Automatic Acquisition of Knowledge About Multiword Predicates

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

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

Abstract

Human interpretation of natural language relies heavily on cognitive processes involving metaphorical and idiomatic meanings. One area of computational linguistics in which such processes play an important, but largely unaddressed, role is the determination of the properties of multiword predicates (MWPs). MWPs such as *give a groan* and *cut taxes* involve metaphorical meaning extensions of highly frequent, and highly polysemous, verbs. Tools for automatically identifying such MWPs, and extracting their lexical and syntactic properties, are crucial to the adequate treatment of text in a computational system, due to the productive nature of MWPs across many languages. This paper gives an overview of our work addressing these issues. We begin by relating linguistic properties of metaphorical uses of verbs to their distributional properties. We devise automatic methods for assessing whether a verb phrase is literal, metaphorical, or idiomatic. Since metaphorical MWPs are generally semi-productive, we also develop computational measures of their individual acceptability and of their productivity over semantically related combinations. Our results demonstrate that combining statistical approaches with linguistic information is beneficial, both for the acquisition of knowledge about metaphorical and idiomatic MWPs, and for the organization of such knowledge in a computational lexicon.

1. Metaphorical Multiword Predicates

Metaphor is a powerful aspect of language, enabling creative expression in terms of familiar concepts, usually ones which are easily visualizable (Lakoff and Johnson, 1980; Johnson, 1987; Nunberg et al., 1994). Indeed, metaphor 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 metaphorical uses (Newman, 1996). 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 (MWPs) to form around certain high frequency verbs that easily undergo a process of 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*, *give*, *put*, *sit*). Their metaphorical uses yield a range of meaning extensions, exhibited in MWPs such as those in 1(a–d):

1. (a) cut taxes, cut in line, cut a dash
   (b) give permission, give a toss, give a groan, give sb. their due
   (c) put sth. to rest, put one's finger (on), put a gloss (on)
   (d) sit in judgment, sit on sth., sit tight, sit on the fence

As seen in the examples above, a wide variety of MWPs result from basic verbs in combination with various types of complements, lying along a continuum of less to more metaphorical usage, culminating in idiomatic expressions.

Multiword predicates of this type are widespread in many different languages, including, but not limited to, English, German, French, Spanish, Persian, Urdu, Hindi, Chinese, and Japanese (Seaton and Macaulay, 2002; Kearns, 2002; Fellbaum, 2002; Desbiens and Simon, 2003; Alba-Salas, 2002; Karimi, 1997; Butt, 2003; Lin, 2001; Butt and Scott, 2002; Miyamoto, 2000). Such MWPs contrast with verb-particle constructions (e.g., *give up* and *figure out*), which form a restricted set of MWPs that are common in English but not as widely attested crosslinguistically. Nonetheless, MWPs with basic verbs have been granted relatively little attention in the computational linguistics community. Consequently, fundamental issues about such expressions are only beginning to be addressed, such as: automatic extraction and acquisition of relevant properties; adequate representation in a computational lexicon; and appropriate treatment in various natural language processing tasks.

Because expressions with basic verbs fall on a continuum of literal to metaphorical to idiomatic, they cannot be treated uniformly in computational systems. Metaphorical expressions vary in the degree to which the meaning of the basic verb differs from its literal semantics, and even expressions considered idiomatic can vary in their level of semantic transparency. Such expressions also exhibit varying levels of tolerance for morphosyntactic variations, so that their appropriate syntactic treatment is also not uniform. We address these issues through a careful linguistic analysis of the properties of MWPs and their relation to statistical behaviour, to support the automatic acquisition of semantic and syntactic knowledge about MWPs.

2. Automatic Lexical Acquisition of Properties of MWPs

In this article, we give an overview of our research on the automatic acquisition of semantic and syntactic properties of English MWPs involving a basic verb. For simplicity, we use the general term MWP to refer to such expressions in the remainder of the paper. Because many MWPs are comprised of the verb and a noun in its direct object position (Cowie et al., 1983; Nunberg et al., 1994), we particularly focus on verb+noun combinations, or VNCs for short. In our recent body of work, we have addressed two overall problems in lexical acquisition of the properties of VNCs: first, of determining where on the continuum of literal to metaphorical to idiomatic a particular VNC lies; and second, of inferring the possible complements a basic verb can combine with to form an acceptable MWP.

Recognizing VNCs as literal, metaphorical, or idiomatic is complicated by the fact that non-literal VNCs conform to the grammar rules for verb phrases. Hence they are indistinguishable on the surface from regular compositional verb phrases. For example, the three expressions *take a lunch*, *take a powder*, and *take a walk* appear similar at first glance, but a closer look at their semantics reveals significant differences. Each of these three represents a different class of expressions using a basic verb. *Take a lunch* is a compositional combination of a verb and

a noun, both contributing their literal semantics. *Take a powder* has an idiomatic meaning ("to leave abruptly") that has nothing to do with either *take* or *powder*. *Take a walk* stands somewhere between the idiomatic and compositional expressions. It is not completely idiomatic because the noun constituent determines the primary meaning of the expression—that is, *take a walk* can be roughly paraphrased by the verb *walk*. However, it is not fully compositional either, since the verb constituent does not contribute its literal meaning, but rather a metaphorical extension of it. Because the verb is "semantically bleached" in such expressions (Butt, 2003), it is referred to as a light verb. *Take a walk* is one of a general class of VNCs known as light verb constructions (LVCs).

It is evident that the distinction between idiomatic and literal expressions is essential to NLP applications that require some degree of semantic interpretation. A machine translation system, for example, should translate the idiom *kick the bucket* as a single unit of meaning, while this is not the case for the literal phrase *kick the pail*. It is thus necessary to develop means for automatically distinguishing idiomatic VNCs from literal ones. This distinction is also important for LVCs. For example, *give a groan* is translated to the French verb *gémir* ("to groan"), while the literal *give a present* has a word-for-word translation. However, LVCs differ from idiomatic phrases in that, in most cases, their semantics (and hence their translation) is more predictable. Specifically, the meaning of an LVC typically corresponds to a verb related to the noun complement, as with the French translation *gémir* related to the English noun complement *groan*. In Section 3, we describe our work on the first step in the appropriate handling of VNCs: to automatically distinguish expressions that are literal, metaphorical (as in LVCs), and idiomatic.

The second issue regarding lexical acquisition of VNCs concerns the determination of which complements are semantically compatible with a basic verb in forming a multiword predicate. While idiomatic combinations are by definition semantically idiosyncratic, LVCs show a greater degree of compositionality and hence also semantic predictability. However, it is still extremely difficult to determine which nouns can combine with a given light verb to form an LVC. This depends on the semantic properties of both the noun and the light verb; for example, one can *take a walk* and *give a groan*, but it is less natural to ?*take a groan* or ?*give a walk*. Interestingly, light verbs often tend to combine with semantically similar complements to form families of semantically-related LVCs (Wierzbicka, 1982; Nunberg et al., 1994; Sag et al., 2002). For example, one can *take a walk*, *take a stroll*, or *take a run*; similarly it is acceptable to *give a groan*, *give a smile*, and *give a wink*. It is important in a computational system to recognize the allowable patterns of combination for a light verb, so that previously unseen LVCs can be handled appropriately. In Section 4, we describe our two-pronged approach to this problem, which involves computational measures for the individual acceptability of potential LVCs, as well as for the assessment of productivity over a class of semantically related potential complements.

3. Literal, Metaphorical, or Idiomatic VNCs

Here we summarize our work on the determination of whether a VNC is literal, metaphorical, or idiomatic. We have thus far broken down the problem into two parts. First, we address the endpoints on the scale of metaphoricity by distinguishing idiomatic expressions from literal ones (Section 3.1). Then we tackle the distinction of degree of metaphoricity of a potential LVC, which also helps to distinguish LVCs from literal expressions (Section 3.2). Our current work focuses on combining these approaches into one measure that places a VNC on a scale of literal to metaphorical to idiomatic.

3.1. Literal vs. Idiomatic

Compared to literal expressions, idiomatic VNCs are notable for their semantic idiosyncrasy. For example, *kick the bucket* has a meaning ("to die") that is relevant neither to the independent meaning of *kick* nor to that of *bucket* (Cacciari, 1993; Sag et al., 2002). Semantic idiosyncrasy is a matter of degree, however: the idiom *spill the beans* ("to reveal a secret") is often argued to be less idiomatic than *kick the bucket*, because it can be analyzed as *spill* corresponding to "reveal" and *beans* referring to "secret(s)". Such idioms are referred to as semantically analyzable.[1]

There is evidence in the linguistic literature that the idiosyncrasy of idiomatic combinations is not limited to their semantics, but extends to their lexical and/or syntactic behaviour. Idiomatic VNCs are known to be lexically fixed (non-productive) to a large extent (Gibbs, 1993; Glucksberg, 1993). Neither *kick the pail* nor *hit the bucket* have meanings related to that of *kick the bucket*. Similarly, while *spill the beans* has an idiomatic interpretation, *spill the peas* and *spread the beans* are literal phrases. Interestingly, when an idiomatic VNC does show some (limited) lexical productivity, it is typically with respect to the verb constituent—that is, in some cases, a few semantically similar basic verbs can be used in a particular idiom (e.g., *keep one's cool, lose one's cool*).

Idiomatic combinations are also syntactically fixed to some extent, i.e., they typically cannot appear in syntactic variations while at the same time retaining their idiomatic interpretations (Stock et al., 1993; Fellbaum, 1993; Nunberg et al., 1994). The idiom *kick the bucket*, for example, is generally unacceptable in other syntactic forms:

2. (a) John kicked the bucket.
   (b) ?? John kicked a bucket.
   (c) ?? John kicked the buckets.
   (d) ?? John kicked the fast bucket.
   (e) ?? The bucket was kicked by John.

Syntactic flexibility of an idiom is argued to be strongly related to its semantic analyzability. The idiom *spill the beans*, as noted above, is considered to be semantically more transparent, and is correspondingly more flexible:

3. (a) John spilled the beans.
   (b) ? John spilled some beans.
   (c) ?? John spilled the bean.
   (d) John spilled the official beans.
   (e) The beans were spilled by John.

As sentences in 2 and 3 show, some idioms (e.g., those that are semantically analyzable) may be syntactically more flexible than others. Nonetheless, in general they are expected to appear in restricted syntactic constructions.

The lexical and syntactic fixedness of idiomatic VNCs contrasts with the behaviour of regular compositional verb phrases that tend to be more productive and appear in a wider range

[1] Semantic analyzability is also referred to as decomposability or compositionality in the linguistic literature.

of syntactic constructions. Examining the degree of fixedness of a verb+noun combination can thus be regarded as a way of determining its idiomaticity, i.e., the degree to which it is idiomatic. In our recent work on idioms (Fazly and Stevenson, 2005), we propose statistical measures for quantifying the lexical, syntactic, and overall fixedness of a VNC. Each measure brings together aspects of the above-mentioned linguistic properties of idioms and their distributional behaviour in a corpus.

We assume that a target VNC is lexically fixed, and hence idiomatic, if it is not open to lexical substitution. This means that we expect to see a notable difference between the idiomaticity level of the target VNC and that of the variants generated by replacing one of its constituents (the verb or noun) with a semantically (and syntactically) similar word. The paradox of this assumption is that it requires knowledge about the idiomaticity of the variant combinations. Inspired by Lin (1999), we moderate this assumption by examining the association strengths of the target combination and its variants, as an indirect cue to their idiomaticity. Target VNCs that are significantly greater in association strength than their variants are assumed to be more lexically fixed, and therefore more idiomatic. Our contribution is a novel technique for incorporating these association strengths into a single measure of lexical fixedness for the target expression.

We assume a target VNC is syntactically fixed, and hence idiomatic, if it mainly appears in restricted syntactic constructions. We thus extract, from the linguistic literature, a set of syntactic patterns expected to capture the difference in behaviour of idiomatic and literal VNCs (Nunberg et al., 1994; Fellbaum, 1993). Syntactic fixedness of a VNC is measured by comparing its probability distribution over the selected set of patterns, to that of a "typical" verb+noun combination. The more the distributional behaviour of the target VNC deviates from that of the typical VNC, the more likely it is to be idiomatic.

We evaluate our fixedness measures by applying them to the task of separating idiomatic from literal combinations, and comparing their performance to that of a collocation-based measure, pointwise mutual information (PMI) (Church et al., 1991). We find in our results that both the lexical and syntactic fixedness measures have good performance. While the lexical fixedness measure works comparably to PMI, the syntactic fixedness measure substantially outperforms both. Moreover, we show that combining the lexical and syntactic fixedness measures into a measure of overall fixedness results in a notable gain in performance. Our measures are also less sensitive to the frequency of the expressions than PMI; this is important since many idioms have low frequency of occurrence in traditional corpora.

3.2. Literal vs. Metaphorical

Compared to idiomatic VNCs, light verb constructions are more semantically analyzable, and thus lie between idioms and literal phrases on the continuum of metaphoricity. In an LVC, the noun constituent generally contributes its literal meaning to the expression, while the light verb contributes a more or less metaphorical meaning. A light verb can also be used in a literal expression with its core meaning. Hence the challenge is to determine the level of metaphoricity of a use of a light verb. For example, *give* in *give sb. a present* has a literal meaning, i.e., "transfer of possession" of a THING to a RECIPIENT. In an LVC such as *give a speech*, *give* has a metaphorical meaning, while at the same time keeping aspects of its literal semantics: an abstract entity (*a speech*) is "transferred" to the audience, but no "possession" is involved. In other LVCs such as *give a groan*, *give* has a highly metaphorical meaning: here the notions of

"transfer" and "possession" are almost completely diminished.

As with idiomatic VNCs, syntactic fixedness plays an important role in identifying degree of semantic analyzability. The noun complement of a literal use of a light verb (e.g., *present* in *give sb. a present*) acts as a direct object and hence can move around freely. LVCs whose noun constituent can be treated, possibly metaphorically, as the direct object of the light verb also exhibit syntactic flexibility to a large extent. In these, the noun may be introduced by a definite article, pluralized, passivized, or relativized, as in 4(b–e):

4. (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.

In contrast, LVCs involving highly metaphorical uses of the light verb enforce certain restrictions on the syntactic freedom of their noun constituents (Kearns, 2002), as in 5(b–e):

5. (a) Azin gave a groan just now.
   (b) ?? Azin gave the groan just now.
   (c) ? Azin gave a couple of groans last night.
   (d) ?? A groan was given by Azin just now.
   (e) ?? The groan that Azin gave was very long.

In general, the degree to which the light verb retains aspects of its literal meaning—and contributes them compositionally to the LVC—is reflected in the degree of freedom exhibited by the noun component. We have proposed a statistical measure that uses this insight to situate an LV+N combination (a potential LVC) on a scale of literal to metaphorical usage of the light verb (metaphoricity continuum) (Fazly et al., 2005). Our measure assigns a score to each potential LVC, reflecting its syntactic fixedness as determined by the degree of freedom of the noun component. This is approximated as the difference between the strength of association of the potential LVC with two types of syntactic patterns: those that are preferred by LVCs (as in 5(a)), and those that are less preferred by more metaphorical LVCs (as in 5(b–e)).

We evaluate our measure of metaphoricity by comparing the ratings it assigns to a set of candidate LV+N combinations with those assigned by human judges. Using the Spearman rank correlation coefficient, we show that the ratings of our measure achieve significant high correlations with the human judgments. Moreover, when the measure is applied to a restricted subset of LVCs, i.e., those of the form "LV a/an N", the correlation scores are particularly high. Since the more metaphorical LVCs are similar in some respects to idioms, we also compare our results to correlations attained by PMI, as a simple measure of collocation. Our measure outperforms PMI (on all the data, as well as on the restricted subset), indicating that degree of metaphoricity cannot be simply treated as degree of collocation.

4. Semantic Patterns in MWP Formation

Here we give an overview of our work on determining the semantic class of complements that can combine with a basic verb to form multiword predicates. Because LVCs (in contrast to idiomatic expressions) exhibit some predictability in terms of the allowable complements of a light verb, we have focused on this subclass of MWPs in our research on this topic. As mentioned above, we have developed computational measures for addressing two related aspects of this problem: individual acceptability of LV+N combinations, and assessment of productivity of a light verb with respect to a semantic class of potential complements.

4.1. Individual LVC Acceptability

First, we focus on the individual acceptability of potential LVCs. Although light verbs tend to have similar patterns of cooccurrence with semantically similar complements, they fail to exhibit full generality in their combinations with a semantic class of nouns. Hence, deciding precisely which light verbs combine with which nouns to form acceptable LVCs is a difficult task. Our first solution to this problem treated LV+N combinations as syntactically-dependent collocations, using a PMI-based measure (Stevenson et al., 2004). This measure outperformed the standard PMI measure because it incorporated some knowledge of the preferred LVC pattern (cf. the examples in 5 above). However, while common LVCs typically appear as good collocations, our collocation-based measure failed to take into account other important properties of LVCs.

We have devised an improved acceptability measure that brings together some of the linguistic properties of LVCs into a probability formula to determine the likelihood of a given light verb and noun forming an acceptable LVC (Fazly et al., 2005, 2006). The measure captures the general tendency of the noun to form LVCs with any light verb, as well as its specific inclination towards the particular light verb. We evaluate this measure by comparing its ratings on a set of expressions formed from combining candidate light verbs and nouns, with human judgments of acceptability on these LV+N combinations. Using the Spearman rank correlation coefficient, we show that our probabilistic acceptability measure generally achieves good (and significant) correlations with the human judgments. Moreover, we compare this measure to our earlier PMI-based measure, as well as to standard PMI. We find that our new measure performs better and more consistently across expressions formed from different light verbs and candidate nouns from different semantic classes.

4.2. Class-based LVC Acceptability

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

Moreover, each group of semantically similar nouns combining with a particular light verb is often deemed to distinguish a possible meaning extension for the light verb (Newman, 1996).

For example, in *give advice*, *give orders*, *give a speech*, etc., *give* contributes a notion of "abstract transfer", while in *give a groan*, *give a moan*, *give a rasp*, etc., *give* contributes a notion of "emission". Capturing the class-based tendency of light verbs in forming LVCs is thus essential to refining the semantic space of the metaphorical usages of these (and other) basic verbs, which are all highly polysemous (Fazly et al., 2005).

We examine the class-based tendency of LVC formation by focusing on potential LVCs formed from the combination of candidate light verbs with nouns from a number of selected semantic classes (Fazly et al., 2006). We draw on two different classifications, that of Levin (1993), and that of WordNet (Fellbaum, 1998). We define the productivity level of each semantic class with respect to a particular light verb to be the proportion of class members that form acceptable LVCs with the light verb. To extend our acceptability measure for assessing the productivity of a class, we set a threshold on the ratings assigned by the measure; LV+N combinations with ratings higher than the threshold are considered to be acceptable LVCs.

A good acceptability measure should accurately predict the individual acceptability level of an LV+N combination, as well as the collective acceptability (productivity level) of a semantic class with respect to a particular light verb. We compare, for each semantic class and light verb, the proportion of nouns that form acceptable LVCs according to our probabilistic measure (see Section 4.1), to the same figure as determined by the human judges. The divergence of the former from the latter is estimated by calculating the sum of squared errors between the two sets of numbers, averaged across all light verbs and semantic classes. The results show that our linguistically-motivated measure of acceptability has smaller divergence when compared with the collocation-based PMI measures.

5. General Discussion

Recently there has been a growing understanding both of the need for the appropriate handling of multiword expressions, and of the complexities involved in the task (Sag et al., 2002). Most research, however, has concentrated on the automatic extraction of these expressions (Grefenstette and Teufel, 1995; Dras and Johnson, 1996; Melamed, 1997; Baldwin and Villavicencio, 2002; Seretan et al., 2003; Moirón, 2004). Previous studies that focus on learning about semantic properties of multiword expressions, such as their compositionality, have mainly covered compound nouns (Wermter and Hahn, 2005), and verb-particle constructions (McCarthy et al., 2003; Bannard et al., 2003; Baldwin et al., 2003). Our work differs in focusing on those multiword predicates (MWPs) that involve a "basic" verb, a broadly-documented class of expressions that has received relatively little attention within the computational linguistics community.

Most previous work on compositionality of multiword expressions either treats them as collocations (Smadja, 1993), or examines the distributional similarity between an 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 (2005) 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 (VNCs) according to their compositionality.

Our work differs from such approaches in presenting an alternative view on compositionality. More specifically, we carefully examine several linguistic properties of metaphorical and idiomatic MWPs, those that distinguish them from literal (compositional) combinations. These

include characteristics related not only to lexical fixedness, but to syntactic fixedness as well. Widdows and Dorow (2005) also draw on the notion of syntactic fixedness for idiom detection, though their method is specific to a highly constrained type of idiom. Our work examines a broader range of syntactic patterns associated with a large class of MWPs. We then suggest novel techniques for translating these lexical and syntactic characteristics into measures that predict the level of metaphoricity or idiomaticity of MWPs.

Work indicating acceptability of multiword expressions is largely limited to collocational analysis using PMI-based measures (Lin, 1999; Stevenson et al., 2004). In our recent work, we have proposed a linguistically-motivated acceptability measure for light verb constructions that enables flexible integration of LVC-specific properties. In addition to distinguishing literal from metaphorical usages of a light verb, we aim to determine finer-grained distinctions among the identified metaphorical usages. In most cases, the finer-grained distinctions appear to relate to the semantic properties of the complement that combines with the light verb. Not only does a light verb tend to combine with semantically similar complements, it tends to contribute similar metaphorical meaning to the resulting LVC.

Semantic class knowledge may thus enable us to further refine the semantic space of a light verb by elucidating its relation with complements from different classes. Wanner (2004) attempts to perform a similar task by classifying VNCs into predefined groups, each corresponding to a particular semantic relation between the verb and the noun. However, his approach requires manually labelled training data. Villavicencio (2003) uses class-based knowledge to extend a lexicon of verb-particle constructions, but assumes that an unobserved expression is not acceptable. We instead propose that more robust application of class-based knowledge can be achieved with a better estimate of the acceptability of various expressions. While we focus on light verb constructions, we believe that similar techniques can be useful in dealing with other semi-productive MWPs, such as verb-particle constructions.

The significance of the role metaphor plays in language has long been recognized. However, due to the peculiarities in the behaviour of metaphorical and idiomatic expressions, they have been mostly overlooked by researchers in computational linguistics. Previous studies recognize the challenges these constructions impose on NLP systems (see, e.g., Fellbaum, 2005), but often lack proposals for robust and wide-coverage mechanisms to handle them. Some work has relied on the existence of expensive resources such as manually-built knowledge bases (Fass, 1991; Villavicencio et al., 2004). Mason (2004) incorporates automatically-induced knowledge about the domain of use of a verb to help identify different metaphorical meanings. However, highly polysemous verbs cannot be easily associated with particular domains. Hence such an approach overlooks the great potential of basic verbs in forming metaphorical and idiomatic MWPs.

Our work demonstrates that combining statistical approaches with linguistic information is beneficial in devising reliable techniques, both for the acquisition of knowledge about metaphorical and idiomatic MWPs, and for the organization of such knowledge in a computational lexicon. Given the crosslinguistic prominence of MWPs, our future work aims to extend these techniques to similar constructions in languages other than English. Moreover, while we have focused here on LVCs and idiomatic VNCs, we believe that similar techniques can be useful in dealing with other MWPs, and possibly other types of multiword expressions in general.

6. Acknowledgements

Thanks especially to our colleague, Ryan North, for his contributions toward a fruitful and enjoyable collaboration. We are also grateful to the rest of the computational linguistics research group at the University of Toronto for helpful comments and discussion on this and earlier papers on this topic. This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under a grant to the second author, and by Ontario Graduate Scholarships to the first author.

7. References

Alba-Salas, J. (2002). *Light Verb Constructions in Romance: A Syntactic Analysis*. PhD thesis, Cornell University.

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (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.

Baldwin, T. and Villavicencio, A. (2002). Extracting the unextractable: A case study on verb-particles. In *Proceedings of CoNLL'02*, pages 98–104, Taipei, Taiwan.

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

Butt, M. (2003). The light verb jungle. Workshop on Multi-Verb Constructions.

Butt, M. and Scott, B. (2002). Chinese directionals. Talk held as part of the Workshop on Complex Predicates, Particles and Subevents.

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

Cacciari, C. and Tabossi, P., editors (1993). *Idioms: Processing, Structure, and Interpretation*. Lawrence Erlbaum Associates, Publishers.

Church, K., Gale, W., Hanks, P., and Hindle, D. (1991). Using statistics in lexical analysis. In Zernik, U., editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Lawrence Erlbaum.

Cowie, A. P., Macking, P., and McCaig, I. R. (1983). *Oxford Dictionary of Current Idiomatic English*, volume 2. Oxford University Press.

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

Dras, M. and Johnson, M. (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*.

Fass, D. (1991). met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

Fazly, A., North, R., and Stevenson, S. (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.

Fazly, A., North, R., and Stevenson, S. (2006). Automatically determining allowable combinations of a class of flexible multiword expressions. Submitted.

Fazly, A. and Stevenson, S. (2005). Automatically constructing a lexicon of verb phrase idioms. In preparation.

Fellbaum, C. (1993). The determiner in English idioms. In Cacciari and Tabossi (1993), pages 271–.395.

Fellbaum, C., editor (1998). *WordNet, An Electronic Lexical Database*. MIT Press.

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

Fellbaum, C. (2005). The ontological loneliness of verb phrase idioms. In Schalley, A. and Zaefferer, D., editors, *Ontolinguistics*. Mouton de Gruyter. Forthcomming.

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

Glucksberg, S. (1993). Idiom meanings and allusional content. In Cacciari and Tabossi (1993), pages 3–.36.

Grefenstette, G. and Teufel, S. (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)*.

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

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

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

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

Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press.

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

Lin, T.-H. (2001). *Light Verb Syntax and the Theory of Phrase Structure*. PhD thesis, University of California, Irvine.

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

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

Melamed, I. D. (1997). 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.

Miyamoto, T. (2000). *The Light Verb Construction in Japanese: the Role of the Verbal Noun*. John Benjamins.

Moirón, M. B. V. (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.

Newman, J. (1996). *Give: A Cognitive Linguistic Study*. Mouton de Gruyter.

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

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

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

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (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.

Seaton, M. and Macaulay, A., editors (2002). *Collins COBUILD Idioms Dictionary*. HarperCollins Publishers, second edition.

Seretan, V., Nerima, L., and Wehrli, E. (2003). Extraction of multi-word collocations using syntactic bigram composition. In *Proceedings of the International Conference RANLP'03*, Bulgaria.

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

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

Stock, O., Slack, J., and Ortony, A. (1993). Building castles in the air. some computational and theoretical issues in idiom comprehension. In Cacciari and Tabossi (1993), pages 229–.347.

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

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

Villavicencio, A., Baldwin, T., and Waldron, B. (2004). A multilingual database of idioms. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 1127–30, Lisbon, Portugal.

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

Wermter, J. and Hahn, U. (2005). Paradigmatic modifiability statistics for the extraction of complex multi-word terms. In *Proceedings of HLT-EMNLP'05*, pages 843–850.

Widdows, D. and Dorow, B. (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.

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