a variety of features to group verb–noun combinations according to their level of compositionality. They use some linguistically-motivated features, as well as collocation-based and distributional similarity measures for the task. Their work has the advantage of not using a lot of training data, which is required to be manually labelled for such applications. However, their classes are defined on the basis of compositionality only. Our proposed work is different in that it brings in a new group of features—i.e., the fixedness features—which prove to be one of the most effective set of features in determining degree of figurativeness. Moreover, our predefined set of classes include items from particular linguistic constructions whose distinct properties are well-recognized. Their distinct properties also suggest that it is important to distinguish these classes for many NLP applications, such as machine translation and natural language generation.

4.7 Summary of Contributions

In this chapter, we have provided an analysis of the various characteristics of figurative language, as discussed in linguistic and lexicographical studies. We also elaborate on the relationship between these properties and four coherent, linguistically plausible classes of verb+noun combinations, falling on a continuum from literal to metaphorical to idiomatic. On the basis of such analysis, we have developed statistical, corpus-based measures to quantify each of the properties pertaining to figurativeness in a systematic way.

The suggested measures are then used as features in a classification task, to determine the effectiveness and relevance of each property for separating items from the four predefined classes. Our goal in this study has been mainly to provide a clear analysis of the interaction between the feature groups and the individual classes. Our results are intended to provide guidelines concerning which features are most or least relevant for identifying a particular class of verb+noun combinations.

As mentioned previously, each of the feature groups we use to classify verb+noun combinations is intended to relate to one property of figurative language. Here, we also show that combining evidence from all the different sources is beneficial. The best performance overall, and on most individual classes, belong to the classifier that uses all the features.

Chapter 5

Summary and Outlook

In the research presented in this thesis, we have shown that it is possible to acquire more accurate knowledge about multiword expressions by looking into their deeper linguistic properties. Our results show that on one hand, linguistic knowledge appropriately incorporated into statistical methods greatly benefits them. On the other hand, the scalability of statistical approaches adds much to the existing linguistic theories: Predictions provided by such theories have mainly emerged through introspection and hence can be enriched by being empirically tested against real corpus data (as also noted by corpus linguists). Our work successfully blends together the flexibility of statistical methods and the predictability power inherent in linguistic theories.

Specifically, we have developed several statistical, corpus-based measures that relate to some of the important properties of figurative language as manifest in the surface behaviour of multiword lexical units. Our focus in this work has been on a particular class of multiword expressions, i.e., verb+noun combinations. We have introduced new measures for quantifying the degree of fixedness of these combinations, and have presented new ways of using some existing techniques for capturing other characteristics of figurative language, such as institutionalization and non-compositionality. Each group of measures thus relates to one property pertaining to figurativeness (fixedness, institutionalization, non-compositionality). We show that by combining evidence from all these different sources, we can successfully distinguish

different types of verb+noun combinations, such as literal phrases, abstract combinations, light verb constructions and idiomatic combinations. Moreover, these statistical measures help with the acquisition of important knowledge about these combinations to be included in their lexical representation in a computational lexicon.

Even though our experiments are limited to comparisons against human judgments of various properties, we believe that the kinds of knowledge our measures acquire and the kinds of distinctions they are capable of making will be useful in many NLP tasks. For example, distinguishing among different classes of verb+noun combinations can be beneficial for machine translation, information extraction, and text summarization systems, to name a few. Members of these verb+noun classes exhibit different syntactic and semantic behaviour, and hence need to be treated differently in such NLP applications. In addition, providing an automatic mechanism for summarizing evidence from lexical, syntactic, and semantic behaviour of figurative expressions can greatly benefit corpus linguistics as well as research on the diachronic aspects of language change, particularly idiomatization and lexicalization processes. We also expect our proposed measures and techniques to be useful as tools for assisting lexicographers. Statistical measures can be used to automatically acquire knowledge about the behaviour of multiword expressions from large bodies of text. Such knowledge could be further used by human experts to draw more general conclusions on the behaviour of these expressions.

5.1 Summary of Contributions

A brief summary of the contributions of this thesis has been presented in Section 1.3 (Chapter 1). Here, we expand it, adding more emphasis on the results from our experimental evaluation.

Detecting a continuum of figurativeness in metaphorical expressions: In Chapter 2, we focused on the more metaphorical part of the figurativeness continuum that mainly involves light verb constructions or LVCs. We proposed statistical measures that examine the degree to

which a verb usage is "similar" to the prototypical LVC, as an inverse indicator of the degree to which the verb retains aspects of its literal semantics. These measures identify a continuum of literal to figurative usages of a verb. We evaluated our measures by comparing them against human judgments on the same property, and showed that they correlate well with the literal–figurative spectrum represented in these judgments. We also showed that the more linguistic knowledge a measure incorporates, the stronger the correlations are with the human judgments. For example, on a set of verb+noun combinations—i.e., those that involve the indefinite determiner *a* and the verb *give*—the measure incorporating most linguistic knowledge about LVCs achieves a correlation score of .77. This is in comparison with a correlation score of .68 for a less-informed measure, and a score of .62 for a measure of association strength with no specific knowledge about LVCs.

Even for humans, determining the degree of figurativeness of the verb constituent of a verb+noun combination is not an easy task. We thus developed a careful strategy for collecting data required for evaluation. Instead of directly asking our human judges to provide ratings of figurativeness, we asked them a set of yes/no questions that indirectly captured the degree to which aspects of the literal meaning of the verb constituent were retained in the meaning of an expression. For each expression, we then translated these answers into numerical ratings reflecting its degree of figurativeness. We get moderate inter-annotator agreement (as measured by linearly weighted kappa) on one set of expressions and reasonably high agreement on the other. We believe that both the annotation procedure and the resulting annotated expressions can be useful for conducting further research on metaphorical verb+noun combinations.

Development of novel techniques for handling idiomatic combinations: In Chapter 3, our focus moved to the more idiomatic end of the figurativeness continuum, i.e., on verb+noun idiomatic combinations or VNICs. We proposed statistical measures that quantify the degree of lexical, syntactic, and overall fixedness of verb+noun pairs. We used scores assigned by these measures to a given verb+noun combination as an indicator of the degree of idiomaticity of the

combination. We evaluated the fixedness measures by using them to: (i) separate VNICs from similar-on-the-surface literal verb phrases; (ii) rank a mixed list of verb+noun combinations according to their degree of idiomaticity.

In both tasks, our measures perform substantially better than a widely used measure of collocation extraction. In the first task, the lexical, syntactic, and overall fixedness measures achieve accuracies of 68%, 71%, and 74%, respectively (compared to 63% for the collocation extraction measure). Moreover, unlike the collocation extraction measure, our measures are not adversely sensitive to frequency. Of more interest is their performance in the second task for which we look at the 3-point interpolated average precision (average of precisions at 3 levels of recall). The average precision for the fixedness measures is very high, ranging from around 75% to around 85% (compared to around 63% for the collocation extraction measure).

We also introduced a method for determining the canonical forms of idioms needed to be included in their lexical representation. We showed that for a set of 100 idiomatic verb–noun pairs, our method could determine the canonical forms with an average precision of 81.2% and an average recall of 88%.

Classification of verb+noun combinations: In Chapter 4, we turned our attention to the whole continuum of figurativeness, identifying linguistically plausible classes of verb+noun combinations on that continuum. These are literal phrases, abstract combinations, LVCs, and VNICs. We looked at several properties of figurative language in addition to fixedness, and put forward a relationship between each such property and a set of statistical measures used to quantify it. We evaluated these measures by using them for classifying a mixed set of verb+noun combinations into the above-mentioned predefined classes. We showed that combining all the measures yields the highest classification performance, an accuracy of over 58% on a task with a chance baseline of 25%—i.e., over 44% reduction in error rate.

To produce a gold standard solution for our evaluation, we needed to collect data annotated by human judges. We thus developed a careful strategy for the task, and asked four expert

judges to annotate a set of experimental items (automatically extracted from the BNC). We get high inter-annotator agreements, measured using both the percentage of observed agreement and the kappa score. The high agreements suggest that the classes are in fact linguistically plausible, and that the annotation procedure is accurate. Given the lack of sufficient annotated data for the evaluation of lexical acquisition techniques for multiword expressions, we believe that both our annotation procedure and the resulting set of annotated items will be useful for the community.

5.2 Future Directions

Our work also has limitations that need to be addressed in the future. For one, our work focuses on a particular class of MWEs, i.e., verb+noun combinations. Also, in developing our fixedness measures, we ignore the fact that like single words, MWEs can also be ambiguous. Because of this limiting assumption, we expect our measures to reflect the properties of the most dominant sense of a target MWE. Moreover, because we look into the deep linguistic properties of verb+noun combinations, our fixedness measures are language-specific to a certain extent. Work on the underlying semantic properties of MWEs is a relatively new area of research in computational linguistics, hence evaluation is a great challenge. Like many others, we have also evaluated our work by comparing the predictions of our statistical measures with those of human judges. A more comprehensive evaluation in the context of real NLP applications is necessary. The following sections expound on techniques for overcoming such limitations, as well as directions for extending the current work.

5.2.1 Short-term Extensions

Extension to other figurative MWEs: Figurative language is widespread in many different forms. In the work presented in this thesis, we have focused on the acquisition of knowledge for a largely overlooked class of figurative multiword expressions, i.e., verb+noun combina-

tions. There are many other classes of expressions with distinct syntactic and semantic properties that require specific attention. A possible future direction of the work is thus to extend the statistical measures to other figurative verb phrase (VP) combinations, such as *give in to (someone/something), keep to (something), take (something) into account, have a bun in the oven, fly in the face of*, and *keep one's lips sealed*.

It is also interesting to determine the effectiveness of similar techniques to those proposed in this thesis in identifying figurative constructions other than VPs. Examples are figurative noun phrase (NP) combinations, such as *a red herring* ("a deliberately misleading object") and *golden handshake* ("a good financial package someone is given when they leave a company"), figurative adjective phrase (AP) combinations, such as *all too brief* ("much briefer than suitable") and *fast asleep* ("deeply asleep"), and complex prepositions, such as *in search of* and *in conformity with*.

Like verb+noun combinations, other figurative MWEs are also known to exhibit lexical and syntactic fixedness to a certain extent. For example, complex prepositions are known to exhibit, to some degree: (i) lexical fixedness, i.e., the substitution of either preposition, (ii) fixedness with respect to modification, i.e., the pre-modification of the noun constituent, and (iii) syntactic fixedness, i.e, change in the number of the noun (singular vs. plural), as well as change in the determiner introducing the noun (Brinton and Akimoto, 1999). Our fixedness measures, although construction-specific to some extent, are sufficiently general to be used for MWEs other than verb+noun combinations. Nonetheless, for each type of construction, a linguistic analysis of the properties is required.

Improving the classification: Our work on the classification of verb+noun combinations demonstrates the usefulness of combining evidence from different sources to determine where on the figurativeness continuum a given combination resides. A possible extension to this work is to use additional sources of information to measure the degree of each property pertaining to figurativeness. One useful source is the translations of verb+noun combinations into another

language, as others have also noted. For example, it is often assumed that non-compositional MWUs are usually not translated word-for-word to another language (Melamed, 1997a). Such information can thus be used as another piece of evidence for determining whether a particular verb+noun combination is an MWU or a literal or abstract combination.

5.2.2 Long-term Goals

MWE token disambiguation: The statistical measures used in this thesis, either individually or combined, measure the degree of figurativeness of a given verb+noun type out of context. More specifically, the way these measures are designed is that they gather evidence from all instances of the target verb+noun combination (tokens) to infer something about the lexical, syntactic, and/or semantic properties of the verb+noun type. Nonetheless, not all instances (tokens) of a particular verb+noun combination have exactly the same properties: *kick the bucket* is often used as an idiom, but in *Joe kicked the red bucket instead of the blue one* it is used as a literal verb phrase. Similarly, *make a face* is ambiguous between an idiom, as in *The little girl made a funny face at her mother*, and a literal verb phrase, as in *She made a face on the snowman using a carrot and two buttons.*. Token disambiguation is thus an important problem for multiword expressions. One possible approach is to use measures suggested in this work to gather some prior information on the behaviour of a multiword type, and augment it with context-specific knowledge for each multiword token (see Lapata and Brew, 2004, for a similar approach on verb class token disambiguation).

MWEs in machine translation and language generation: Because of their peculiar syntactic and semantic behaviour, multiword expressions pose a serious challenge to current translation models. Most existing machine translation systems are based on a statistical alignment between words from the source and the target language (Melamed, 2000; Och et al., 1999). Even phrase-based translation models mostly try to find alignments between contiguous sequences of words across the two languages (Koehn et al., 2003). We believe there is room for

improvement in the existing machine translation systems.

A straightforward approach is to perform a preprocessing of source and target texts to identify potential MWEs. We can then modify an existing phrase alignment algorithm to give priority to the identified MWEs as linguistically meaningful phrases, rather than just looking at contiguous sequences of words (see Venkatapathy and Joshi, 2006, for one such approach). A by-product of such a technique is a bilingual lexicon of MWEs, created by looking at the alignments between MWEs from the two languages. Such a lexicon can be used by a machine translation system or as an aid for human translators (Smadja et al., 1996).

Another interesting but more complex approach is to incorporate the statistical measures for identifying MWEs into the language model used by a translation system. Most current systems use simple *n*-gram language models that look only at the contiguity among words for the generation of natural-sounding sentences in the target language. By also looking at syntactic dependencies we can increase the likelihood of generating linguistically valid MWEs, where possible.

It is also possible to use evidence from the translation model (the word or phrase alignments) to more accurately identify MWE tokens in context. In their work on identifying idioms and LVCs using word alignments, Villada Moirón and Tiedemann (2006) find that alignments produced for the constituents of an MWE are often too diverse. They propose a method, based on the idea of translational entropy proposed by Melamed (1997b), to use this diversity for recognizing MWEs. One potential future project is to investigate the possibility of developing a bootstrapping strategy that uses information about potential MWE *types* to improve a translation model, and at the same time uses information inherent in the translation model to more accurately identify MWE *tokens* in context.

Extending to other languages: Figurative speech is common across all languages. Another continuation of this work is thus to see how well the linguistic and psycholinguistic observations on the syntax-semantics interface for figurative language, as well as the statistical

measures that draw on this relation, extend to languages other than English. Previous studies on Hindi, German, and Dutch figurative VP combinations have produced encouraging results (Venkatapathy and Joshi, 2005b; Krenn and Evert, 2005; Villada Moirón and Tiedemann, 2006). It is nonetheless important to note that some of our measures, such as the syntactic fixedness measure, may not be directly applicable to some languages, including those with relatively free word order. Hence, it might be necessary to look for other surface manifestations of semantic figurativeness for these languages, by analysing their specific linguistic properties.

Comparative studies across languages: So far, we have focused on predefined sets of basic verbs with the hypothesis that they tend to appear in many diverse expressions. An important follow-up on this work is thus to automatically identify such verbs. Simple ways are looking at their frequency of occurrence and/or the number of different types of arguments they combine with. An interesting observation is that basic verbs with similar meanings have been documented to form MWEs in many diverse languages, including those that are genetically unrelated (Butt, 1997), e.g., *make* in English, *faire* ("to do/make") in French, *kardan* ("to do/make") in Persian, and *suru* ("to do/make") in Japanese. One interesting approach to the identification of basic verbs is thus to develop an accurate measure of their productivity in forming MWEs, by looking at evidence from all the above-mentioned sources—that is by looking both at their productivity within a language and across different languages.

Analysis of differences in the choice of the basic verb in figurative combinations across different languages, or even across different dialects of the same language, is also important in the context of machine translation and language generation. For example, one would *set the table* or *take a nap* in America, whereas in Britain one would *lay the table* or *have a nap* (examples taken from Smadja et al., 1996; Wierzbicka, 1982). For such comparative studies, both monolingual and parallel texts are needed. The international corpus of English (ICE)[1] consists of one million words of spoken and written English for a variety of dialects—e.g., Australian, British,

[1] http://www.ucl.ac.uk/english-usage/ice/

Canadian, East African, and Indian—produced to ensure compatibility among the different components. These are considerably small-sized corpora; moreover, monolingual corpora are not available for many of these dialects. It is thus necessary to come up with techniques for grouping figurative expressions in order to look at more reliable evidence for choosing a particular verb over another. One approach is to group expressions according to the semantic class of the item that combines with the verb, as noted by Fazly et al. (2006).

Appendix A

List of abbreviations

CV	Complex Predicate
NLP	Natural Language Processing
MWE	Multiword Expression
MWU	Multiword Lexical Unit
LVC	Light Verb Construction
PMI	Pointwise Mutual Information
VNIC	Verb+Noun Idiomatic Combination
ODCIE	Oxford Dictionary of Current Idiomatic English
CCID	The Collins COBUILD Idioms Dictionary
Acc	Accuracy
RER	Reduction in Error Rate
IAP	Interpolated Average Precision
INST	INSTitutionalization
FIXD	FIXeDness
COMP	COMPositionality
NSEM	Noun SEMantics

Appendix B

On the annotation task from Chapter 2

This appendix contains information on the procedure of acquiring human judgments for the development and test expressions used in experiments of Chapter 2. Table B.1 presents the questions that the judges were asked on expressions with *take*. Tables B.2 and B.3 show how the judges' answers to the questions are translated into numerical ratings. Higher numerical ratings express higher degrees of literalness, hence lower degrees of figurativeness. Expressions for which no numerical rating is listed in the tables are removed from the final set of experimental expressions.

Table B.1: Questions for expressions containing *take*.

	Question	Possible answers
As a result of the *event* expressed by the given *expression*:		
I.	Does "SUBJ **take in** a physical object", or "AP[a] **transfer** a physical object to SUBJ"?	yes, no, maybe, ?
II.	Does "SUBJ **move**"?	yes, no, maybe, ?
III.	Does "AP **transfer** something (non-physical) to SUBJ"?	yes, no, maybe, ?
IV.	Does "SUBJ **take in** or **adopt** something (non-physical)"?	yes, no, maybe, ?

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

Table B.2: Interpretation of answers to the questions for expressions with *take*.

Q(I)	Q(II)	Q(III)	Q(IV)	Rating
yes/maybe	no	no	no	4
yes/maybe	–	yes/maybe	no	3
maybe	–	no	maybe	3
no	–	yes/maybe	no	2
no	–	no/maybe	yes/maybe	1
maybe	–	no	yes	1
no	–	no	no	0
yes/maybe	yes	no	no	0
any combination other than above				-

Table B.3: Interpretation of answers to the questions for expressions with *give*.

Q(I)	Q(II)	Q(III)	Rating
yes	no	no	4
yes/maybe	yes/maybe	no	3
no	yes	no	2
no	no/maybe	yes	1
no	no	no	0
any combination other than above			-

Appendix C

Experimental expressions from Chapter 2

Table C.1: Development expressions with their individual and consensus (average) human ratings, sorted by the latter.

give					*take*				
expression	human ratings				expression	human ratings			
give thought	0	0	0	0	*take a grip*	0	0	0	0
give a hand	2	0	0	0.67	*take a ride*	0	0	0	0
give a smile	1	1	0	0.67	*take a seat*	0	0	0	0
give a squeeze	2	0	0	0.67	*take a shower*	0	0	0	0
give a thrill	2	0	0	0.67	*take a taxi*	0	0	0	0
give cause	1	1	0	0.67	*take a turn*	0	-	0	0
give pause	2	0	0	0.67	*take exercise*	0	0	0	0
give a kick	2	1	0	1	*take leave*	0	0	0	0
give a laugh	1	1	1	1	*take practice*	0	-	0	0
give a nod	1	2	0	1	*take revenge*	0	0	0	0
give a profile	1	0	2	1	*take a gamble*	0	0	1	0.33
give a rating	1	2	0	1	*take a photograph*	0	1	0	0.33
give a shudder	1	1	1	1	*take a risk*	0	0	1	0.33
give a speech	2	1	0	1	*take an appearance*	1	0	0	0.33
give a start	2	-	0	1	*take precedence*	0	0	1	0.33
give a taste	3	0	0	1	*take retirement*	0	0	1	0.33
give an impression	1	2	0	1	*take root*	0	1	0	0.33
give backing	1	2	0	1	*take shape*	0	1	0	0.33
give clearance	1	2	0	1	*take time*	1	0	0	0.33
give priority	1	2	0	1	*take a shine*	1	0	-	0.5
give a kiss	2	2	0	1.33	*take a toll*	0	1	-	0.5
give a nudge	2	2	0	1.33	*take a decision*	1	0	1	0.67
give a quality	2	0	2	1.33	*take a significance*	0	1	1	0.67
give a range	1	1	2	1.33	*take advantage*	1	0	1	0.67
give a say	2	0	2	1.33	*take comfort*	1	0	1	0.67

Table C.1: Development expressions with their individual and consensus (average) human ratings, sorted by the latter.

give a sentence	2	2	0	1.33	*take delight*	1	0	1	0.67
give an outline	2	2	0	1.33	*take office*	1	0	1	0.67
give assistance	2	2	0	1.33	*take a position*	1	1	1	1
give help	2	0	2	1.33	*take a stance*	1	1	-	1
give room	2	0	2	1.33	*take a view*	1	1	1	1
give satisfaction	2	2	0	1.33	*take action*	1	-	1	1
give service	2	2	0	1.33	*take an interest*	1	1	1	1
give a score	1	2	-	1.5	*take employment*	2	0	1	1
give a base	1	0	4	1.67	*take heed*	1	-	1	1
give a flavour	1	2	2	1.67	*take notice*	1	1	1	1
give coverage	2	2	1	1.67	*take pride*	1	1	1	1
give offence	2	1	2	1.67	*take a job*	2	2	0	1.33
give a deal	2	2	2	2	*take training*	2	2	0	1.33
give a diet	2	2	2	2	*take management*	-	2	1	1.5
give a dimension	2	2	2	2	*take a course*	2	1	2	1.67
give a guide	2	2	-	2	*take a subscription*	4	1	0	1.67
give a job	2	2	2	2	*take a vote*	2	2	2	2
give a name	2	2	2	2	*take a knife*	-	1	4	2.5
give an example	2	2	2	2	*take a loan*	4	2	2	2.67
give an explanation	2	2	2	2	*take sugar*	4	0	4	2.67
give an order	2	2	2	2	*take a drink*	4	1	4	3
give body	2	4	0	2	*take a couple*	4	2	4	3.33
give credit	2	2	2	2	*take an amount*	4	2	4	3.33
give energy	2	2	2	2	*take a proportion*	3	4	4	3.67
give guidance	2	2	2	2	*take a quarter*	3	4	4	3.67
give notice	2	2	2	2	*take a bottle*	4	-	4	4
give permission	2	2	2	2	*take a piece*	4	4	4	4
give colour	2	4	2	2.67	*take a sip*	4	4	4	4
give evidence	3	2	3	2.67	*take delivery*	4	4	4	4
give a share	2	4	3	3	*take a newspaper*	5	5	5	5
give a treat	3	4	3	3.33	*take a notebook*	5	5	5	5
give a dose	4	4	4	4	*take a packet*	5	5	5	5
give a note	4	-	4	4	*take an envelope*	5	5	5	5
give an award	5	5	5	5	*take a bag*	5	5	5	5
give a shilling	5	5	5	5	*take a ball*	5	5	5	5
give a thing	5	5	5	5	*take a book*	5	5	5	5
give a bag	5	5	5	5	*take biscuit*	5	5	5	5
give a ball	5	5	5	5	*take a box*	5	5	5	5
give a book	5	5	5	5	*take a basket*	5	5	5	5
give a cigarette	5	5	5	5	*take a cup*	5	5	5	5
give a gift	5	5	5	5	*take a coin*	5	5	5	5
give a meal	5	5	5	5	*take a cigarette*	5	5	5	5
give a prize	5	5	5	5	*take a camera*	5	5	5	5
give a bowl	5	5	5	5	*take a dog*	5	5	5	5
give a box	5	5	5	5	*take a stick*	5	5	5	5

Table C.1: Development expressions with their individual and consensus (average) human ratings, sorted by the latter.

give a cake	5	5	5	5	*take aspirin*	5	5	5	5
give a car	5	5	5	5	*take a bowl*	5	5	5	5
give a card	5	5	5	5					
give a coat	5	5	5	5					
give a coin	5	5	5	5					
give bread	5	5	5	5					
give cash	5	5	5	5					
give coffee	5	5	5	5					

Table C.2: Test expressions with their individual and consensus (average) human ratings, sorted by the latter.

give					*take*				
expression	human ratings				expression	human ratings			
give a go	0	0	0	0	*take a cruise*	0	0	0	0
give chase	1	0	0	0.33	*take a trip*	0	0	0	0
give preference	1	0	0	0.33	*take a twist*	0	-	0	0
give a fright	1	1	0	0.67	*take an example*	0	-	0	0
give a shake	2	0	0	0.67	*take flight*	0	0	0	0
give a toss	2	0	0	0.67	*take hold*	0	-	0	0
give attention	1	2	0	1	*take stock*	0	0	0	0
give recognition	2	1	0	1	*take shelter*	0	0	0	0
give a cry	1	1	1	1	*take stage*	-	0	0	0
give a glimpse	1	2	0	1	*take trouble*	1	0	0	0.33
give aid	2	2	0	1.33	*take credit*	1	0	1	0.67
give exposure	2	2	0	1.33	*take a holiday*	0	2	0	0.67
give pride	2	2	0	1.33	*take aim*	1	0	1	0.67
give rein	2	0	2	1.33	*take note*	0	1	1	0.67
give a sense	2	2	0	1.33	*take fright*	1	1	1	1
give an edge	2	2	0	1.33	*take issue*	1	1	1	1
give an opinion	2	2	0	1.33	*take offence*	1	-	1	1
give a lecture	2	1	2	1.67	*take side*	1	1	1	1
give a course	2	2	2	2	*take a glass*	5	5	5	5
give a number	2	2	2	2	*take a key*	5	5	5	5
give an excuse	2	2	2	2	*take an attitude*	1	1	1	1
give an indication	2	2	2	2	*take a post*	1	2	-	1.5
give feedback	2	2	2	2	*take staff*	-	0	4	2
give power	2	2	2	2	*take a statement*	2	2	2	2
give a figure	2	2	3	2.33	*take a number*	1	2	4	2.33
give a set	2	2	3	2.33	*take a room*	4	0	4	2.6
give a lease	2	4	2	2.67	*take a lot*	4	2	4	3.33
give a sum	4	2	3	3	*take a sample*	4	4	4	4
give change	4	4	4	4	*take a slice*	4	4	4	4
give a grant	5	5	5	5	*take a swig*	4	4	4	4

Table C.2: Test expressions with their individual and consensus (average) human ratings, sorted by the latter.

give a letter	5	5	5	5		*take a letter*	5	5	5	5
give a penny	5	5	5	5		*take a ticket*	5	5	5	5
give a key	5	5	5	5		*take a card*	5	5	5	5
give a present	5	5	5	5		*take a pen*	5	5	5	5
give a ticket	5	5	5	5						
give a cup	5	5	5	5						
give an envelope	5	5	5	5						

Appendix D

Experimental pairs from Chapter 3

Table D.1: Test *verb–noun* pairs and their frequencies in the BNC, grouped by *verb* and divided into idiomatic and literal.

Idiomatic		Literal	
pair	frequency	pair	frequency
blow fuse	18	*blow bridge*	16
blow gasket	10		
blow hole	26		
blow mind	14		
blow trumpet	40		
		bring bag	30
		bring cup	41
catch fire	12	*catch insect*	10
catch attention	45	*catch rabbit*	20
catch breath	199	*catch trout*	26
catch fancy	127	*catch horse*	11
catch imagination	187		
cut rate	201	*cut tree*	77
cut dash	15	*cut wood*	38
cut cloth	20	*cut hand*	36
cut throat	85	*cut cake*	52
cut cord	16	*cut grass*	67
		cut rope	22
		cut wire	26
find tongue	16	*find bottle*	31
		find box	41
get wind	33	*get money*	1266
get hook	15	*get pudding*	16
get drift	30	*get wire*	15
get bird	28	*get book*	301

Table D.1: Test *verb–noun* pairs and their frequencies in the BNC, grouped by *verb* and divided into idiomatic and literal.

get hump	14	*get box*	117
get nod	28	*get brush*	20
get push	16	*get camera*	40
		get farm	17
		get mug	12
		get finger	58
		get glass	98
		get ball	141
		get car	526
		get chair	70
give birth	608	*give drug*	81
give notice	581	*give gift*	78
give way	1211	*give mug*	15
give creep	27	*give land*	90
give sack	29	*give ticket*	81
give slip	46	*give drink*	90
give flick	14		
give lift	285		
give push	82		
give whirl	10		
have moment	195	*have cash*	113
have misfortune	78	*have leg*	262
have nerve	127	*have shell*	37
have wit	56	*have window*	188
have chip	90	*have bed*	139
have fling	21	*have flat*	53
have future	221	*have pool*	89
		have showroom	14
		have telephone	139
hit ceiling	10	*hit man*	40
hit deck	17		
hit headlines	81		
hit jackpot	33		
hit spot	22		
hit wall	67		
hold fire	32	*hold bowl*	17
hold ground	38	*hold tray*	12
hold hand	1163	*hold baby*	81
hold horse	26	*hold bird*	12
hold tongue	43	*hold key*	149
		hold plate	18
keep watch	203	*keep pig*	34
keep grip	50	*keep horse*	41
keep tab	62		
keep cool	50		

Table D.1: Test *verb–noun* pairs and their frequencies in the BNC, grouped by *verb* and divided into idiomatic and literal.

keep end	18		
keep hand	179		
keep head	203		
keep secret	177		
keep word	68		
lay waste	37	*lay block*	10
		lay carpet	20
		lay pipe	12
lose face	48	*lose money*	326
lose ground	105	*lose deposit*	21
lose touch	116	*lose home*	86
lose head	62	*lose ticket*	10
lose rag	15		
lose shirt	11		
lose temper	238		
make history	160	*make biscuit*	20
make peace	167	*make custard*	20
make beeline	14	*make pancake*	11
make hit	27	*make pie*	51
make killing	41	*make plastic*	17
make pile	28	*make scone*	14
make debut	587	*make toy*	24
make mark	260	*make cake*	104
move house	160	*move car*	37
move mountain	18		
pull finger	18	*pull box*	16
pull hair	80	*pull chair*	89
pull leg	76	*pull shirt*	19
pull weight	51		
pull chain	18		
push luck	37	*push barrow*	11
push boat	30	*push trolley*	33
push paper	17	*push bike*	24
put flesh	17	*put box*	46
put gloss	10	*put candle*	22
		put car	60
		put helmet	10
		put key	50
see daylight	17	*see woman*	166
see red	21	*set tank*	49
see sight	47		
set fire	283	*set carriage*	24
set cap	11		
set stage	78		
shoot bolt	16		

Table D.1: Test *verb–noun* pairs and their frequencies in the BNC, grouped by *verb* and divided into idiomatic and literal.

smell rat	19		
take air	58	*take lunch*	50
take biscuit	31	*take box*	83
take ease	17	*take handkerchief*	59
take heart	87	*take notebook*	26
		take arm	296
		take plate	57
		take boat	61
		take folder	14
		take gun	54
		take prize	83
		throw brick	38
		throw hat	11
		touch forehead	14
		touch shoulder	58
		touch finger	27
		throw towel	63

Appendix E

Rankings over test verb–noun pairs

This appendix contains top and bottom portions of two ranked lists of test verb–noun pairs, given by two measures of idiomaticity from Chapter 3, i.e., PMI and $\text{Fixedness}_{\text{overall}}$, respectively. For the ease of comparison, idiomatic pairs are displayed in boldface.

Table E.1: Test pairs ranked by PMI.

Top 30 pairs	Bottom 30 pairs
blow trumpet	***take heart***
blow gasket	***hold horse***
hit jackpot	*have pool*
smell rat	***have moment***
blow fuse	*give mug*
push barrow	*get mug*
push trolley	*take lunch*
hit headline	*take boat*
lose temper	*give drink*
catch trout	*make biscuit*
throw towel	*have flat*
shoot bolt	***keep word***
keep cool	*lose ticket*
keep tab	*hold bird*
touch forehead	*take box*
hit deck	*have shell*
throw brick	*have leg*
cut grass	*give land*
keep watch	***lose head***
catch breath	*get finger*
touch shoulder	*get farm*
cut throat	*take gun*
push luck	*cut hand*
catch fire	***take air***
catch imagination	*have cash*
lay waste	*have window*
lay carpet	*put car*
pull chair	***get bird***
catch fancy	*have bed*
cut dash	***keep end***

Table E.2: Test pairs ranked by $\text{Fixedness}_{\text{overall}}$, with $M = 50$ and $\alpha = .6$.

Top 30 pairs	Bottom 30 pairs
lose rag	***get bird***
give creep	*catch horse*
hit headline	***keep secret***
smell rat	***cut rate***
push luck	*find bottle*
take ease	*have leg*
lose temper	*take boat*
blow trumpet	*have bed*
keep cool	*give drug*
get drift	*get camera*
blow gasket	*put car*
make beeline	*make pie*
cut dash	*get car*
hit jackpot	*find box*
catch breath	*bring bag*
lose shirt	*take gun*
pull finger	*give ticket*
hold tongue	*get chair*
keep tab	*take prize*
give flick	*make cake*
give birth	*give drink*
throw hat	*get box*
pull weight	*take box*
give whirl	*give land*
find tongue	*give gift*
catch fire	*get book*
shoot bolt	*put box*
make debut	*keep horse*
push barrow	*hit man*
touch finger	*see woman*

Appendix F

Canonical forms from Chapter 3

This appendix contains canonical forms (Cforms) for the 100 idiomatic test pairs, as given by our method and the two idiom dictionaries, Oxford Dictionary of Current Idiomatic English (Cowie et al., 1983), and the Collins COBUILD Idioms Dictionary (Seaton and Macaulay, 2002). Each test item appears in its most dominant form according to the BNC; each Cform is a number that represents a pattern in Table 3.1. For each test item, precision and recall of our Cform identification method are also given.

Table F.1: Individual precisions and recalls of automatically identifying canonical forms.

Test expression	Automatically determined Cforms	Cforms taken from dictionaries	%*Precision*	%*Recall*
take the biscuit	3	3	100	100
take the air	3	3	100	100
take one's ease	5	5	100	100
take heart	1	1	100	100
smell a rat	2	2	100	100
shoot one's bolt	5	5	100	100
set one's cap	5	5	100	100
set fire	1	1	100	100
see red	1	1	100	100
see daylight	1	1	100	100
put flesh	1	1	100	100
push the boat	3	3	100	100
push one's luck	5	5	100	100
pull one's weight	5	5	100	100
pull one's hair	5	5	100	100

Table F.1: Individual precisions and recalls of automatically identifying canonical forms.

move mountains	6	6	100	100
move house	1	1	100	100
make one's debut	5	5	100	100
make a killing	2	2	100	100
make a hit	2	2	100	100
make a beeline	2	2	100	100
make history	1	1	100	100
lose one's temper	5	5	100	100
lose one's shirt	5	5	100	100
lose one's rag	5	5	100	100
lose one's head	5	5	100	100
lose ground	1	1	100	100
lose face	1	1	100	100
lay waste	1	1	100	100
keep one's word	5	5	100	100
keep one's head	5	5	100	100
keep one's end	5	5	100	100
keep one's cool	5	5	100	100
keep a grip	2	2	100	100
keep tabs	6	6	100	100
hold one's tongue	5	5	100	100
hold one's hand	5	5	100	100
hold one's ground	5	5	100	100
hold one's fire	5, 1	5, 1	100	100
hit the jackpot	3	3	100	100
hit the headlines	7	7	100	100
hit the deck	3	3	100	100
hit the ceiling	3	3	100	100
have the misfortune	3	3	100	100
give the slip	3	3	100	100
give the sack	3	3	100	100
give the creeps	7	7	100	100
give a whirl	2	2	100	100
give a lift	2	2	100	100
give way	1	1	100	100
give birth	1	1	100	100
get the push	3	3	100	100
get the nod	3	3	100	100
get the hump	3	3	100	100
get wind	1	1	100	100
find one's tongue	5	5	100	100
cut the cord	3	3	100	100
cut a dash	2	2	100	100
cut rates	6	6	100	100
catch one's fancy	5	5	100	100

Table F.1: Individual precisions and recalls of automatically identifying canonical forms.

catch one's breath	5	5	100	100
catch fire	1	1	100	100
blow one's trumpet	5	5	100	100
blow one's mind	5	5	100	100
blow a hole	2	2	100	100
blow a gasket	2	2	100	100
blow a fuse	2	2	100	100
have a future	2	2	100	100
make one's mark	5	5, 2	100	50
have the nerve	3	3, 2	100	50
catch the imagination	3	3, 5	100	50
have a fling	2	2, 5	100	50
set the stage	3, 10	3	50	100
see the sights	7, 6	7	50	100
put a gloss	2, 3	2	50	100
pull one's fingers	9, 5	5	50	100
pull one's leg	5, 3	5	50	100
keep watch	1, 2	1	50	100
hit the wall	3, 2	3	50	100
hit the spot	3, 7	3	50	100
have the wit	3, 9	9	50	100
have a moment	2, 9	9	50	100
get the bird	3, 2	3	50	100
get one's drift	5, 3	5	50	100
get a hook	2, 9	9	50	100
cut one's throat	5, 11	5	50	100
catch one's attention	5, 3	5	50	100
hold one's horses	9, 3, 7	9	33	100
cut one's cloth	5, 1, 3	5	33	100
make a pile	2, 6	2, 5	50	50
give notice	1, 11	1, 10	50	50
secrets are kept	10	2	0	0
push the papers	7, 3, 6	1	0	0
pull the chain	3	5	0	0
make peace	1	5	0	0
lose touch	1	5	0	0
keep one's hands	9	5	0	0
have chips	6	5	0	0
give a push	2	3	0	0
give a flick	2	3	0	0

Appendix G

Annotation guide

Thank you for your participation! Please read this guide carefully before starting the task. If any of the instructions or examples are unclear, please contact me:

[contact information]

Overview

In this task, you are asked to annotate entries in a list, by assigning a class label to each. Each entry is an English expression of the form "*verb* [*det*] *noun*", in which *det* can be any determiner (or no determiner). In this guide, the expressions are also referred to as verb+noun combinations. Examples are *make a cake*, *cut taxes*, *find happiness*, *give a groan*, *make one's mark*, and *hit the road*. Our final goal is to use the annotated verb+noun combinations for the evaluation of a computational model. A description of the classes, and more details on the procedure of assigning class labels to the expressions, are explained in this guide.

Classes

In our experiments, we consider four classes of verb+noun combinations, categorized by the extent and type of the semantic contribution of the constituents to the whole meaning of the expression. Here is a brief description of each class:

A verb+noun combination can have a literal or a non-literal interpretation. We define a

literal combination to be one in which the verb contributes its "basic" meaning that involves a physical action, as in *cut the bread*, *find a pen*, *make a cake*, and *give a present*. Note that this definition of the term *literal* may not match other definitions you might be familiar with. Nonetheless, for the sake of clarity and consistency, we choose this conservative definition for our annotation task. For the basic meanings of the verbs, you should use the ones that we provide at the end of this guide (see last page).

In a non-literal combination, the verb or the noun or both contribute a meaning that to some degree deviates from the basic meaning of the constituent. The following paragraphs explain three classes of non-literal combinations.

One class of non-literal combinations are those in which the verb contributes a metaphorical extension of its basic literal (physical) semantics, as in *cut taxes* and *find happiness*. Like literal combinations, these are compositional verb phrases in which the verb is the main source of predication, and the noun contributes its nominal semantics. However, they are different from literal combinations, since the verb contributes an abstract meaning (which is a metaphorical extension of its basic physical semantics). For example, *cut* in *cut taxes* has an abstract meaning ("decrease something" rather than "physically cut something"). We refer to these as **abstract combinations**.

Abstract combinations contrast with another group of non-literal verb+noun combinations called **light verb constructions (LVCs)**. LVCs, such as *make a suggestion* and *give a groan*, are also largely-compositional—i.e., their meaning can be derived from that of their constituents, although mainly from the noun. In fact, in LVCs, the main source of predication is the noun constituent, e.g., *make a suggestion* roughly means *suggest* and *give a groan* roughly means *groan*. The verb contributes a highly abstract meaning (or very little meaning).

Idiomatic combinations form another class of non-literal verb+noun combinations. In contrast to abstract combinations and LVCs, idiomatic combinations are non-compositional to a large extent. In other words, there is no clear direct relation between the meaning of an idiomatic expression and the meanings of its constituents outside that expression (except per-

haps a rather indirect metaphorical and/or historical relation). Examples are *cut corners* ("to do something in the easiest way"), *make a killing* ("to earn a lot of money very easily"), and *give a whirl* ("to try"). Note that even when an idiomatic combination has a somewhat compositional semantics, there are key connotations to the idiomatic expression that are an essential part of its meaning and that are not predictable from the individual components. Examples are *make a beeline* ("to move quickly towards") and *take the bait* ("to accept something that was offered to get one to do something"). In these cases, it is possible to draw a metaphorical relation between the idiomatic meaning and the meanings of the components. However, the idiomatic meaning also involves some extra connotations that emerge from the combination and not from the individual components.

Here is a pictorial representation of the above-mentioned characterizations that help define the classes. For each class, an abbreviated class label (in boldface) and an example are given in parentheses:

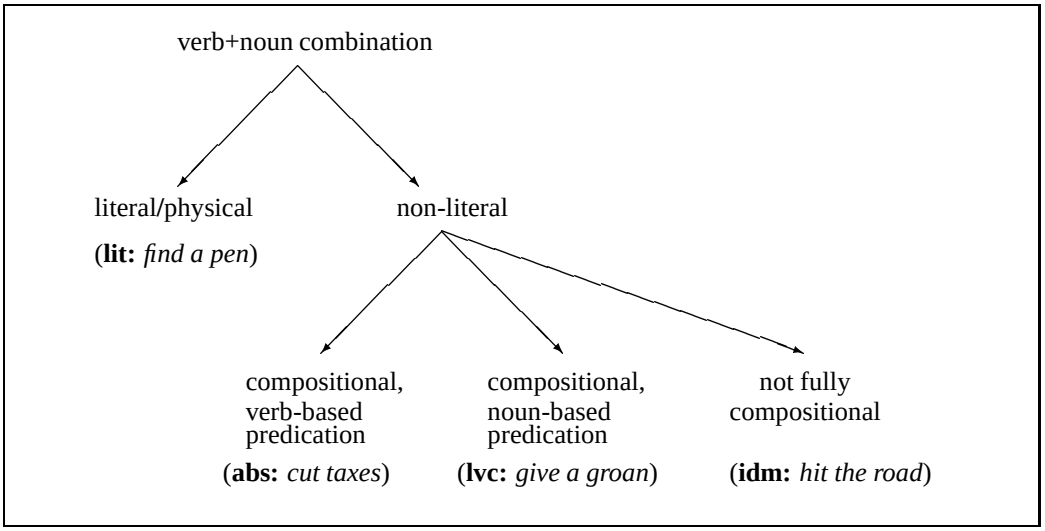


Figure G.1: A pictorial representation of classes and their properties.

Note that properties such as compositionality are a matter of degree. Hence, in reality verb+noun combinations often fall on a continuum with literal (physical) and idiomatic expressions on the two ends, and abstract combinations and LVCs in between. Deciding which

of these four classes a particular expression belongs to thus may not always be as clear-cut as one may desire. Although there are definitely some borderline cases, the goal here is to find core members of each class.

Also note that an expression can be ambiguous; e.g., *kick the bucket* means "to die" when used as an idiom, but it can also be used as a literal expression, as in "*Jill kicked the red bucket instead of the yellow one.*". Another example is *give a speech* that can be used as an LVC, as in "*Tim gave a speech to a thousand students*.", or as a literal combination, as in "*Clinton's PR person gave a speech to Hillary to use since she did not have time to write her own.*". It is important that you annotate the typical (common) use of an expression, e.g., even though *give a speech* can have a literal interpretation, it is typically used as an LVC and hence it should be given the label **lvc**. Similarly, since *kick the bucket* is typically used as an idiom, it should be labeled **idm**. You should use your judgment about which usage of an ambiguous expression is most common.

Task Description

You are given a list with entries of the following form:

verb [*det*] *noun* (referred to as *expression* or *verb+noun combination*)
related_verb

where *related_verb* is a verb morphologically related to *noun*, if such a verb exists according to an existing online lexical resource; it is "<>" otherwise. Here is a sample list:

---

see a film
film

make a cake
<>

cut taxes

tax

give a speech

speak

take a walk

walk

make a killing

kill

You are asked to assign a class label (**lit**, **abs**, **lvc**, **idm**) to each *expression*. This section gives you hints on how some properties of an expression and its constituents may help you with the classification task.

lit: If *expression* is largely compositional with *verb* contributing its basic predicative meaning that involves a physical action, then *expression* is most likely a **literal combination** (use the basic meanings of verbs provided on last page). In most (but not necessarily all) such expressions, *noun* is a concrete noun, as in *get a jacket*, *keep tools*, and *make a cake*.

abs: If *expression* is largely compositional with *verb* being the main source of predication, contributing a metaphorical (abstract) extension of its literal (physical) meaning, then *expression* is most likely an **abstract combination**. In most such expressions, *noun* is an abstract noun referring to a state or property, as in *find happiness*, *give a reason*, *bring awareness*, and *lose concentration*. Although note that it is not necessarily the case that the noun is a state or property, as in *cut taxes*, *hold an auction*, *take the stairs*, and *put a smile*. Also note that in some abstract combinations, the abstract semantics contributed by the verb also includes its basic literal meaning, as in *see a doctor* ("seeing" the doctor in the visit sense generally also involves a physical perception with eyes).

lvc: If *expression* is largely compositional with *noun* contributing a predicative meaning, *expression* is most likely a **light verb construction**. In an LVC, *noun* is the main source of predication and has a morphologically related verb, as in *make a pledge* (*pledge*), *make a suggestion* (*suggest*), *give a speech* (*speak*), *take a puff* (*puff*), *give a profile* (*profile*), and *take pleasure* (*please*).

An important clue to the recognition of LVCs is the existence of an entailment relation between *expression* and the corresponding *related_verb*. In most LVCs, "*expression*" entails "*related_verb*"—i.e., "*expression*" can be roughly paraphrased by "*related_verb*" (except for some differences in aspectual properties).

In some cases, it might be difficult to distinguish between abstract combinations and LVCs. Recall that in an abstract combination, *verb* is the primary source of predication, while in LVCs *noun* is the main determiner of the predicative meaning of the expression. The degree of "lightness" of *verb* (i.e., the degree to which it is devoid of semantic content) should give hints for making this distinction.

In other words, in an LVC, *verb* is not contributing much as a predicate (although it may contribute some aspectual properties); rather it is *noun* that has an argument structure and contributes it to the whole expression.

idm: If *expression* has a largely non-compositional and/or non-transparent meaning, then *expression* is most likely an **idiomatic combination**. This means that the expression as a whole has a particular meaning that is difficult to get from the components. For example, it is hard for non-native speakers of English to get the (idiomatic) meaning of the expression if they had not heard it before. Here are some examples with their idiomatic interpretation given in parentheses: *spill the beans* ("to reveal a secret"), *take a powder* ("to run away"), *make a killing* ("to earn a lot of money very easily"), and *take the bait* ("to accept something that was offered to get one to do something").

Note that idiomatic combinations often have an opaque (non-transparent) meaning. This

is true even when some relation can be thought of between the constituents of the idiom and those of the corresponding idiomatic interpretation. For example, linguists often state that *spill* in *spill the beans* means "reveal", and *beans* refers to "secret(s)". Although drawing such relations may help understand the idiom, it does not imply that *spill the beans* is compositional.

Moreover, it is important that you separate idioms from abstract combinations, since there are idiomatic expressions in which the verb contributes a metaphorical meaning, e.g., *take the bait*. However, as explained previously, the whole expression has extra connotations that cannot be derived by combining the metaphorical meaning of the verb with the meaning of the noun.

Summary of the Classes

lit: literal combination (physical-action use of the verb);
examples: *cut the bread*, *find a pen*, *make a cake*, *give a present*, *keep tools*, *get a jacket*, *give a measurement*, *see a film*.

abs: abstract combination (non-literal, compositional, verb-based predication);
examples: *cut taxes*, *find happiness*, *find fulfillment*, *bring awareness*, *bring excitement*, *put a smile*, *hold an auction*, *take the stairs*, *give a reason*.

lvc: light verb construction (non-literal, compositional, noun-based predication);
examples: *make a suggestion* (*suggest*), *make a pledge* (*pledge*), *give a speech* (*speak*), *give a groan* (*groan*), *take a walk* (*walk*), *take pleasure* (*please*), *make a proposal* (*propose*).

idm: idiomatic combination (non-literal, largely non-compositional).
examples: *cut corners* ("to do something in the easiest way"), *make a killing* ("to earn a lot of money very easily"), *give a whirl* ("to try"), *take the bait* ("to accept something that was offered to get one to do something"), *make a beeline* ("to move quickly towards"), *spill the beans* ("to reveal a secret"), *take a powder* ("to run away"), *make one's mark* ("to gain distinction").

General Guidelines

- Both *expression* and *related_verb* are given to you out of context. In most cases, you may need to put them in context and possibly add arguments (complements) in order to interpret them so you can annotate them. For example, the idiomatic combination *give a whirl*

requires a second object to be interpretable: "*to give* <**something**> *a whirl*" means "*to try* <**something**>". Also, the LVC *take delight* requires a PP argument with the preposition *in*: "*to take delight in* <**something**>". Note that in some cases, the arguments may (seemingly) be arguments of the noun, as in "*get wind of* <**something**>".

- When putting an expression in context, you should remember that *noun* can only be the **head noun** of a **direct object**. In other words, you should avoid turning an expression into one that is perhaps more familiar, but uses *noun* as a part of a larger noun phrase (though not as its head), or as an indirect object (recipient).

- Some expressions may be grammatical only in a negative context, e.g., "*he did not give a hoot about us*".

- Some expressions that are given with no determiner may require a quantifier such as *some, any, no* to be grammatical, e.g., "*she saw no sign*", "*that makes some/no sense to me*".

- Some expressions with *get* are only acceptable when the verb is used in the form "*have got*", e.g., "*she's got a problem*", "*he's got sense*". (This is the usage of the second sense of *get* on last page.)

- *expression* and *related_verb*, even when they have close or related meanings, may require different arguments (in terms of number and/or position). Note the following examples:

  People take pleasure in music.
  Music pleases people.

  Jill demonstrated the new software. ("*the new software*" is obligatory)
  Jill gave a demonstration [*of the new software*]. ("*the new software*" is optional)

- When comparing the meaning of *expression* with that of *related_verb*, you should **not** stretch to find some very specific context(s) in which *expression* and *related_verb* have closely related meanings. Instead, you should place them in natural-sounding contexts.

- Since the related verbs are extracted automatically, there are cases where the given related verb is not the best you can think of. There might also be expressions for which no related verb is given (<>), but you know of one. If you think either of these situations is the case for a given expression, please write down your suggested related verb in

parentheses in front of the given related verb, and base your judgment on the verb you supply. Also you should consider the given related verbs as suggestions only. If they do not make any sense, you can simply ignore them (or use your own suggested related verb).

- The expressions are automatically collected from a corpus of English (i.e., the British National Corpus (BNC)) with a restriction on their frequency of occurrence, and thus they are all expected to be acceptable expressions. Also, the expressions are presented in their most dominant form according to the BNC. This involves the choice of the determiner (e.g., *a/an, the*) and the number of the noun (i.e., plural or singular). If a particular item (*expression* or *related_verb*) does not sound acceptable to you, try searching it on **Google**; this may give you some hints about its possible meanings, although be aware of the noise on **Google**!! You might also try searching for "negative" uses, e.g., for the pair *see sign*, you may need to search for *see no sign* or *see any sign* to find natural-sounding expressions.

- For a given item (*expression* or *related_verb*), you may simultaneously think of two or more different senses. Pick the one sense that seems more prominent (common) to you and base your judgment on that.

- Our preference is that you do not leave any of the expressions without annotation. However, if you cannot think of any meanings for an expression, or have no clue as to what class it should be assigned to, mark it with "**?**". However, please avoid being indecisive by using **?** for too many expressions.

- Annotate as many expressions as you can in a single round. This helps you to be consistent in your annotation.

- We ask you to go back through the annotated expressions and revise them if needed. This way, you can check for internal consistency: your annotations are consistent if "similar" expressions are assigned "similar" class labels. To make this easier for you, we will take your annotations and group the items according to the class labels you assign to them. We will give you back 4 different lists, one for each class (**lit**, **abs**, **lvc**, **idm**). Each list is sorted by the verb, so expressions with the same verb are grouped together. You should go through these lists and make sure expressions that are in one class are similar, and that you make consistent judgments within and across classes for the same verb.

Important Notes on Formatting

- Expressions are given to you in files with raw text format that can be edited using Microsoft Word. Since we need to automatically process your annotations, you should edit and save the files in raw text format (i.e., you should not change the original format of the files).

- Please write down the class label for each expression on a separate line after the related verb, and leave one empty line after this line. To make it easier for you, we have already provided 2 empty lines between each two entries. One is for you to provide the class label, the other one should be left as is. Here is a sample of how the input and output files should look like:

Input	Output
see a film	see a film
film	film
	lit
cut taxes	cut taxes
tax	tax
	abs
make a plea	make a plea
<>	<> **(plead)**
	idm

  Items specified in boldface in Output are things that you (the annotator) will type in. But note that you should **not** type them in boldface!

Do not forget to look at next page!

Basic Meanings of the Verbs

bring: physical taking of something with oneself (to someplace).

find: physical discovering, locating, or coming upon something (that one had not known the location of).

get: (1) physical obtaining or gaining of possession of something.
(2) physical having of possession of something (i.e., "*have got*" usage).

give: physical transferring of possession from oneself to another.

hold: physical having or maintaining of something within one's grasp (usually meaning with the hands, but could be any part of the body or the body itself).

keep: (1) physical maintaining of something.
(2) physical retaining of possession of something.

lose: physical missing or removal of possession of something.

make: physical creating or assembling of something.

put: physical placing of something in a location or position.

see: physical perception with one's eyes.

set: (1) physical placing of something in a location or position.
(2) physical adjusting of something to a certain position or value.

take: (1) physical obtaining of something or moving of something to oneself.
(2) physical moving of something with oneself (to someplace).

Appendix H

Experimental pairs from Chapter 4

This appendix contains the set of 84 unseen test verb–noun pairs used in the experiments reported in Chapter 4.

Table H.1: Test *verb–noun* pairs and their class labels as determined by the primary annotator.

verb–noun pair	assigned class label	verb–noun pair	assigned class label
get chop	**idm**	*bring charge*	**lvc**
get hang	**idm**	*bring prosecution*	**lvc**
get nerve	**idm**	*get help*	**lvc**
give blessing	**idm**	*get sentence*	**lvc**
hold tongue	**idm**	*give appearance*	**lvc**
hold water	**idm**	*give debut*	**lvc**
keep cool	**idm**	*give explanation*	**lvc**
keep tab	**idm**	*give salute*	**lvc**
lose cool	**idm**	*give snort*	**lvc**
lose ground	**idm**	*give vent*	**lvc**
make difference	**idm**	*make announcement*	**lvc**
make face	**idm**	*make comment*	**lvc**
make getaway	**idm**	*make evaluation*	**lvc**
make track	**idm**	*make impact*	**lvc**
put match	**idm**	*make reduction*	**lvc**
see fit	**idm**	*make reply*	**lvc**
set seal	**idm**	*make splash*	**lvc**
take air	**idm**	*make survey*	**lvc**
take liberty	**idm**	*make total*	**lvc**
take pain	**idm**	*take bath*	**lvc**
take root	**idm**	*take pride*	**lvc**
bring trouble	**abs**	*find evidence*	**lit**
find comfort	**abs**	*get car*	**lit**

Table H.1: Test *verb–noun* pairs and their class labels as determined by the primary annotator.

find variation	**abs**	*get cash*	**lit**
get hand	**abs**	*get drive*	**lit**
get kid	**abs**	*get fish*	**lit**
get result	**abs**	*get gun*	**lit**
give guide	**abs**	*get hat*	**lit**
give idea	**abs**	*get phone*	**lit**
give say	**abs**	*give tool*	**lit**
give time	**abs**	*keep level*	**lit**
hold share	**abs**	*make frame*	**lit**
hold stock	**abs**	*make protein*	**lit**
keep word	**abs**	*put leg*	**lit**
lose status	**abs**	*put shoulder*	**lit**
make crossing	**abs**	*see bird*	**lit**
make profit	**abs**	*see launch*	**lit**
put price	**abs**	*see tear*	**lit**
see need	**abs**	*take cigarette*	**lit**
set structure	**abs**	*take drug*	**lit**
take chance	**abs**	*take girl*	**lit**
take possession	**abs**	*take sandwich*	**lit**

Appendix I

Precision and recall values for the classification task

This appendix contains recall (%*R*) and precision (%*P*) values per class, for the classification task described in Chapter 4. Performance figures are given for the five individual feature groups, as well as for all features combined.

Table I.1: Individual recall and precision values on TEST pairs, for individual feature groups, as well as all features combined.

	Only the features in group											
	INST		FIXD		COMP		VERB		NSEM		ALL	
Class	%*R*	%*P*	%*R*	%*P*	%*R*	%*P*	%*R*	%*P*	%*R*	%*P*	%*R*	%*P*
lit	47.6	47.6	42.9	40.9	52.4	50	61.9	48.2	76.2	45.7	71.4	51.7
abs	38.1	42.1	28.6	35.3	14.3	20	19	44.4	81	34.7	38.1	57.1
lvc	19	23.5	66.7	51.9	57.1	40	90.5	39.6	0	–	71.4	65.2
idm	38.1	29.6	61.9	72.2	38.1	47	0	–	0	–	52.4	61.1

Appendix J

Per-class inter-annotator agreements

This appendix contains the observed agreement and kappa scores among annotators for the four individual classes from Chapter 4.

Table J.1: Per-class observed agreement and kappa score between PA and each of the three annotators.

	ANNOTATOR$_1$		ANNOTATOR$_2$		ANNOTATOR$_3$	
	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

Bibliography

Anne Abeillé. 1995. The flexibility of French idioms: A representation with lexicalized Tree Adjoining Grammar. In Everaert et al. (1995), pages 15–42.

Minoji Akimoto. 1999. Collocations and idioms in Late Modern English. In Brinton and Akimoto (1999), pages 207–238.

Josep Alba-Salas. 2002. *Light Verb Constructions in Romance: A Syntactic Analysis*. Ph.D. thesis, Cornell University.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96, Sapporo, Japan.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of the Sixth Conference on Computational Natural Language Learning (CoNLL'02)*, pages 98–104, Taipei, Taiwan.

Colin Bannard. 2005. Learning about the meaning of verb-particle constructions from corpora. *Computer Speech and Language*, 19(4):467–478.

Colin Bannard, Timothy Baldwin, and 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*, pages 65–72, Sapporo, Japan.

Laurie Bauer. 1983. *English Word-formation*. Cambridge University Press.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for the nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 329–336, Trento, Italy.

BNC. 2000. *Reference Guide for the British National Corpus (World Edition)*, second edition.

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

Miriam Butt. 1997. Aspectual complex predicates, passives and disposition/ability. Talk held at the 1997 meeting of the Linguistics Association of Great Britain (LAGB'97).

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

Cristina Cacciari. 1993. The place of idioms in a literal and metaphorical world. In Cacciari and Tabossi (1993), pages 27–53.

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

Ray Cattell. 1984. *Composite Predicates in English*, volume 17 of *Syntax and Semantics*. Academic Press Australia.

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*, pages 115–164. Lawrence Erlbaum.

Claudia Claridge. 2000. *Multi-word Verbs in Early Modern English: A Corpus-based Study*. Editions Rodopi B. V., Amsterdam–Atlanta.

Eve V. Clark. 1978. Discovering what words can do. *Papers from the Parasession on the Lexicon*, 14:34–57.

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

Paul Cook and Suzanne Stevenson. 2006. Classifying particle semantics in English verb-particle constructions. In *Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 45–53, Sydney, Australia.

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory*. John Wiley and Sons, Inc.

Anthony P. Cowie. 1992. Multiword lexical units and communicative language teaching. In Pieree J. L. Arnaud and Henri Béjoint, editors, *Vocabulary and Applied Linguistics*, pages 1–12. Anthony Rowe Ltd.

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

Ido Dagan, Fernando Pereira, and Lillian Lee. 1994. Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the 32nd Anuual Meeting of the Association for Computational Linguistics (ACL'94)*, pages 272–278, Las Cruces, NM.

Giovanni B. Flores d'Arcais. 1993. The comprehension and semantic interpretation of idioms. In Cacciari and Tabossi (1993), pages 79–98.

Paul Deane. 2005. A nonparametric method for extraction of candidate phrasal terms. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 605–613, Ann Arbor, Michigan, June.

Marguerite Champagne Desbiens and Mara Simon. 2003. Déterminants et locutions verbales. Manuscript.

Mark Dras and Mike Johnson. 1996. Death and lightness: Using a demographic model to find support verbs. In *Proceedings of the 5th International Conference on the Cognitive Science of Natural Language Processing*.

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

Martin Everaert, Erik-Jan van der Linden, André Schenk, and Rob Schreuder, editors. 1995. *Idioms: Structural and Psychological Perspectives*. Lawrence Erlbaum Associates, Publishers.

Stefan Evert, Ulrich Heid, and Kristina Spranger. 2004. Identifying morphosyntactic preferences in collocations. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Afsaneh Fazly, Ryan North, and Suzanne Stevenson. 2005. Automatically distinguishing literal and figurative usages of highly polysemous verbs. In *Proceedings of the ACL'05 Workshop on Deep Lexical Acquisition*, pages 38–47, Ann Arbor, USA.

Afsaneh Fazly and 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)*, pages 337–344, Trento, Italy.

Afsaneh Fazly, Suzanne Stevenson, and Ryan North. 2006. Automatically learning semantic knowledge about multiword predicates. *Journal of Language Resources and Evaluation*. Accepted for Publication.

Christiane Fellbaum. 1993. The determiner in English idioms. In Cacciari and Tabossi (1993), pages 271–295.

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

Christiane Fellbaum. 2002. VP idioms in the lexicon: Topics for research using a very large corpus. In S. Busemann, editor, *Proceedings of the KONVENS 2002 Conference*, Saarbruecken, Germany.

Christiane Fellbaum. 2005. The ontological loneliness of verb phrase idioms. In Andrea Schalley and Dietmar Zaefferer, editors, *Ontolinguistics*. Mouton de Gruyter. Forthcomming.

Raffaella Folli, Heidi Harley, and Simin Karimi. 2003. Determinants of event type in Persian complex predicates. *Cambridge Working Papers in Linguistics*.

Bruce Fraser. 1970. Idioms within a transformational grammar. *Foundations of Language*, 6:22–42.

Raymond W., Jr. Gibbs. 1993. Why idioms are not dead metaphors. In Cacciari and Tabossi (1993), pages 57–77.

Raymond W., Jr. Gibbs. 1995. Idiomaticity and human cognition. In Everaert et al. (1995), pages 97–116.

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

Raymond W., Jr. Gibbs, Nandini P. Nayak, J. Bolton, and M. Keppel. 1989. Speaker's assumptions about the lexical flexibility of idioms. *Memory and Cognition*, 17:58–68.

Sam Glucksberg. 1993. Idiom meanings and allusional content. In Cacciari and Tabossi (1993), pages 3–26.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. The University of Chicago Press.

Lynn E. Grant. 2005. Frequency of 'core idioms' in the British National Corpus (BNC). *International Journal of Corpus Linguistics*, 10(4):429–451.

Gregory Grefenstette and Simone Teufel. 1995. Corpus-based method for automatic identification of support verbs for nominalization. In *Proceedings of the 7th Meeting of the European Chapter of the Association for Computational Linguistics (EACL'95)*.

Risto Hiltunen. 1999. Verbal phrases and phrasal verbs in Early Modern English. In Brinton and Akimoto (1999), pages 133–165.

Jerry R. Hobbs. 1979. Metaphor, metaphor schemata and selective inferencing. Technical Report Technical Note No. 204, SRI International, Menlo Park, CA.

Ray Jackendoff. 1997. *The architecture of the language faculty*. MIT Press.

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

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

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using Latent Semantic Analysis. In *Proceedings of the ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia.

Jerrold J. Katz. 1973. Compositionality, idiomaticity, and lexical substitution. In S. Anderson and P. Kiparsky, editors, *A Festschrift for Morris Halle*, pages 357–376. New York: Holt, Rinehart and Winston.

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

Parviz Khanlari. 1973. *Tarikh-e Zaban-e Farsi (The History of Persian Language)*. Bonyad-e Farhang.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'03)*, Edmonton, Canada.

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

Brigitte Krenn and Stefan Evert. 2005. Separating the wheat from the chaff – corpus-driven evaluation of statistical association measures for collocation extraction. *Sprache, Sprechen und Computer/Computer Studies in Language and Speech*, 8:104–117.

Merja Kytö. 1999. Collocational and idiomatic aspects of verbs in Early Modern English. In Brinton and Akimoto (1999), pages 167–206.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. The University of Chicago Press.

Mirella Lapata and Chris Brew. 2004. Verb class disambiguation using informative priors. *Computational Linguistics*, 30(1):45–73.

Dekang Lin. 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)*, pages 768–774, Montreal, Canada.

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

Tzong-Hong Lin. 2001. *Light Verb Syntax and the Theory of Phrase Structure*. Ph.D. thesis, University of California, Irvine.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts; London, England.

Alec Marantz. 1984. *On the nature of grammatical relations*. MIT Press.

Zachary J. Mason. 2004. CorMet: A computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44.

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

I. Dan Melamed. 1997a. Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the 2nd Conference on Empirical Methods for Natural Language Processing (EMNLP'97)*, Providence, USA.

I. Dan Melamed. 1997b. Measuring semantic entropy. In *Proceedings of the ACL-SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How*, pages 41–46, Washington, USA.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

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

Saif Mohammad and Graeme Hirst. 2005. Distributional measures as proxies for semantic relatedness. Submitted.

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

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

John Newman and 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.

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

Tim Nicolas. 1995. Semantics of idiom modification. In Everaert et al. (1995), pages 233–252.

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

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the Joint Conference on Empirical Methods for Natural Language Processing and Very Large Corpora*, pages 20–28.

Charles Kay Ogden. 1968. *Basic English, International Second language*. Harcourt, Brace, and World, New York.

Paul Pauwels. 2000. *Put, Set, Lay and Place: A Cognitive Linguistic Approach to Verbal Meaning*. LINCOM EUROPA.

J. Ross Quinlan. 1993. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.

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

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research (JAIR)*, (11):95–130.

Julia Ritz and Ulrich Heid. 2006. Extraction tools for collocations and their morphosyntactic specificities. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy.

Douglas L. T. Rohde. 2004. TGrep2 User Manual. URL: http://tedlab.mit.edu/d̃r/Tgrep2/.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and 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)*, pages 1–15, Mexico City.

Maggie Seaton and Alison Macaulay, editors. 2002. *Collins COBUILD Idioms Dictionary*. HarperCollins Publishers, second edition.

Violeta Seretan, Luka Nerima, and Eric Wehrli. 2003. Extraction of multi-word collocations using syntactic bigram composition. In *Proceedings of the International Conference RANLP'03*, Bulgaria.

Sidney Siegel and John, Jr. Castellan. 1988. *Nonparametric Statistics for the Behavioural Sciences*. McGraw-Hill.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

Frank Smadja, K. R. McKeown, and V. Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1).

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

Harumi Tanabe. 1999. Composite predicates and phrasal verbs in *The Paston Letters*. In Brinton and Akimoto (1999), pages 97–132.

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

Mohammad-Mehdi Vahedi-Langrudi. 1996. *The Syntax, Semantics and Argument Structure of Complex Predicates in Modern Farsi*. Ph.D. thesis, Université d'Ottawa, June.

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

Sriram Venkatapathy and Aravind Joshi. 2005b. Relative compositionality of noun+verb multi-word expressions in Hindi. In *Proceedings of International Conference on Natural Language Processing (ICON'05)*, IIT Kanpur, India.

Sriram Venkatapathy and Aravind Joshi. 2006. Using information about multi-word expressions for the word-alignment task. In *Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 53–60, Sydney, Australia.

Begoña Villada Moirón. 2004. Discarding noise in an automatically acquired lexicon of support verb constructions. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Begoña Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL'06 Workshop on Multiword Expressions in a Multilingual Context*, pages 33–40, Trento, Italy.

Aline Villavicencio, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. 2004. Lexical encoding of multiword expressions. In *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, pages 80–87, Barcelona, Spain.

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. 2005. Paradigmatic modifiability statistics for the extraction of complex multi-word terms. In *Proceedings of HLT-EMNLP'05*, pages 843–850.

Dominic Widdows and Beate Dorow. 2005. Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In *Proceedings of ACL'05 Workshop on Deep Lexical Acquisition*, pages 48–56.

Anna Wierzbicka. 1982. Why can you Have a Drink when you can't *Have an Eat? *Language*, 58(4):753–799.