Chapter 1

Multiword Expressions

1.1 Introduction

Languages are made up of words, which combine via morphosyntax to encode meaning in the form of phrases and sentences. While it may appear relatively innocuous, the question of what constitutes a "word" is a surprisingly vexed one. First, are *dog* and *dogs* two separate words, or variants of a single word? The traditional view from lexicography and linguistics is to treat them as separate inflected *wordforms* of the *lexeme* *dog*, as any difference in the syntax/semantics of the two words is predictable from the general process of noun pluralisation in English. Second, what is the status of expressions like *top dog* and *dog days*? A speaker of English who knew *top*, *dog* and *day* in isolation but had never been exposed to these two expressions would be hard put to predict the semantics of "person who is in charge" and "period of inactivity", respectively.[1] To be able to retrieve the semantics of these expressions, they must have lexical status of some form in the mental lexicon, which encodes their particular semantics. Expressions such as these which have surprising properties not predicted by their component words are referred to as *multiword expressions* (MWEs).[2] The focus of this chapter is the precise nature and types of MWEs, and the current state of MWE research in NLP.

Armed with our informal description of MWEs, let's first motivate this chapter with a brief overview of the range of MWEs, and complexities asso-

[1] All glosses in this paper are taken from WordNet 3.0 (Fellbaum 1998).

[2] Terms which are largely synonymous with "multiword expression" are "multiword unit", "multiword lexical item", "phraseological unit" and "fixed expression"; there is also variation in the hyphenation of "multiword", with "multi-word" in common use.

Baldwin, Timothy and Su Nam Kim (2010) Multiword Expressions, in Nitin Indurkhya and Fred J. Damerau (eds.) Handbook of Natural Language Processing, Second Edition, CRC Press, Boca Raton, USA, pp. 267-292.

ciated with them. We return to define MWEs formally in Section 1.2.

(1a)–(1b) include a number of MWEs, underlined.

(1) a. In a nutshell, the administrator can take advantage of the database's many features through a single interface.

b. You should also jot down the serial number of your television video.

As we can see, analogously to simple words, MWEs can occur in a wide range of lexical and syntactic configurations (e.g. nominal, verbal and adverbial). Semantically, we can observe different effects: in some cases (e.g. *serial number* and *television video*), the component words preserve their original semantics, but the MWE encodes extra semantics (e.g. the fact that a *television video* is a single-unit device, and usually designed to be portable); in other cases (e.g. *in a nutshell*, meaning "summed up briefly"), the semantics of one or more of the component words has no obvious bearing of the semantics of the MWE.

While all of the MWE examples we have seen to date have occurred as contiguous units, this is not always the case:

(2) a. She likes to take a long bath for relaxation after exams.

b. Kim hates to put her friends out.

For example, in (2a), *long* is an internal modifier and not a component of the base MWE *take a bath*, as there is nothing surprising about the syntax of the modified MWE or the resulting semantics (c.f. *take a short/leisurely/warm/mud/... bath*).

How big an issue are MWEs, though? The number of MWEs is estimated to be of the same order of magnitude as the number of simplex words in a speaker's lexicon (Jackendoff 1997; Tschichold 1998; Pauwels 2000). At the type level, therefore, MWEs are as much of an issue as simple words. Added to this, new (types of) MWE are continuously created as languages evolve (e.g. *shock and awe*, *carbon footprint*, *credit crunch*) (Gates 1988; Tschichold 1998; Fazly, Cook, and Stevenson 2009).

Crosslingually, MWEs have been documented across a broad spectrum of the world's languages (see the companion web site for this chapter for a detailed listing of references). In fact, MWEs are such an efficient way of providing nuance and facilitating lexical expansion with a relatively small simplex lexicon, it is highly doubtful that any language would evolve without MWEs of some description.

MWEs are broadly used to enhance fluency and understandability, or mark the register/genre of language use (Fillmore, Kay, and O'Connor 1988; Liberman and Sproat 1992; Nunberg, Sag, and Wasow 1994; Dirven 2001). For example, MWEs can make language more or less informal/colloquial (c.f. *London Underground* vs. *Tube*, and *piss off* vs. *annoy*). Regionally, MWEs vary considerably. For example, *take away* and *take out* are identical in meaning, but the former is the preferred expression in British/Australian English, while the latter is the preferred expression in American English. Other examples

are *phone box* vs. *phone booth*, *lay the table* vs. *set the table*, and *no through road* vs. *not a through street*, respectively.

There is a modest body of research on modelling MWEs which has been integrated into NLP applications, e.g. for the purposes of fluency, robustness or better understanding of natural language. One area where MWEs have traditionally been used heavily (either explicitly or implicitly) is machine translation, as a means of capturing subtle syntactic, semantic and pragmatic effects in the source and target languages (Miyazaki, Ikehara, and Yokoo 1993; Gerber and Yang 1997; Melamed 1997; Matsuo, Shirai, Yokoo, and Ikehara 1997). Understanding MWEs has broad utility in tasks ranging from syntactic disambiguation to conceptual (semantic) comprehension. Explicit lexicalised MWE data helps simplify the syntactic structure of sentences that include MWEs, and conversely, a lack of MWE lexical items in a precision grammar is a significant source of parse errors (Baldwin, Bender, Flickinger, Kim, and Oepen 2004). Additionally, it has been shown that accurate recognition of MWEs influences the accuracy of semantic tagging (Piao, Rayson, Archer, Wilson, and McEnery 2003), and word alignment in machine translation (MT) can be improved through a specific handling of the syntax and semantics of MWEs (Venkatapathy and Joshi 2006).

1.2 Linguistic Properties of MWEs

We adopt the following formal definition of *multiword expression*, following (Sag, Baldwin, Bond, Copestake, and Flickinger 2002):

(3) Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity

In languages such as English, the conventional interpretation of the requirement of decomposability into lexemes is that MWEs must in themselves be made up of multiple whitespace-delimited words. For example, *marketing manager* is potentially a MWE as it is made up of two lexemes (*marketing* and *manager*), while fused words such as *lighthouse* are conventionally not classified as MWEs.[3] In languages such as German, the high productivity of compound nouns such as *Kontaktlinse* "contact lens" (the concatenation of *Kontakt* "contact" and *Linse* "lens"), without whitespace delimitation, means that we tend to relax this restriction and allow for single-word MWEs. In non-segmenting languages such as Japanese and Chinese (Baldwin and Bond

[3] In practice, a significant subset of research on English noun compounds (see Section 1.3.1) has considered both fused and whitespace-separated expressions.

2002; Xu, Lu, and Li 2006), we are spared this artificial consideration. The ability to decompose an expression into multiple lexemes is still applicable, however, and leads to the conclusion, e.g. that *fukugō-hyōgen* "multiword expression" is a MWE (both *fukugō* "compound" and *hyōgen* "expression" are standalone lexemes), but *buchō* "department head" is *not* (*bu* "department" is a standalone lexeme, but *chō* "head" is not).

The second requirement on a MWE is for it to be idiomatic. We provide a detailed account of idiomaticity in its various manifestations in the following section.

1.2.1 Idiomaticity

In the context of MWEs, *idiomaticity* refers to markedness or deviation from the basic properties of the component lexemes, and applies at the lexical, syntactic, semantic, pragmatic, and/or statistical levels. A given MWE is often idiomatic at multiple levels (e.g. syntactic, semantic and statistical in the case of *by and large*), as we return to illustrate in Section 1.2.3.

Closely related to the notion of idiomaticity is *compositionality*, which we consider to be the degree to which the features of the parts of a MWE combine to predict the features of the whole. While compositionality is often construed as applying exclusively to semantic idiomatic (hence by "non-compositional MWE", researchers tend to mean a semantically-idiomatic MWE), in practice it can apply across all the same levels as idiomaticity. Below, we present an itemised account of each sub-type of idiomaticity.

1.2.1.1 Lexical Idiomaticity

Lexical idiomaticity occurs when one or more components of an MWE are not part of the conventional English lexicon. For example, *ad hoc* is lexically marked in that neither of its components (*ad* and *hoc*) are standalone English words.[4] Lexical idiomaticity inevitably results in syntactic and semantic idiomaticity because there is no lexical knowledge associated directly with the parts from which to predict the behaviour of the MWE. As such, it is one of the most clear-cut and predictive properties of MWEhood.

1.2.1.2 Syntactic Idiomaticity

Syntactic idiomaticity occurs when the syntax of the MWE is not derived directly from that of its components (Katz and Postal 2004; Chafe 1968; Bauer 1983; Sag, Baldwin, Bond, Copestake, and Flickinger 2002). For example, *by and large*, is syntactically idiomatic in that it is adverbial in nature, but made up of the anomalous coordination of a preposition (*by*) and an adjective

[4]Note that the idiomaticity is diminished if the speaker has knowledge of the Latin origins of the term. Also, while the component words don't have status as standalone lexical items, they do occur in other MWEs (e.g. *ad nauseum*, *post hoc*).

(*large*). On the other hand, *take a walk* is not syntactically marked as it is a simple verb–object combination which is derived transparently from a transitive verb (*walk*) and a countable noun (*walk*). Syntactic idiomaticity can also occur at the constructional level, in classes of MWEs having syntactic properties which are differentiated from their component words, e.g. verb particle constructions Section 1.3.2.1 and determinerless prepositional phrases Section 1.3.3.2.

1.2.1.3 Semantic Idiomaticity

Semantic idiomaticity is the property of the meaning of a MWE not being explicitly derivable from its parts (Katz and Postal 2004; Chafe 1968; Bauer 1983; Sag, Baldwin, Bond, Copestake, and Flickinger 2002). For example, *middle of the road* usually signifies "non-extremism, especially in political views", which we could not readily predict from either *middle* or *road*. On the other hand, *to and fro* is not semantically marked as its semantics is fully predictable from its parts. Many cases are not as clear cut as these, however. The semantics of *blow hot and cold* ("constantly change opinion"), for example, is partially predictable from *blow* ("move" and hence "change"), but not as immediately from *hot and cold*. There are also cases where the meanings of the parts are transparently inherited but there is additional semantic content which has no overt realisation. One such example is *bus driver* where, modulo the effects of word sense disambiguation, *bus* and *driver* both have their expected meanings, but there is additionally the default expectation that a *bus driver* is "one who drives a bus" and not "one who drives *like* a bus" or "an object for driving buses with", for example.

Closely related to the issue of semantic idiomaticity is the notion of *figuration*, i.e. the property of the components of a MWE having some metaphoric (e.g. *take the bull by the horns*), hyperbolic (e.g. *not worth the paper it's printed on*) or metonymic (e.g. *lend a hand*) meaning in addition to their literal meaning (Fillmore, Kay, and O'Connor 1988; Nunberg, Sag, and Wasow 1994). As an illustration of decomposability via metaphorical figuration, consider the English idiom *spill the beans*. Assuming a formal semantic representation of `reveal′(secret′)` for the MWE, we can coerce the semantics of *spill* and *beans* into `reveal′` and `secret′`, respectively, to arrive at a figurative interpretation of the MWE semantics. A compositionality analysis would not be able to predict this regularity as these senses for *spill* and *beans* are not readily available outside this particular MWE. Predictably, MWEs vary in the immediacy of their decomposability — with *get the nod* being more transparently decomposable than *spill the beans*, e.g. — and not all MWEs are decomposable (c.f. *kick the bucket*). We return to discuss the interaction between decomposability and syntactic flexibility in Section 1.3.2.4.

One intriguing aspect of semantic idiomaticity is that higher-usage MWEs are generally perceived to be less semantically idiomatic, or at least more readily decomposable (Keysar and Bly 1995).

1.2.1.4 Pragmatic Idiomaticity

Pragmatic idiomaticity is the condition of a MWE being associated with a fixed set of situations or a particular context (Kastovsky 1982; Jackendoff 1997; Sag, Baldwin, Bond, Copestake, and Flickinger 2002). *Good morning* and *all aboard* are examples of pragmatic MWEs: the first is a greeting associated specifically with mornings[5] and the second is a command associated with the specific situation of a train station or dock, and the imminent departure of a train or ship. Pragmatically idiomatic MWEs are often ambiguous with (non-situated) literal translations; e.g. *good morning* can mean "pleasant morning" (c.f. *Kim had a good morning*).

1.2.1.5 Statistical Idiomaticity

Statistical idiomaticity occurs when a particular combination of words occurs with markedly high frequency, relative to the component words or alternative phrasings of the same expression (Cruse 1986; Sag, Baldwin, Bond, Copestake, and Flickinger 2002). For example, in Table 1.1, we present an illustration of statistical idiomaticity, adapted from Cruse (1986, p281). The example is based on the cluster of near-synonym adjectives (*flawless*, *immaculate*, *impeccable* and *spotless*), and their affinity to pre-modify a range of nouns. For a given pairing of adjective and noun, we indicate the compatibility in the form of discrete markers ("+" indicates a positive lexical affinity, "?" indicates a neutral lexical affinity, and "−" indicates a negative lexical affinity). For example, *immaculate* has a strong lexical affinity with *performance* (i.e. *immaculate performance* is a relatively common expression), whereas *spotless* has a negative affinity with *credentials* (i.e. *spotless credentials* is relatively infrequent). There may, of course, be phonological, semantic or other grounds for particular adjective–noun combinations being more or less frequent; statistical idiomaticity is simply an observation of the relative frequency of a given combination. It is also important to note that statistical idiomaticity is a continuously-graded phenomenon, and our predictions about lexical affinity in Table 1.1 are most naturally interpreted as a ranking of the propensity for each of the adjectives to occur as a pre-modifier of *record*; for example, *impeccable* and *spotless* are more probable choices than *immaculate*, which is in turn more probable than *flawless*.

Another striking case of statistical idiomaticity is with binomials such as *black and white* — as in *black and white television* — where the reverse noun ordering does not preserve the lexicalised semantics of the word combination (c.f. *?white and black television*) (Benor and Levy 2006). The arbitrariness of the preferred noun order in English is poignantly illustrated by it being reversed in other languages, e.g. *shirokuro* "white and black" and *blanco y negro* "white and black" in Japanese and Spanish, respectively.

[5] Which is not to say that it can't be used ironically at other times of the day!

	flawless	immaculate	impeccable	spotless
condition	+	−	+	+
credentials	−	−	+	−
hair	−	+	?	−
house	?	+	?	+
logic	+	−	+	−
timing	?	+	+	−

Table 1.1: Examples of statistical idiomaticity ("+" = strong lexical affinity, "?" = neutral lexical affinity, "−" = negative lexical affinity)

Statistical idiomaticity relates closely to the notion of *institutionalisation* (a.k.a. *conventionalisation*), i.e. a particular word combination coming to be used to refer a given object (Fernando and Flavell 1981; Bauer 1983; Nunberg, Sag, and Wasow 1994; Sag, Baldwin, Bond, Copestake, and Flickinger 2002). For example, *traffic light* is the conventionalised descriptor for "a visual signal to control the flow of traffic at intersections". There is no *a priori* reason why it shouldn't instead be called a *traffic director* or *intersection regulator*, but the simple matter of the fact is that it is not referred to using either of those expressions; instead, *traffic light* was historically established as the canonical term for referring to the object. Similarly, it is an arbitrary fact of the English language that we say *many thanks* and not **several thanks*, and *salt and pepper* in preference to *pepper and salt*.[6] We term these *anti-collocations* of the respective MWEs (Pearce 2001): lexico-syntactic variants of MWEs which have unexpectedly low frequency, and in doing so, contrastively highlight the statistical idiomaticity of the target expression.[7]

1.2.2 Other Properties of MWEs

Other common properties of MWE are: single-word paraphrasability, proverbiality and prosody. Unlike idiomaticity, where some form of idiomaticity is a necessary feature of MWEs, these other properties are neither necessary nor sufficient. Prosody relates to semantic idiomaticity, while the other properties are independent of idiomaticity as described above.

- Crosslingual variation

[6]Which is not to say there wasn't grounds for the selection of the canonical form at its genesis, e.g. for historical, crosslingual or phonological reasons.

[7]The term anti-collocation originated in the context of collocation research (see Section 1.2.4). While noting the potential for confusion, we use it in the broader context of MWEs as a tool for analysing the statistical idiomaticity of a candidate MWE relative to alternative forms of the same basic expression.

There is remarkable variation in MWEs across languages (Villavicencio, Baldwin, and Waldron 2004). In some cases, there is direct lexico-syntactic correspondence for a crosslingual MWE pair with similar semantics. For example, *in the red* has a direct lexico-syntactic correlate in Portuguese with the same semantics: *no vermelho*, where *no* is the contraction of *in* and *the*, *vermelho* means *red*, and both idioms are prepositional phrases (PPs). Others have identical syntax but differ lexically. For example, *in the black* corresponds to *no azul* ("in the blue") in Portuguese, with a different choice of colour term (*blue* instead of *black*). More obtusely, *Bring the curtain down on* corresponds to the Portuguese *botar um ponto final em* (lit. "put the final dot in"), with similar syntactic make-up but radically different lexical composition. Other MWEs again are lexically similar but syntactically differentiated. For example, *in a corner* (e.g. *The media has him in a corner*) and *encurralado* ("cornered") are semantically equivalent but realised by different constructions – a PP in English and an adjective in Portuguese.

There are of course many MWEs which have no direct translation equivalent in a second language. For example, the Japanese MWE *zoku-giiN*, meaning "legistors championing the causes of selected industries" has no direct translation in English (Tanaka and Baldwin 2003). Equally, there are terms which are realised as MWEs in one language but single-word lexemes in another, such as *interest rate* and its Japanese equivalent *ririsu*.

- Single-word paraphrasability

  Single-word paraphrasability is the observation that significant numbers of MWEs can be paraphrased with a single word (Chafe 1968; Gibbs 1980; Fillmore, Kay, and O'Connor 1988; Liberman and Sproat 1992; Nunberg, Sag, and Wasow 1994). While some MWEs are single-word paraphrasable (e.g. *leave out* = *omit*), others are not (e.g. *look up* = *?*). Also, MWEs with arguments can sometimes be paraphrasable (e.g. *take off clothes* = *undress*), just as non-MWEs comprised of multiple words can be single-word paraphrasable (e.g. *drop sharply* = *plummet*).

- Proverbiality

  Proverbiality is the ability of a MWE to "describe and implicitly to explain a recurrent situation of particular social interest in the virtue of its resemblance or relation to a scenario involving homely, concrete things and relations" (Nunberg, Sag, and Wasow 1994). For example, verb particle constructions and idioms are often indicators of more informal situations (e.g. *piss off* is an informal form of *annoy*, and *drop off* is an informal form of *fall asleep*).

- Prosody

	Lexical	Syntactic	Semantic	Pragmatic	Statistical
all aboard	−	−	−	+	+
bus driver	−	−	+	−	+
by and large	−	+	+	−	+
kick the bucket	−	−	+	−	+
look up	−	−	+	−	+
shock and awe	−	−	−	+	+
social butterfly	−	−	+	−	+
take a walk	−	−	+	−	?
to and fro	?	+	−	−	+
traffic light	−	−	+	−	+
eat chocolate	−	−	−	−	−

Table 1.2: Classification of MWEs in terms of their idiomaticity

MWEs can have distinct *prosody*, i.e. stress patterns, from compositional language (Fillmore, Kay, and O'Connor 1988; Liberman and Sproat 1992; Nunberg, Sag, and Wasow 1994). For example, when the components do not make an equal contribution to the semantics of the whole, MWEs can be prosodically marked, e.g. *soft spot* is prosodically marked (due to the stress on *soft* rather than *spot*), although *first aid* and *red herring* are not. Note that *prosodic* marking can equally occur with non-MWEs, such as *dental operation*.

1.2.3 Testing an Expression for MWEhood

Above, we described five different forms of idiomaticity, along with a number of other properties of MWEs. We bring these together in categorising a selection of MWEs in Table 1.2.

Taking the example of the verb particle construction *look up* (in the sense of "seek information from", as in *Kim looked the word up in the dictionary*), we first observe that it is made up of multiple words (*look* and *up*), and thus satisfies the first requirement in our MWE definition. In terms of idiomaticity: (1) it is not lexically idiomatic, as both *look* and *up* are part of the standard English lexicon; (2) while it has peculiar syntax relative to its component words, in *up* being separable from *look*, this is a general property of transitive verb particle constructions (see Section 1.3.2.1) rather than this particular word combination, so it is not syntactically idiomatic; (3) it is semantically idiomatic, as the semantics of "seek information from" is not predictable from the standard semantics of *look* and *up*; (4) it is not pragmatically idiomatic, as it doesn't generally evoke a particular situation; and (5) it is statistically

marked, as it contrasts with anti-collocations such as **see/watch up*[8] and is a relatively frequent expression in English. That is, it is semantically and statistically idiomatic; in combination with its multiword composition, this is sufficient to classify it as a MWE.

In Table 1.2, *kick the bucket* (in the sense of "die") has only one form of idiomaticity (semantic), while all the other examples have at least two forms of idiomaticity. *Traffic light*, for example, is statistically idiomatic in that it is both a common expression in English and stands in opposition to anti-collocations such as **vehicle light/traffic lamp*, and it is semantically idiomatic in that the particular semantics of "a visual signal to control the flow of traffic" is not explicitly represented in the component words (e.g. interpretations such as "a visual signal to indicate the flow of traffic", "a device for lighting the way of traffic" or "a lamp which indicates the relative flow of data" which are predicted by the component words are not readily available). Other noteworthy claims about idiomaticity are: *shock and awe* is pragmatically idiomatic because of its particular association with the commencement of the Iraq War in 2003; *take a walk* is semantically idiomatic because this sense of *take* is particular to this and other *light verb constructions* (see Section 1.3.2.3), and distinct from the literal sense of the verb; and *to and fro* is syntactically idiomatic because of the relative syntactic opacity of the antiquated *fro*, and (somewhat) lexically idiomatic as it is used almost exclusively in the context of *to and fro*.[9]

Table 1.2 includes one negative example: *eat chocolate*. While it satisfies the requirement for multiword decomposability (i.e. it is made up of more than one word), it clearly lacks lexical, syntactic, semantic and pragmatic idiomaticity. We would claim that it is also not statistically idiomatic. One possible counter-argument could be that *eat* is one of the most common verbs associated with *chocolate*, but the same argument could be made for almost any foodstuff in combination with *eat*. Possible anti-collocations such as *consume chocolate* or *munch on chocolate* are also perfectly acceptable.

1.2.4 Collocations and MWEs

A common term in NLP which relates closely to our discussion of MWEs is *collocation*. A widely-used definition for collocation is "an arbitrary and recurrent word combination" (Benson 1990), or in our terms, a statistically idiomatic MWE (esp. of high frequency). While there is considerable varia-

[8] Under the constraint that *up* is a particle; examples such as *see you up the road* occur readily, but are not considered to be anti-collocations as *up* is a (transitive) preposition.

[9] Words such as this which occur only as part of a fixed expression are known variously as *cranberry words* or *bound words* (Aronoff 1976; Moon 1998; Trawiński, Sailer, Soehn, Lemnitzer, and Richter 2008) (other examples are *tenterhooks* and *caboodle*), and the expressions that contain them are often termed *cranberry expressions* (e.g. *on tenterhooks* and *the whole caboodle*).

tion between individual researchers, collocations are often distinguished from "idioms" or "non-compositional phrases" on the grounds that they are not syntactically idiomatic, and if they are semantically idiomatic, it is through a relatively transparent process of figuration or metaphor (Choueka 1988; Lin 1998; McKeown and Radev 2000; Evert 2004). Additionally, much work on collocations focuses exclusively on predetermined constructional templates (e.g. adjective–noun or verb–noun collocations). In Table 1.2, e.g. *social butterfly* is an uncontroversial instance of a collocation, but *look up* and *to and fro* would tend not to be classified as collocations. As such, collocations form a proper subset of MWEs.

1.2.5 A Word on Terminology and Related Fields

It is worth making mention of a number of terms which relate to MWEs.

The term *idiom* varies considerably in its usage, from any kind of multiword item to only those MWEs which are semantically idiomatic; even here, there are those who consider idioms to be MWEs which are *exclusively* semantically idiomatic (also sometimes termed *pure idioms*), and those who restrict the term to particular syntactic sub-types of semantically idiomatic MWEs (Fillmore, Kay, and O'Connor 1988; Nunberg, Sag, and Wasow 1994; Moon 1998; Huddleston and Pullum 2002). To avoid confusion, we will avoid using this term in this chapter.

The field of terminology has a rich history of research on multiword terms, which relates closely to MWEs (Sager 1990; Justeson and Katz 1995; Frantzi, Ananiadou, and Mima 2000; Kageura, Daille, Nakagawa, and Chien 2004). The major difference is that terminology research is primarily interested in identifying and classifying technical terms specific to a particular domain (both MWE and simplex lexemes) and predicting patterns of variation in those terms. It is thus broader in scope than MWEs in the sense that simple lexemes can equally be technical terms, and narrower in the sense than non-technical MWEs are not of interest to the field.

Phraseology is another field with a rich tradition history relating to MWEs (Cowie and Howarth 1996; Cowie 2001). It originally grew out of the work of Mel'čuk and others in Russia on Meaning-Text Theory (Mel'čuk and Polguère 1987), but more recently has taken on elements from the work of Sinclair and others in the context of corpus linguistics and corpus-based lexicography (Sinclair 1991). Phraseology is primarily interested in the description and functional classification of MWEs (including "sentence-like" units, such as phrases and quotations), from a theoretical perspective.

1.3 Types of MWE

In this section, we detail a selection of the major MWE types which have received particular attention in the MWE literature. We will tend to focus on English MWEs for expository purposes, but provide tie-ins to corresponding MWEs in other languages where possible.

1.3.1 Nominal MWEs

Nominal MWEs are one of the most common MWE types, in terms of token frequency, type frequency, and their occurrence in the world's languages (Tanaka and Baldwin 2003; Lieber and Štekauer 2009). In English, the primary type of nominal MWE is the *noun compound* (NC), where two or more nouns combine to form a $\bar{\text{N}}$, such as *golf club* or *computer science department* (Lauer 1995; Sag, Baldwin, Bond, Copestake, and Flickinger 2002; Huddleston and Pullum 2002); the rightmost noun in the NC is termed the *head noun* (i.e. *club* and *department*, respectively) and the remainder of the component(s) *modifier(s)* (i.e. *golf* and *computer science*, respectively).[10] Within NCs, there is the subset of *compound nominalisations*, where the head is deverbal (e.g. *investor hesitation* or *stress avoidance*). There is also the broader class of nominal MWEs where the modifiers aren't restricted to be nominal, but can also be verbs (usually present or past participles, such as *connecting flight* or *hired help*) or adjectives (e.g. *open secret*). To avoid confusion, we will term this broader set of nominal MWEs *nominal compounds*. In Romance languages such as Italian, there is the additional class of *complex nominals* which include a preposition or other marker between the nouns, such as *succo di limone* "lemon juice" and *porta a vetri* "glass door".[11]

One property of noun compounds which has put them in the spotlight of NLP research is their underspecified semantics. For example, while sharing the same head, there is little semantic commonality between *nut tree*, *clothes tree* and *family tree*: a *nut tree* is a tree which bears edible nuts; a *clothes tree* is a piece of furniture shaped somewhat like a tree, for hanging clothes on; and a *family tree* is a graphical depiction of the genealogical history of a family (which can be shaped like a tree). In each case, the meaning of the compound relates (if at times obtusely!) to a sense of both the head and the modifier, but the precise relationship is highly varied and not represented

[10] In fact, the norm amongst Germanic languages (e.g. Danish, Dutch, German, Norwegian and Swedish) is for noun compounds to be realised as a single compound word (Bauer 2001). *Solar cell*, for example, is *zonnecel* in Dutch, *Solarzelle* in German, and *solcell* in Swedish. See Section 1.2 for comments on their compatibility with our definition of MWE.

[11] Our use of the term *complex nominal* for MWEs of form N P N should not be confused with that of Levi (1978), which included NCs and nominal compounds.

explicitly in any way. Furthermore, while it may be possible to argue that these are all lexicalised noun compounds with explicit semantic representations in the mental lexicon, native speakers generally have reasonably sharp intuitions about the semantics of novel compounds. For example, a *bed tree* is most plausibly a tree that beds are made from or perhaps for sleeping in, and a *reflection tree* could be a tree for reflecting in/near or perhaps the reflected image of a tree. Similarly, context can evoke irregular interpretations of high-frequency compounds (Downing 1977; Spärck Jones 1983; Copestake and Lascarides 1997; Gagné, Spalding, and Gorrie 2005). This suggests that there is a dynamic interpretation process that takes place, which complements encyclopedic information about lexicalised compounds.

One popular approach to capturing the semantics of compound nouns is via a finite set of relations. For example, *orange juice*, *steel bridge* and *paper hat* could all be analysed as belonging to the MAKE relation, where HEAD is made from MODIFIER. This observation has led to the development of a bewildering range of semantic relation sets of varying sizes, based on abstract relations (Vanderwende 1994; Barker and Szpakowicz 1998; Rosario and Hearst 2001; Moldovan, Badulescu, Tatu, Antohe, and Girju 2004; Nastase, Sayyad-Shirabad, Sokolova, and Szpakowicz 2006), direct paraphrases, e.g. using prepositions or verbs (Lauer 1995; Lapata 2002; Grover, Lapata, and Lascarides 2004; Nakov 2008), or various hybrids of the two (Levi 1978; Vanderwende 1994; Ó Séaghdha 2008). This style of approach has been hampered by issues including low inter-annotator agreement (especially for larger semantic relation sets), coverage over data from different domains, the impact of context on interpretation, how to deal with "fringe" instances which don't quite fit any of the relations, and how to deal with interpretational ambiguity (Downing 1977; Spärck Jones 1983; Ó Séaghdha 2008).

An additional area of interest with nominal MWEs (especially noun compounds) is the syntactic disambiguation of MWEs with 3 or more terms. For example, *glass window cleaner* can be syntactically analysed as either *(glass (window cleaner))* (i.e. "a window cleaner made of glass", or similar) or *((glass window) cleaner)* (i.e. "a cleaner of glass windows"). Syntactic ambiguity impacts on both the semantic interpretation and prosody of the MWE. The task of disambiguating syntactic ambiguity in nominal MWEs is called *bracketing*. We return to discuss the basic approaches to bracketing in Section 1.5.3.

1.3.2 Verbal MWEs

1.3.2.1 Verb-particle constructions

Verb-particle constructions (*VPCs*, also sometimes termed *particle verbs* or *phrasal verbs*) are made up of a verb and an obligatory particle, typically in the form of an intransitive preposition (e.g. *play around*, *take off*), but including adjectives (e.g. *cut short*, *band together*) and verbs (e.g. *let go*, *let fly*) (Bolinger 1976; Jackendoff 1997; Huddleston and Pullum 2002; McIntyre

2007). English VPCs relate closely to *particle verbs* (a.k.a. *separable verbs*) in languages such as German (Lüdeling 2001), Dutch (Booij 2002) and Estonian (Kaalep and Muischnek 2008), but the construction has its own peculiarities in each language which go beyond the bounds of this chapter. To avoid confusion, we will focus exclusively on English VPCs in our discussion here.

The distinguishing properties of English VPCs are:

- Transitive VPCs can occur in either the *joined* (e.g. *Kim put on the sweater*) or *split* (e.g. *Kim put the sweater on*) word order in the case that the object NP is not pronominal
- Transitive VPCs must occur in the split word order if the object NP is pronominal (e.g. *Kim polished it off* vs. **Kim polished off it*).
- Manner adverbs do not readily occur between the verb and particle, in both intransitive and transitive VPCs (e.g. ?**Kim played habitually around*, **Kim made quickly up her mind*). Note, there is a small set of degree adverbs that readily premodify particles, notably *right* (e.g. *My turn is coming right up*) and *back* (e.g. *Kim put the sweater back on*)

All of these properties are defined at the construction level and common to all VPCs, however, begging the question of where the idiomaticity comes in that allows us to define them as MWEs. The answer is, in the main, semantic and statistical idiosyncrasy. For example, the semantics of *polish* in *polish off* (e.g. *polish off dessert*, *polish off the hitman*, *polish off my homework*) is differentiated from that of the simplex lexeme. Conversely, *swallow down* (e.g. *swallow down the drink*) preserves the semantics of both *swallow* and *down* (i.e. the liquid is swallowed, and as a result goes down [the oesophagus]), and is thus conventionally not considered be a MWE.

VPCs are highly frequent in English text, but the distribution is highly skewed towards a minority of the VPC types, with the majority of VPCs occurring very infrequently (Baldwin 2005a). This is bad news if we want to build a parser with full coverage, e.g., as we need to capture the long tail of VPC types. Compounding the problem, the construction is highly productive. For example, the completive *up* (e.g. *eat/finish/rest/... up*) can combine productively with a large array of action verbs to form a VPC with predictable syntax and semantics, which we could never hope to exhaustively list. Having said this, there are large numbers of semantically-idiomatic VPCs which need to be recorded in the lexicon if we wish to capture their semantics correctly. Even here, VPCs populate the spectrum of compositionality relative to their components (Lidner 1983; Brinton 1985; Jackendoff 2002; Bannard, Baldwin, and Lascarides 2003; McCarthy, Keller, and Carroll 2003; Cook and Stevenson 2006), so while some VPCs are clear candidates for lexicalisation in terms of their semantic idiomaticity (e.g. *make out*, as in *Kim made out the cheque to Sandy* or *Kim and Sandy made out*), others are semantically closer to the semantics of their component words (e.g. *check out*, *blow over*)

and to some degree derivable from their component words. One approach to representing this continuum of VPC semantics is that of Bannard, Baldwin, and Lascarides (2003), who subclassify VPCs into four compositionality classes based on the independent semantic contribution of the verb and particle: (1) the VPC inherits its semantics from the verb and particle (i.e. is not semantically idiomatic); (2) the VPC inherits semantics from the verb only; (3) the VPC inherits semantics from the particle only; and (4) the VPC inherits semantics from neither the verb nor the particle. A second approach is to employ a one-dimensional classification of holistic VPC compositionality (e.g. in the form of a integer scale of 0 to 10 (McCarthy, Keller, and Carroll 2003)).

1.3.2.2 Prepositional verbs

Prepositional verbs (PVs) relate closely to VPCs in being comprised of a verb and selected preposition, with the crucial difference that the preposition is transitive (e.g. *refer to*, *look for*) (Jackendoff 1973; O'Dowd 1998; Huddleston and Pullum 2002; Baldwin 2005b; Osswald, Helbig, and Hartrumpf 2006). English PVs occur in two basic forms: (1) *fixed preposition PVs* (e.g. *come across*, *grow on*), where there is a hard constraint of the verb and selected preposition being strictly adjacent; and (2) *mobile preposition PVs* (e.g. *refer to*, *send for*), where the selected preposition is adjacent to the verb in the canonical word order, but undergoes limited syntactic alternation. For example, mobile preposition PVs allow limited coordination of PP objects (e.g. *refer to the book and to the DVD* vs. **come across the book and across the DVD*), and the NP object of the selected preposition can be passivised (e.g. *the book was referred to* vs. **I was grown on by the book*).

PVs are highly frequent in general text, and notoriously hard to distinguish from VPCs and simple verb–preposition combinations, e.g. in parsing applications.

1.3.2.3 Light-Verb Constructions

Light-verb constructions (i.e. LVCs) are made up of a verb and a noun complement, often in the indefinite singular form (Jespersen 1965; Abeillé 1988; Miyagawa 1989; Grefenstette and Tapanainen 1994; Hoshi 1994; Sag, Baldwin, Bond, Copestake, and Flickinger 2002; Huddleston and Pullum 2002; Butt 2003; Stevenson, Fazly, and North 2004). The name of the construction comes from the verb being semantically bleached or "light", in the sense that their contribution to the meaning of the LVC is relatively small in comparison with that of the noun complement. In fact, the contribution of the light verb is so slight that in many cases, the LVC can be paraphrased with the verbal form of the noun complement (e.g. *take a walk* vs. *walk* or *take a photograph* vs. *photograph*). LVCs are also sometimes termed *verb-complement pairs* (Tan, Kan, and Cui 2006) or *support verb constructions* (Calzolari, Fillmore, Grishman, Ide, Lenci, MacLeod, and Zampolli 2002).

The following are the principle light verbs in English:

- *do*, e.g. *do a demo*, *do a drawing*, *do a report*
- *give*, e.g. *give a wave*, *give a sigh*, *give a kiss*
- *have*, e.g. *have a rest*, *have a drink*, *have pity (on)*
- *make*, e.g. *make an offer*, *make an attempt*, *make a mistake*
- *take*, e.g. *take a walk*, *take a bath*, *take a photograph*

There is some disagreement in the scope of the term LVC, most notably in the membership of verbs which can be considered "light". Calzolari, Fillmore, Grishman, Ide, Lenci, MacLeod, and Zampolli (2002), e.g., argued that the definition of LVCs (or support verb constructions in their terms) should be extended to include: (1) verbs that combine with an event noun (deverbal or otherwise) where the subject is a participant in the event most closely identified with the noun (e.g. *ask a question*); and (2) verbs with subjects that belong to some scenario associated with the full understanding of the event type designated by the object noun (e.g. keep a promise).

Morphologically, the verb in LVCs inflects but the noun complement tends to have fixed number and a preference for determiner type. For example, *make amends* undergoes full verbal inflection (*make/makes/made/making amends*), but the noun complement cannot be singular (e.g. **make amend*.[12] Syntactically, LVCs are highly flexible, undergoing passivization (e.g. *an offer was made*), extraction (e.g. *How many offers did Kim make?*) and internal modification (e.g. *make an irresistible offer*). On the other hand, there are hard constraints on what light verbs a given noun complement can be combined with (c.f. **give/do/put/take an offer*), noting that some noun complements combine with multiple light verbs (e.g. *do/give a demo*), often with different semantics (e.g. *make a call* vs. *take a call* vs. *have a call*). Also, what light verb a given noun will combine with to form an LVC is often consistent across semantically-related noun clusters (e.g. *give a cry/moan/howl* vs. **take a cry/moan/howl*.

LVCs occur across a large number of the world's languages, including Japanese (Grimshaw and Mester 1988; Baldwin and Bond 2002), Korean (Ahn 1991), Hindi (Mohanan 1994) and Persian (Karimi-Doostan 1997).

1.3.2.4 Verb–Noun Idiomatic Combinations

Verb–Noun Idiomatic Combinations (VNICs, also known as *VP idioms*) are composed of a verb and noun in direct object position, and are (at

[12] But also note other examples where the noun complement can be either singular or plural, e.g. *take a bath* vs. *take baths*.

least) semantically idiomatic (e.g. *kick the bucket*, *shoot the breeze*) (Nunberg, Sag, and Wasow 1994; Fellbaum 2002; Sag, Baldwin, Bond, Copestake, and Flickinger 2002; Fazly, Cook, and Stevenson 2009). They are a notable subclass of MWE because of their crosslingual occurrence, and high lexical and semantic variability.

VNICs (along with other semantically idiomatic MWEs) are often categorised into two groups, based on their semantic decomposability (see Section 1.2.1.3) (Nunberg, Sag, and Wasow 1994; Riehemann 2001). With *decomposable VNICs*, given the interpretation of the VNIC, it is possible to associate components of the VNIC with distinct elements of the VNIC interpretation, based on semantics not immediately accessible from the component lexemes. Assuming an interpretation of *spill the beans* such as `reveal'(`x`,secret')`,[13] e.g., we could analyse *spill* as having the semantics of `reveal'` and *beans* having the semantics of `secret'`, through a process of figuration. Other examples of decomposable VNICs are *pull strings* (c.f. `exert'(`x`,influence')`) and *touch a nerve* (c.f. `cause'(`x`,reaction')`). With *non-decomposable VNICs* (e.g. *get the hang (of)*, *kick the bucket*), such a semantic decomposition is not possible. The reason we make this distinction is that decomposable VNICs tend to be syntactically flexible, in a manner predicted by the nature of the semantic decomposition; non-decomposable VNICs, on the other hand, tend not to be syntactically flexible (Cruse 1986; Nunberg, Sag, and Wasow 1994; Jackendoff 1997; Sag, Baldwin, Bond, Copestake, and Flickinger 2002). For example, *spill the beans* can be passivised (*It's a shame the beans were spilled*) and internally modified (*AT&T spilled the Starbucks beans*), similarly to a conventional verb–direct object pair (c.f. *Sandy is loved by Kim* and *Kim loves the inimitable Sandy*); this is predicted by its decomposability.

VNICs generally occur with low frequency, but are notoriously hard to distinguish from literal usages of the same word combination (e.g. *Kim made a face at the policeman* vs. *Kim made a face in pottery class*). An accurate means of disambiguation is thus important in tasks which require semantic interpretation, but generally fraught by low volumes of training data.

1.3.3 Prepositional MWEs

1.3.3.1 Determinerless-Prepositional Phrases

Determinerless prepositional phrases (PP-Ds) are MWEs that are made up of a preposition and a singular noun without a determiner (Quirk, Greenbaum, Leech, and Svartvik 1985; Huddleston and Pullum 2002; Sag, Baldwin, Bond, Copestake, and Flickinger 2002; Baldwin, Beavers, Van Der Beek, Bond, Flickinger, and Sag 2006).

Syntactically, PP-Ds are highly diverse, and display differing levels of syntactic markedness, productivity and modifiability (Chander 1998; Ross 1995).

[13] I.e., `reveal'` is a 2-place predicate, with x binding to the subject.

That is, some PP-Ds are non-productive (e.g. *on top* vs. **on bottom*) and non-modifiable (e.g. *on top* vs. **on table top*), whereas others are fully-productive (e.g. *by car/foot/bus/...*) and highly modifiable (e.g. *at high expense*, *on summer vacation*). In fact, while some PP-Ds are optionally modifiable (e.g. *on vacation* vs. *on summer vacation*), others require modification (e.g. **at level* vs. *at eye level*, and **at expense* vs. *at company expense*) (Baldwin, Beavers, Van Der Beek, Bond, Flickinger, and Sag 2006).

Syntactically-marked PP-Ds can be highly productive (Ross 1995; Grishman, Macleod, and Myers 1998). For example, *by* combines with a virtually unrestricted array of countable nouns (e.g. *by bus/car/taxi/...*) but less readily with uncountable nouns (e.g. **by information/linguistics/...*).

Semantically, PP-Ds have a certain degree of semantic markedness on the noun (Haspelmath 1997; Mimmelmann 1998; Stvan 1998; Bond 2005). For example, *in* combines with uncountable nouns which refer to a social institution (e.g. *school, church, prison* but not *information*) to form syntactically-*un*marked PP-Ds with marked semantics, in the sense that only the social institution sense of the noun is evoked (e.g. *in school/church/prison/...* vs. **in information*) (Baldwin, Beavers, Van Der Beek, Bond, Flickinger, and Sag 2006).

PP-Ds occur with surprising frequency and cause problems during parsing and generation, in terms of achieving the right balance between over- and under-generation (Baldwin, Bender, Flickinger, Kim, and Oepen 2004).

1.3.3.2 Complex prepositions

Another common form of prepositional MWE is complex prepositions (e.g. *on top of*, *in addition to*), and other forms of complex markers (Villada Moirón 2005; Tsuchiya, Shime, Takagi, Utsuro, Uchimoto, Matsuyoshi, Sato, and Nakagawa 2006; Trawiński, Sailer, and Soehn 2006). Complex prepositions can take the form of fixed MWEs (e.g. *in addition to*), or alternatively semi-fixed MWEs, for example optionally allowing internal modification (e.g. *with (due/particular/special/...) regard to*) or determiner insertion (e.g. *on (the) top of*).

1.4 MWE Classification

In developing a lexicon of MWEs, it is crucially important to develop a classification which captures the general properties of MWE classes, but at the same time allows for the encoding of information particular to a given MWE instance. In this section, we present a commonly-used high-level classification, based particularly on the syntactic and semantic properties of MWEs outlined in Figure 1.1 (Bauer 1983; Sag, Baldwin, Bond, Copestake, and Flickinger

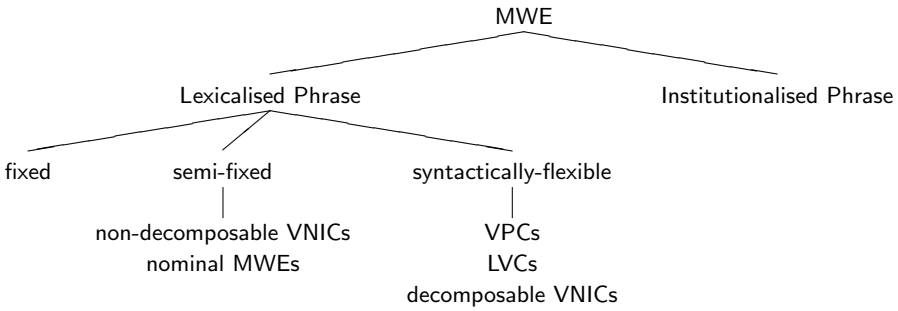

FIGURE 1.1: A classification of MWEs

2002).

The classification of MWEs into *lexicalised phrases* and *institutionalised phrases* hinges on whether the MWE is lexicalised (i.e. explicitly encoded in the lexicon), or a simple collocation (i.e. only statistically idiomatic).

Lexicalised phrases are MWEs with lexical, syntactic, semantic or pragmatic idiomaticity. Lexicalised phrases can be further split into: *fixed expressions* (e.g. *ad hoc*, *at first*), *semi-fixed expressions* (e.g. *spill the beans*, *car dealer*, *Chicago White Socks*) and *syntactically-flexible expressions* (e.g. *add up*, *give a demo*).

- *fixed expressions* are fixed strings that undergo neither morphosyntactic variation nor internal modification, often due to fossilisation of what was once a compositional phrase. For example, *by and large* is not morphosyntactically modifiable (e.g. **by and larger*) or internally modifiable (e.g. **by and very large*). Non-modifiable determinerless prepositional phrases such as *on air* are also fixed expressions.

- *semi-fixed expressions* are lexically-variable MWEs that have hard restrictions on word order and composition, but undergo some degree of lexical variation such as inflection (e.g. *kick/kicks/kicked/kicking the bucket* vs. **the bucket was kicked*), variation in reflexive pronouns (e.g. *in her/his/their shoes*) and determiner selection (e.g. *The Beatles vs. a Beatles album*[14]). Non-decomposable VNICs (e.g. *kick the bucket*, *shoot the breeze*) and nominal MWEs (e.g. *attorney general*, *part of speech*) are also classified as semi-fixed expressions.

- *syntactically flexible expressions* are MWEs which undergo syntactic variation, such as VPCs, LVCs and decomposable VNICs. The nature of

[14]The determiner *the* in *The Beatles* is obligatory in the case that *The Beatles* forms a noun phrase (i.e. *Beatles* can only be quantified by *the*), but in cases where *Beatles* forms a $\bar{\text{N}}$, e.g. in $[_{NP}\ a\ [_{N'}[_{N'}\ Beatles'\ ]\ album\ ]]$, the lexical item is realized without a determiner.

the flexibility varies significantly across construction types. VPCs, for example, are syntactically flexible with respect to the word order of the particle and NP in transitive usages: *hand in the paper* vs. *hand the paper in*. They are also usually compatible with internal modification, even for intransitive VPCs: *the plane took right off*. LVCs (e.g. *give a demo*) undergo full syntactic variation, including passivisation (e.g. *a demo was given*), extraction (e.g. *how many demos did he give?*) and internal modification (e.g. *give a clear demo*). Decomposable VNICs are also syntactically flexible to some degree, although the exact form of syntactic variation is predicted by the nature of their semantic decomposability.

Note that many of our MWE construction types can be assigned to a unique sub-category of lexicalised phrase, namely: non-decomposable VNICs, NCs, VPCs and LVCs. Determinerless PPs, on the other hand, cut across all three sub-categories: non-modifiable PP-Ds (e.g. *at first*) are fixed expressions, PP-Ds with strict constraints on modifiability (e.g. *at level*) are semi-fixed expressions, and highly-productive PP-Ds (e.g. *as president/coach/father of the bride/...*) are syntactically-flexible.

The class of *institutionalised phrases* corresponds to MWEs which are exclusively statistically idiomatic, as described in Section 1.2.4. Examples include *salt and pepper* and *many thanks*.

1.5 Research Issues

The major NLP tasks relating to MWEs are: (1) identifying and extracting MWEs from corpus data, and disambiguating their internal syntax, and (2) interpreting MWEs. Increasingly, these tasks are being pipelined with parsers and applications such as machine translation (Venkatapathy and Joshi 2006; Zhang, Kordoni, Villavicencio, and Idiart 2006; Blunsom 2007).

Depending on the type of MWE, the relative import of these syntactic and semantic tasks varies. For example, with noun compounds, the identification and extraction tasks are relatively trivial, whereas interpretation is considerably more difficult. Below, we discuss the challenges and review the key research on MWEs in NLP. For a listing of relevant resources (especially datasets and toolkits), we refer the reader to the companion web site (`http://...`).

1.5.1 Identification

Identification is the task of determining individual occurrences of MWEs in running text. The task is at the token (instance) level, such that we may

identify 50 distinct occurrences of *pick up* in a given corpus. To give an example of an identification task, given the corpus fragment in (4) (taken from "The Frog Prince", a children's story), we might identify the MWEs in (4):

(4) One fine evening a young princess put on her bonnet and clogs, and went out to take a walk by herself in a wood; ... she ran to pick it up; ...

In MWE identification, a key challenge is in differentiating between MWEs and literal usages for word combinations such as *make a face* which can occur in both usages (*Kim made a face at the policeman* [MWE] vs. *Kim made a face in pottery class* [non-MWE]). Syntactic ambiguity is also a major confounding factor, e.g. in identifying VPCs in contexts such as *Have the paper in today*. For example, in the sentence *Kim signed in the room*, there is ambiguity between a VPC interpretation (*sign in* = "check in/announce arrival") and an intransitive verb + PP interpretation ("Kim performed the act of signing in the room").

MWE identification has tended to take the form of customised methods for particular MWE construction types and languages (e.g. English VPCs, LVCs and NVICs), but there have been attempts to develop generalised techniques, as outlined below.

Perhaps the most obvious method of identifying MWEs is via a part-of-speech (POS) tagger, chunker or parser, in the case that lexical information required to identify MWEs is contained within the parser output. For example, in the case of VPCs, there is a dedicated tag for (prepositional) particles in the Penn POS tagset, such that VPC identification can be performed simply by POS tagging a text, identifying all particle tags, and further identifying the head verb associated with each particle (e.g. by looking left for the first main verb, within a word window of fixed size) (Baldwin and Villavicencio 2002; Baldwin 2005a). Similarly, a chunker or phrase structure parser can be used to identify constructions such as noun compounds or VPCs (McCarthy, Keller, and Carroll 2003; Lapata and Lascarides 2003; Kim and Baldwin pear). This style of approach is generally not able to distinguish MWE and literal usages of a given word combination, however, as they are not differentiated in their surface syntax. Deep parsers which have lexical entries for MWEs and disambiguate to the level of lexical items are able to make this distinction, however, via supertagging or full parsing (Baldwin, Bender, Flickinger, Kim, and Oepen 2004; Blunsom 2007).

Another general approach to MWE identification is to treat literal and MWE usages as different senses of a given word combination. This then allows for the application of word sense disambiguation (WSD) techniques to the identification problem. As with WSD research, both supervised (Patrick and Fletcher 2005; Hashimoto and Kawahara 2008) and unsupervised (Birke and Sarkar 2006; Katz and Giesbrecht 2006; Sporleder and Li 2009) approaches

have been applied to the identification task. The key assumption in unsupervised approaches has been that literal usages will be contextually similar to simplex usages of the component words (e.g. *kick* and *bucket* in the case of *kick the bucket*). Mirroring the findings from WSD research, supervised methods tend to be more accurate, but have the obvious drawback that they requires large numbers of annotated literal and idiomatic instances of a given MWE to work. Unsupervised techniques are therefore more generally applicable.

A third approach, targeted particularly at semantically idiomatic MWEs, is to assume that MWEs occur: (a) in canonical forms, or (b) only in particular syntactic configurations, and do not undergo the same level of syntactic variation as literal usages. This relates to our claims in Section 1.3.2.4 relating to non-decomposable VNICs, where the prediction is that VNICs such as *kick the bucket* will not passivise or be internally modifiable. If we have a method of identifying the limits of syntactic variability of a given MWE, therefore, we can assume that any usage which falls outside these (e.g. *kicked a bucket*) must be literal. The problem, then, is identifying the degree of syntactic variability of a given MWE. This can be performed manually, in flagging individual MWE lexical items with predictions of what variations a given MWE can undergo (Li, Zhang, Niu, Jiang, and Srihari 2003; Hashimoto, Sato, and Utsuro 2006). An alternative which alleviates the manual overhead associated with hand annotation is to use unsupervised learning to predict the "canonical" configurations for a given MWE, which can optionally be complemented with a supervised model to identify literal usages which are used in one of the canonical MWE configurations (e.g. *Kim kicked the bucket in frustration, and stormed out of the room*) (Fazly, Cook, and Stevenson 2009).

In research to date, good results have been achieved for particular MWEs, especially English VPCs. However, proposed methods have tended to rely heavily on existing resources such as parsers and hand-crafted lexical resources, and be tuned to particular MWE types.

1.5.2 Extraction

MWE *extraction* is a type-level task, wherein the MWE lexical items attested in a predetermined corpus are extracted out into a lexicon. For example, we may wish to know whether a given corpus provides evidence for a given verb *take* and preposition *off* combining to form a VPC (i.e. *take off*). To illustrate the difference between identification and extraction, identification would involve the determination of the individual occurrences of *take off* (e.g. each of the 240 in a given corpus), whereas extraction would involve the decision about whether *take off* occurred in the corpus or not (irrespective of the number of occurrences). Clearly there is a close connection between the two tasks, in that if we have identified one or more occurrences of a given MWE we can extract it as a MWE, and conversely, if we have extracted a given MWE, we must be able to identify at least one occurrence in the corpus.

The motivation for MWE extraction is generally lexicon development and

expansion, e.g. recognising newly-formed MWEs (e.g. *ring tone* or *shock and awe*) or domain-specific MWEs

Extracting MWEs is relevant to any lexically-driven application, such as grammar engineering or information extraction. Depending on the particular application, it may be necessary to additionally predict lexical properties of a given MWE, e.g. its syntactic or semantic class. In addition, it is particularly important for productive MWEs or domains which are rich in technical terms (e.g. *bus speed* or *boot up* in the IT domain). MWE extraction is difficult for many of the same reasons as MWE identification, namely syntactic flexibility and ambiguity.

There has been a strong focus on the development of general-purpose techniques for MWE extraction, particularly in the guise of *collocation extraction* (see Section 1.2.4). The dominating view here is that extraction can be carried out via association measures such as pointwise mutual information or the t-test, based on analysis of the frequency of occurrence of a given word combination, often in comparison with the frequency of occurrence of the component words (Church and Hanks 1989; Smadja 1993; Frantzi, Ananiadou, and Mima 2000; Evert and Krenn 2001; Pecina 2008). Association measures provide a score for each word combination, which forms the basis of a ranking of MWE candidates. Final extraction, therefore, consists of determining an appropriate cut-off in the ranking, although evaluation is often carried out over the full ranking.

Collocation extraction techniques have been applied to a wide range of extraction tasks over a number of languages, with the general finding that it is often unpredictable which association measure will work best for a given task. As a result, recent research has focused on building supervised classifiers to combine the predictions of a number of association measures, and shown that this leads to consistently superior results than any one association measure (Pecina 2008). It has also been shown that this style of approach works most effectively when combined with POS tagging or parsing, and strict filters on the type of MWE that is being extracted (e.g. ADJECTIVE–NOUN or VERB–NOUN: Justeson and Katz (1995, Pecina (2008)). It is worth noting that association measures have generally been applied to (continuous) word n-grams, or less frequently, pre-determined dependency types in the output of a parser. Additionally, collocational extraction techniques tend to require a reasonable number of token occurrences of a given word combination to operate reliably, which we cannot always assume (Baldwin 2005a; Fazly 2007).

A second approach to MWE extraction, targeted specifically at semantically and statistically idiomatic MWEs, is to extend the general association measure approach to include substitution (Lin 1999; Schone and Jurafsky 2001; Pearce 2001). For example, in assessing the idiomaticity of *red tape*, explicit comparison is made with lexically-related candidates generated by component word substitution, such as *yellow tape* or *red strip*. Common approaches to determining substitution candidates for a given component word are (near-)synonymy—e.g. based on resources such as WORDNET—and distributional

similarity.

Substitution can also be used to generate MWE candidates, and then check for their occurrence in corpus data. For example, if *clear up* is a known (compositional) VPC, it is reasonable to expect that VPCs such as *clean/tidy/unclutter/... up* are also VPCs (Villavicencio 2005). That is not to say that all of these occur as MWEs, however (c.f. **unclutter up*), so an additional check for corpus attestation is usually used in this style of approach.

A third approach, also targeted at semantically idiomatic MWEs, is to analyse the relative similarity between the context of use of a given word combination and its component words (Schone and Jurafsky 2001; Stevenson, Fazly, and North 2004; Widdows and Dorow 2005). Similar to the unsupervised WSD-style approach to MWE identification (see Section 1.5.1), the underlying hypothesis is that semantically idiomatic MWEs will occur in markedly different lexical contexts to their component words. A bag of words representation is commonly used to model the combined lexical context of all usages of a given word or word combination. By interpreting this context model as a vector, it is possible to compare lexical contexts, e.g. via simple cosine similarity (Widdows 2005). In order to reduce the effects of data sparseness, dimensionality reduction is often carried out over the word space prior to comparison (Schütze 1997).

The same approach has also been applied to extract LVCs, based on the assumption that the noun complements in LVCs are often deverbal (e.g. *bath, proposal, walk*), and that the distribution of nouns in PPs post-modifying noun complements in genuine LVCs (e.g. *(make a) proposal of marriage*) will be similar to that of the object of the underlying verb (e.g. *propose marriage*) (Grefenstette and Teufel 1995). Here, therefore, the assumption is that LVCs will be distributionally *similar* to the base verb form of the noun complement, whereas with the original extraction method, the assumption was that semantically idiomatic MWEs are *dissimilar* to their component words.

A fourth approach is to perform extraction on the basis of implicit identification. That is, (possibly noisy) token-level statistics can be fed into a type-level classifier to predict whether there have been genuine instances of a given MWE in the corpus. An example of this style of approach is to use POS taggers, chunkers and parsers to identify English VPCs in different syntactic configurations, and feed the predictions of the various preprocessors into the final extraction classifier (Baldwin 2005a). Alternatively, a parser can be used to identify PPs with singular nouns, and semantically idiomatic PP-Ds extracted from among them based on distributional (dis)similarity of occurrences with and without determiners across a range of prepositions (van der Beek 2005).

A fifth approach is to use syntactic fixedness as a means of extracting MWEs, based on the assumption that semantically idiomatic MWEs undergo syntactic variation (e.g. passivisation or internal modification) less readily than simple verb–noun combinations (Bannard 2007; Fazly, Cook, and Stevenson 2009).

In addition to general-purpose extraction techniques, linguistic properties of particular MWE construction types have been used in extraction. For example, the fact that a given verb–preposition combination occurs as a noun (e.g. *takeoff*, *clip-on*) is a strong predictor of the fact that combination occurring as a VPC (Baldwin 2005a).

One bottleneck in MWE extraction is the token frequency of the MWE candidate. With a few notable exceptions (e.g. (Baldwin 2005a; Fazly, Cook, and Stevenson 2009)), MWE research has tended to ignore low-frequency MWEs, e.g. by applying a method only to word combinations which occur at least N times in a corpus.

1.5.3 Internal Syntactic Disambiguation

As part of the process of MWE identification and extraction, for some MWE types it is necessary to disambiguate the internal syntax of individual MWEs. A prominent case of this in English is noun compounds with 3 or more terms. For example, *glass window cleaner* has two possible interpretations,[15] corresponding to the two possible bracketings of the compound: (1) "a cleaner of glass windows" (= *[[glass window] cleaner]*), and (2) "a cleaner of windows, made of glass" (= *[glass [window cleaner]]*. In this case, the first case (of left bracketing) is the correct analysis, but *movie car chase*, e.g., is right bracketing (= *(movie (car chase))*). The process of disambiguating the syntax of an NC is called *bracketing*.

The most common approach to bracketing is based on statistical analysis of the components of competing analyses. In the *adjacency* model, for a ternary NC *N1 N2 N3*, a comparison is made of the frequencies of the two modifier–head pairings extracted from the two analyses, namely *N1 N2* and *N1 N3* in the left bracketing case, and *N2 N3* and *N1 N3* in the right bracketing case; as *N1 N3* is common to both, in practice, *N1 N2* is compared directly with *N2 N3*. A left bracketing analysis is selected in the case that *N1 N2* is judged to be more likely, otherwise a right bracketing analysis is selected (Marcus 1980). In the *dependency model*, the NC is instead decomposed into the dependency tuples of *N1 N2* and *N2 N3* in the case of left bracketing, and *N2 N3* and *N1 N3* in the case of right bracketing; once again, the dependency *N2 N3* is common to both, and can be ignored. In the instance that *N1 N2* is more likely than *N1 N3*, the model prefers a left bracketing analysis, otherwise a right bracketing analysis is selected (Lauer 1995). While the dependency model tends to outperform the adjacency model, the best-performing models take features derived from both along with various syntactic and semantic features (Nakov and Hearst 2005; Vadas and Curran 2008).

[15] More generally, for an n item noun compound, the number of possible interpretations is defined by the Catalan number $C_n = \frac{1}{n+1}\binom{2n}{n}$.

1.5.4 MWE Interpretation

The *semantic interpretation* of MWEs is usually performed in one of two ways: (1) relative to a generalised semantic inventory (compatible with both simplex words and MWEs, such as WordNet); and (2) based on a set of semantic relations capturing semantic interplay between component words. When interpreting VPCs or lexicalised PP-Ds, e.g., the former approach would be more appropriate (e.g. to capture the fact that *bow out* is synonymous with *withdraw*, both of which are troponyms of *retire*). Nominal MWEs and productive PP-Ds, on the other hand, are more amenable to interpretation by semantic relations (e.g. to capture the semantics of *apple pie* in terms of the MAKE relation, as in "pie made from apple(s)").

One common approach to MWE interpretation is via component similarity, i.e. comparison of the components of a MWE with corresponding components of annotated MWEs, or alternatively with simplex words. For example, a novel NC can be interpreted by identifying training NCs with similar modifier and head nouns (e.g. in interpreting *grape extract*, *grape* would be compared with similar modifiers, and *extract* with similar heads), as determined relative to a lexical resource or via distributional similarity. We can then extrapolate from the closely-matching training NCs to predict the interpretation of the novel NC (Vanderwende 1994; Moldovan, Badulescu, Tatu, Antohe, and Girju 2004; Kim and Baldwin 2005; Nastase, Sayyad-Shirabad, Sokolova, and Szpakowicz 2006; Kim and Baldwin 2007b; Ó Séaghdha 2008). Alternatively, we may employ contextual similarity to compare a VPC with its simplex verb, to determine if they are sufficiently similar that the VPC can be interpreted compositionally from the verb (Baldwin, Bannard, Tanaka, and Widdows 2003; McCarthy, Keller, and Carroll 2003; Cook and Stevenson 2006).

Crosslinguistic evidence can also provide valuable evidence when interpreting MWEs. For example, analysis of what preposition is used in different Romance languages to translate a given English MWE can provide valuable insights into the range of possible interpretations for the English MWE (Girju 2009). Conversely, semantically idiomatic MWEs can be detected from parallel corpus data by identifying translation divergences in the component words lexical choice (Melamed 1997). For example, knowledge that *balance* and *sheet* are most often translated as *équilibre* and *feuille*, respectively, in French, and yet *balance sheet* is translated as *bilan* suggests that *balance sheet* is semantically idiomatic.

One popular approach to determining the underlying semantic relation associated with a MWE is to identify surface realisations or paraphrases associated with each semantic class (Lapata 2002; Grover, Lapata, and Lascarides 2004; Kim and Baldwin 2006; Nicholson and Baldwin 2006; Nakov and Hearst 2008). For example, in the case of compound nominalisations, there are the two primary classes of SUBJECT and OBJECT, based on whether the modifier acts as the subject (e.g. *investor hesitation* = "investor hesitates") or object (e.g. *product replacement* = "replace (the) product") of the base verb form of

the deverbal head. For a given compound nominalisation and base verb form, it is possible to analyse the relative occurrence of the modifier as subject or object of the base verb, and select the interpretation which is most commonly observed (Lapata 2002; Grover, Lapata, and Lascarides 2004; Nicholson and Baldwin 2006).

Another methodology which has been applied to the interpretation task with success is analysis of the co-occurrence properties of the MWE components. For example, the semantics of particles in VPCs can be interpreted by analysing what types of verbs can combine with a given particle (Cook and Stevenson 2006; Kim and Baldwin 2007a). Similarly, Japanese compound verbs (V-V combinations) can be interpreted by observing what set of verbs each of the component verbs combines with to form a compound verb, optionally including the semantic class of the resulting compound verb (Uchiyama, Baldwin, and Ishizaki 2005).

One overarching assumption made in most semantic interpretation tasks is that it is possible to arrive at a compositional interpretation for each MWE via its component words. Ideally, we of course need to identify instances of semantic idiomaticity, motivating the need for methods which can model the relative compositionality or decomposability of MWEs (Lin 1999; Baldwin, Bannard, Tanaka, and Widdows 2003; McCarthy, Keller, and Carroll 2003; McCarthy, Venkatapathy, and Joshi 2007).

While there has been a healthy interest in MWE interpretation, research has suffered from lack of agreement on semantic inventories, and the relative unavailability of annotated data. One very positive step towards redressing this situation was a shared task at SemEval-2007, on interpreting nominal MWEs in English (Girju, Nakov, Nastase, Szpakowicz, Turney, and Yuret 2007), and an upcoming SemEval-2010 task on the multi-way classification of semantic relations between pairs of nominals. In practice, the SemEval-2007 task took a pair of nouns in a fixed sentential context and attempted to determine if they were interpretable using a set of semantic relations compatible with NCs. As such, the task wasn't specifically on NC interpretation, but NC interpretation methods could be evaluated over the dataset (Kim and Baldwin 2008; Ó Séaghdha 2008). Crucially, the task organisers chose to sidestep the controversy surrounding the precise membership of a broad-coverage set of semantic relations, and instead focused on relations where there is relatively high agreement between researchers. They additionally defused the question of interpretational overlap/ambiguity of a given nominal, by designing the task as a series of binary sub-tasks, where a prediction had to be made about each nominal's compatibility with a given semantic relation (ignoring whether or not it was also compatible with other relations).

1.6 Summary

MWEs are an integral part of language: vast in number and highly varied in nature. They are defined by idiomaticity at the lexical, syntactic, semantic, pragmatic and statistical levels, and occur in a myriad of different constructions in the world's languages. In addition to providing a brief foray into the linguistic complexities of MWEs, we have detailed the key MWEs in MWE research, and outlined various approaches to the primary computational challenges associated with MWEs, namely: identification, extraction and interpretation.

We have deliberately not provided a survey of MWE resources in this paper, choosing instead to maintain an up-to-the-moment snapshot of the field on the companion website at `http://...`. For those interested in pursuing MWE research, we recommend this as your first port of call. For readers who are interested in further reading on MWEs, we particularly recommend the following works: (Moon 1998; McKeown and Radev 2000; Cowie 2001; Sag, Baldwin, Bond, Copestake, and Flickinger 2002; Villavicencio, Bond, Korhonen, and McCarthy 2005).

Bibliography

Abeillé, A. (1988). Light verb constructions and extraction out of NP in a tree adjoining grammar. In *Papers of the 24th Regional Meeting of the Chicago Linguistics Society*.

Ahn, H.-D. (1991). *Light verbs, VP-movement, Negation and Clausal Structure in Korean and English*. Ph. D. thesis, University of Wisconsin-Madison.

Aronoff, M. (1976). *Word Formation in Generative Grammar*. Cambridge, USA: MIT Press.

Baldwin, T. (2005a). The deep lexical acquisition of English verb-particles. *Computer Speech and Language, Special Issue on Multiword Expressions 19*(4), 398–414.

Baldwin, T. (2005b). Looking for prepositional verbs in corpus data. In *Proceedings of the 2nd ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Colchester, UK, pp. 115–126.

Baldwin, T., C. Bannard, T. Tanaka, and D. Widdows (2003). An empirical model of multiword expression decomposability. In *Proceedings of the*

ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, Sapporo, Japan, pp. 89–96.

Baldwin, T., J. Beavers, L. Van Der Beek, F. Bond, D. Flickinger, and I. A. Sag (2006). In search of a systematic treatment of determinerless PPs. In P. Saint-Dizier (Ed.), *Syntax and Semantics of Prepositions*. Springer.

Baldwin, T., E. M. Bender, D. Flickinger, A. Kim, and S. Oepen (2004). Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, pp. 2047–2050.

Baldwin, T. and F. Bond (2002). Multiword expressions: Some problems for Japanese NLP. In *Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing (Japan)*, Keihanna, Japan, pp. 379–382.

Baldwin, T. and A. Villavicencio (2002). Extracting the unextractable: A case study on verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, Taipei, Taiwan, pp. 98–104.

Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the ACL-2007 Workshop on A Broader Perspective on Multiword Expressions*, Prague, Czech Republic, pp. 1–8.

Bannard, C., T. Baldwin, and A. Lascarides (2003). A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, Sapporo, Japan, pp. 65–72.

Barker, K. and S. Szpakowicz (1998). Semi-automatic recognition of noun modifier relationships. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-1998)*, Montreal, Canada, pp. 96–102.

Bauer, L. (1983). *English Word-formation*. Cambridge, UK: Cambridge University Press.

Bauer, L. (2001). Compounding. In M. Haspelmath (Ed.), *Language Typology and Language Universals*. The Hague, Netherlands: Mouton de Gruyter.

Benor, S. B. and R. Levy (2006). The chicken or the egg? a probabilistic analysis of english binomials. *Language 82*(2), 233–278.

Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography 3*(1), 23–35.

Birke, J. and A. Sarkar (2006). A clustering approach for the nearly unsupervised recoginition of nonliteral language. In *Proceedings of the 11th Conference of the EACL (EACL 2006)*, Trento, Italy, pp. 329–336.

Blunsom, P. (2007). *Structured Classication for Multilingual Natural Language Processing*. Ph. D. thesis, University of Melbourne.

Bolinger, D. (1976). *The Phrasal Verb in English*. Boston, USA: Harvard University Press.

Bond, F. (2005). *Translating the Untranslatable: A solution to the Problem of Generating English Determiners*. CSLI Studies in Computational Linguistics. CSLI Publications.

Booij, G. (2002). Separable complex verbs in Dutch: A case of periphrastic word formation. In N. Dehé, R. Jackendoff, A. McIntyre, and S. Urban (Eds.), *Verb-particle explorations*, pp. 21–41. Berlin, Germany / New York, USA: Mouton de Gruyter.

Brinton, L. (1985). Verb particles in English: Aspect or aktionsart. *Studia Linguistica 39*, 157–168.

Butt, M. (2003). The light verb jungle. In *Proceedings of the Workshop on Multi-Verb Constructions*, Trondheim, Norway, pp. 1–49.

Calzolari, N., C. Fillmore, R. Grishman, N. Ide, A. Lenci, C. MacLeod, and A. Zampolli (2002). Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, pp. 1934–1940.

Chafe, W. L. (1968). Idiomaticity as an anomaly in the Chomskyan paradigm. *Foundations of Language 4*, 109–127.

Chander, I. (1998). *Automated postediting of documents*. Ph. D. thesis, University of Southern California.

Choueka, Y. (1988). Lookin for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of RIAO*, pp. 43–38.

Church, K. W. and P. Hanks (1989). Word assication norms, mutual information and lexicography. In *Proceedings of the 27th Annual Meeting of the Association of Computational Linguistics (ACL-1989)*, Vancouver, Canada, pp. 76–83.

Cook, P. and S. Stevenson (2006). Classifying particle semantics in English verb-particle constructions. In *Proceedings of the ACL-2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, Australia, pp. 45–53.

Copestake, A. and A. Lascarides (1997). Integrating symbolic and statistical representations: The lexicon pragmatics interface. In *Proceedings of the 35th Annual Meeting of the Association of Coomputational Linguistics and 8th Conference of the European Chapter of Association of Computational Linguistics (ACL/EACL-1997)*, Madrid, Spain, pp. 136–143.

Cowie, A. (Ed.) (2001). *Phraseology : Theory, Analysis, and Applications*. Oxford, UK: Oxford University Press.

Cowie, A. P. and P. A. Howarth (1996). Phraseology – a select bibliography. *International Journal of Lexicography 9*(1), 38–51.

Cruse, A. D. (1986). *Lexical Semantics*. Cambridge, UK: Cambridge University Press.

Dirven, R. (2001). The metaphoric in recent cognitive approaches to English phrasal verbs. *metaphorik.de 1*, 39–54.

Downing, P. (1977). On the creation and use of English compound nouns. *Language 53*(4), 810–842.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph. D. thesis, University of Stuttgart.

Evert, S. and B. Krenn (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the ACL and 10th Conference of the EACL (ACL-EACL 2001)*, Toulouse, France, pp. 188–195.

Fazly, A. (2007). *Automatic Acquisition of Lexical Knowledge about Multiword Predicates*. Ph. D. thesis, University of Toronto.

Fazly, A., P. Cook, and S. Stevenson (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics 35*(1), 61–103.

Fellbaum, C. (Ed.) (1998). *WordNet, An Electronic Lexical Database*. Cambridge, Massachusetts, USA: MIT Press.

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

Fernando, C. and R. Flavell (1981). *On idioms*. Exeter: University of Exeter.

Fillmore, C., P. Kay, and M. C. O'Connor (1988). Regularity and idiomaticity in grammatical constructions. *Language 64*, 501–538.

Frantzi, K., S. Ananiadou, and H. Mima (2000). Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries 3*(2), 115–130.

Gagné, C. L., T. L. Spalding, and M. C. Gorrie (2005). Sentential context and the interpretation of familiar open-compounds and novel modifier-noun phrases. *Language and Speech 28*(2), 203–221.

Gates, E. (1988). *The treatment of multiword lexemes in some current dictionaries of English*. Snell-Hornby.

Gerber, L. and J. Yang (1997). Systran MT dictionary development. In *Proceedings of the Sixth Machine Translation Summit (MT Summit VI)*, San Diego, USA.

Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory and Cognition 8*(2), 149–156.

Girju, R. (2009). The syntax and semantics of prepositions in the task of automatic interpretation of nominal phrases and compounds: A cross-linguistic study. *Computational Linguistics 35*(2).

Girju, R., P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, and D. Yuret (2007). Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, Czech Republic, pp. 13–18.

Grefenstette, G. and P. Tapanainen (1994). What is a word, what is a sentence? problems of tokenization. In *Proceedings of the 3rd Conference on Computational Lexicography and Text Research*, Budapest, Hungary, pp. 79–87.

Grefenstette, G. and S. Teufel (1995). A corpus-based method for automatic identification of support verbs for nominalizations. In *Proceedings of the 7th European Chapter of Association of Computational Linguistics (EACL-1995)*, Dublin, Ireland, pp. 98–103.

Grimshaw, J. and A. Mester (1988). Light verbs and *theta*-marking. *Linguistic Inquiry 19*(2), 205–232.

Grishman, R., C. Macleod, and A. Myers (1998). COMLEX syntax reference manual.

Grover, C., M. Lapata, and A. Lascarides (2004). A comparison of parsing technologies for the biomedical domain. *Journal of Natural Language Engineering 1*(1), 1–38.

Hashimoto, C. and D. Kawahara (2008). Construction of an idiom corpus and its application to idiom identification based on WSD incorporating idiom-specific features. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, Honolulu, USA, pp. 992–1001.

Hashimoto, C., S. Sato, and T. Utsuro (2006). Japanese idiom recognition: Drawing a line between literal and idiomatic meanings. In *Proceedings of the COLING/ACL 2006 Interactive Poster System*, Sydney, Australia, pp. 353–360.

Haspelmath, M. (1997). *From Space to Time in The World's Languages*. Munich, Germany: Lincorn Europa.

Hoshi, H. (1994). *Passive, Causive, and Light Verbs: A Study of Theta Role Assignment*. Ph. D. thesis, University of Connecticut.

Huddleston, R. and G. K. Pullum (2002). *The Cambridge Grammar of the English Language*. Cambridge, UK: Cambridge University Press.

Jackendoff, R. (1973). The base rules for prepositional phrases. In *A Festschrift for Morris Halle*, pp. 345–356. New York, USA: Rinehart and Winston.

Jackendoff, R. (1997). *The Architecture of the Language Faculty*. Cambridge, USA: MIT Press.

Jackendoff, R. (2002). *Foundations of Language*. Oxford, UK: Oxford University Press.

Jespersen, O. (1965). *A Modern English Grammar on Historical Principles, Part VI, Morphology*. London, UK: George Allen and Unwin Ltd.

Justeson, J. S. and S. M. Katz (1995). Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering 1*(1), 9–27.

Kaalep, H.-J. and K. Muischnek (2008). Multi-word verbs of Estonian: a database and a corpus. In *Proceedings of the LREC 2008 Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, Morocco, pp. 23–26.

Kageura, K., B. Daille, H. Nakagawa, and L.-F. Chien (2004). Recent trends in computational terminology. *Terminology 10*(1), 1–21.

Karimi-Doostan, G. H. (1997). *Light Verb Construction in Persian*. Ph. D. thesis, University of Essex.

Kastovsky, D. (1982). *Wortbildung und Semantik*. Dusseldorf: Bagel/Francke.

Katz, G. and E. Giesbrecht (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL-2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, Australia, pp. 28–35.

Katz, J. J. and P. M. Postal (2004). Semantic interpretation of idioms and sentences containing them. In *Quarterly Progress Report (70), MIT Research Laboratory of Electronics*, pp. 275–282. MIT Press.

Keysar, B. and B. Bly (1995). Intuitions of the transparency of idioms: Can one keep a secret by spilling the beans? *Journal of Memory and Language 34*(1), 89–109.

Kim, S. N. and T. Baldwin (2005). Automatic interpretation of compound nouns using WordNet similarity. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, Jeju, Korea, pp. 945–956.

Kim, S. N. and T. Baldwin (2006). Interpreting semantic relations in noun compounds via verb semantics. In *Proceedings of the COLING/ACL 2006 Interactive Poster System*, Sydney, Australia, pp. 491–498.

Kim, S. N. and T. Baldwin (2007a). Detecting compositionality of English verb-particle constructions using semantic similarity. In *Proceedings of Conference of the Pacific Association for Computational Linguistics*, Melbourne, Australia, pp. 40–48.

Kim, S. N. and T. Baldwin (2007b). Disambiguating noun compounds. In *Proceedings of 22nd AAAI Conference on Artificial Intelligenc*, Vancouver, Canada, pp. 901–906.

Kim, S. N. and T. Baldwin (2008). Benchmarking noun compound interpretation. In *Proceedings of 3rd International Joint Conference on Natual Language Processing (IJCNLP-2008)*, Hyderabad, India, pp. 569–576.

Kim, S. N. and T. Baldwin (to appear). How to pick out token instances of English verb-particle constructions. *Language Resources and Evaluation*.

Lapata, M. (2002). The disambiguation of nominalizations. *Computational Linguistics 28*(3), 357–388.

Lapata, M. and A. Lascarides (2003). Detecting novel compounds: The role of distributional evidence. In *Proceedings of the 11th Conference of the European Chapter for the Association of Computational Linguistics (EACL-2003)*, Budapest, Hungary, pp. 235–242.

Lauer, M. (1995). *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph. D. thesis, Macquarie University.

Levi, J. (1978). *The Syntax and Semantics of Complex Nominals*. New York, USA: Academic Press.

Li, W., X. Zhang, C. Niu, Y. Jiang, and R. K. Srihari (2003). An expert lexicon approach to identifying English phrasal verbs. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, Sapporo, Japan, pp. 513–520.

Liberman, M. and R. Sproat (1992). The stress and structure of modified noun phrases in English. In I. A. Sag and A. Szabolcsi (Eds.), *Lexical Matters – CSLI Lecture Notes No. 24*. Stanford, USA: CSLI Publications.

Lidner, S. (1983). *A lexico-semantic analysis of English verb particle constructions with OUT and UP*. Ph. D. thesis, University of Indiana at Bloomington.

Lieber, R. and P. Štekauer (Eds.) (2009). *The Oxford Handbook of Compounding*. Oxford University Press.

Lin, D. (1998). Extracting collocations from text corpora. In *Proceedings of the 1st Workshop on Computational Terminology*, Montreal, Canada.

Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, USA, pp. 317–324.

Lüdeling, A. (2001). *On Particle Verbs and Similar Constructions in German*. Stanford, USA: CSLI Publications.

Marcus, M. (1980). *A Theory of Syntactic Recognition for Natural Language*. Cambridge, USA: MIT Press.

Matsuo, Y., S. Shirai, A. Yokoo, and S. Ikehara (1997). Direct parse tree translation in cooperation with the transfer method. In D. Joneas and H. Somers (Eds.), *New Methods in Language Processing*, pp. 229–238. London, UK: UCL Press.

McCarthy, D., B. Keller, and J. Carroll (2003). Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, Sapporo, Japan, pp. 73–80.

McCarthy, D., S. Venkatapathy, and A. Joshi (2007). Detecting compositionality of verb-object combinations using selectional preferences. In *Proceedings of the 200 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 369–379.

McIntyre, A. (2007). Particle verbs and argument structure. *Language and Linguistics Compass 1*(4), 350–367.

McKeown, K. R. and D. R. Radev (2000). Collocations. In R. Dale, H. Moisl, and H. Somers (Eds.), *A Handbook of Natural Language Processing*, Chapter 15. Marcel Dekker.

Melamed, I. D. (1997). Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP-97)*, Providence, USA, pp. 97–108.

Mel'čuk, I. A. and A. Polguère (1987). A formal lexicon in the Meaning-Text Theory (or how to do lexica with words). *Computational Linguistics 13*(3–4), 261–275.

Mimmelmann, N. P. (1998). Regularity in irregularity: Article use in adpositional phrases. *Linguistic Typology 2*, 315–353.

Miyagawa, S. (1989). Light verbs and the ergative hypothesis. *Linguistic Inquiry 20*, 659–668.

Miyazaki, M., S. Ikehara, and A. Yokoo (1993). Combined word retrieval for bilingual dictionary based on the analysis of compound word. *Transactions of the Information Processing Society of Japan 34*(4), 743–754. (in Japanese).

Mohanan, T. (1994). *Argument Structure in Hindi*. Stanford, USA: CSLI Publications.

Moldovan, D., A. Badulescu, M. Tatu, D. Antohe, and R. Girju (2004). Models for the semantic classification of noun phrases. In *Proceedings of HLT-NAACL 2004: Workshop on Computational Lexical Semantics*, Boston, USA, pp. 60–67.

Moon, R. E. (1998). *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford, UK: Oxford University Press.

Nakov, P. (2008). Noun compound interpretation using paraphrasing verbs: Feasibility study. In *Proceedings of the 13th International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA'08)*, Varna, Bulgaria, pp. 103–117.

Nakov, P. and M. Hearst (2005). Search engine statistics beyond the *n*-gram: Application to noun compound bracketting. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)*, Ann Arbor, USA, pp. 17–24.

Nakov, P. and M. A. Hearst (2008). Solving relational similarity problems using the web as a corpus. In *Proceedings of the 46th Annual Meeting of the ACL: HLT*, Columbus, USA, pp. 452–460.

Nastase, V., J. Sayyad-Shirabad, M. Sokolova, and S. Szpakowicz (2006). Learning noun-modifier semantic relations with corpus-based and WordNet-based features. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, Boston, USA, pp. 781–787.

Nicholson, J. and T. Baldwin (2006). Interpretation of compound nominalisations using corpus and web statistics. In *Proceedings of the COLING/ACL 2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, Australia, pp. 54–61.

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

O'Dowd, E. M. (1998). *Prepositions and Particles in English*. Oxford University Press.

Ó Séaghdha, D. (2008). *Learning compound noun semantics*. Ph. D. thesis, Computer Laboratory, University of Cambridge.

Osswald, R., H. Helbig, and S. Hartrumpf (2006). The representation of German prepositional verbs in a semantically based computer lexicon. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Patrick, J. and J. Fletcher (2005). Classifying verb particle constructions by verb arguments. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Colchester, UK, pp. 200–209.

Pauwels, P. (2000). *Put, set, lay, and place: a cognitve linguistic approach to verbal meaning*. Munich, Germany: Lincom Europa.

Pearce, D. (2001). Synonymy in collocation extraction. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, USA, pp. 41–46.

Pecina, P. (2008). *Lexical Association Measures*. Ph. D. thesis, Charles University.

Piao, S., P. Rayson, D. Archer, A. Wilson, and T. McEnery (2003). Extracting multiword expressions wth a semantic tagger. In *Proceedings of the ACL2003 Workshop on Multiword Expressions: analysis, acquisition and treatment*, Sapporo, Japan, pp. 49–56.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985). *A Comprehensive Grammar of the English Language*. London, UK: Longman.

Riehemann, S. (2001). *A Constructional Approach to Idioms and Word Formation*. Ph. D. thesis, Stanford University.

Rosario, B. and M. Hearst (2001). Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP-2001)*, Pittsburgh, Pennsylvania, USA, pp. 82–90.

Ross, H. (1995). Defective noun phrases. In *In Papers of the 31st Regional Meeting of the Chicago Linguistics Society*, Chicago, Illinois, USA, pp. 398–440.

Sag, I. A., T. Baldwin, F. Bond, A. Copestake, and D. 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-2002)*, Mexico City, Mexico, pp. 1–15.

Sager, J. C. (1990). *A Practical Course in Terminology Processing*. Amsterdam, Netherlands / Philadelphia, USA: John Benjamins.

Schone, P. and D. Jurafsky (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*, Hong Kong, China, pp. 100–108.

Schütze, H. (1997). *Ambiguity Resolution in Language Learning*. Stanford, USA: CSLI Publications.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford, UK: Oxford University Press.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics 19*(1), 143–77.

Spärck Jones, K. (1983). *Compound noun interpretation problems*. Englewood Cliffes, USA: Prentice-Hall.

Sporleder, C. and L. Li (2009). Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the EACL (EACL 2009)*, Athens, Greece, pp. 754–762.

Stevenson, S., A. Fazly, and R. North (2004). Statistical measures of the semi-productivity of light verb constructions. In *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain, pp. 1–8.

Stvan, L. S. (1998). *The Semantics and Pragmatics of Bare Singular Noun Phrases*. Ph. D. thesis, Northwestern University.

Tan, Y. F., M.-Y. Kan, and H. Cui (2006). Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of the EACL 2006 Workshop on Multi-word-expressions in a multilingual context (MWEmc)*, Trento, Italy.

Tanaka, T. and T. Baldwin (2003). Noun-noun compound machine translation a feasibility study on shallow processing. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 17–24.

Trawiński, B., M. Sailer, and J.-P. Soehn (2006). Combinatorial aspects of collocational prepositional phrases. In P. Saint-Dizier (Ed.), *Computational Linguistics Dimensions of Syntax and Semantics of Prepositions*. Dordrecht, Netherlands: Kluwer Academic.

Trawiński, B., M. Sailer, J.-P. Soehn, L. Lemnitzer, and F. Richter (2008). Cranberry expressions in English and in German. In *Proceedings of the LREC 2008 Workshop: Towards a Shared Task for Multiword Expressions (MWE 2008)*, Marrakech, Morocco, pp. 35–38.

Tschichold, C. (1998). *Multi-word Units in Natural Language Processing*. Ph. D. thesis, University of Basel.

Tsuchiya, M., T. Shime, T. Takagi, T. Utsuro, K. Uchimoto, S. Matsuyoshi, S. Sato, and S. Nakagawa (2006). Chunking Japanese compound functional expressions by machine learning. In *Proceedings of the EACL 06 Workshop on Multi-word-expressions in a Multilingual Context*, Trento, Italy, pp. 25–32.

Uchiyama, K., T. Baldwin, and S. Ishizaki (2005). Disambiguating Japanese compound verbs. *Computer Speech and Language, Special Issue on Multiword Expressions 19*(4), 497–512.

Vadas, D. and J. R. Curran (2008). Parsing noun phrase structure with CCG. In *Proceedings of the 46th Annual Meeting of the ACL: HLT*, Columbus, USA, pp. 335–343.

van der Beek, L. (2005). The extraction of determinerless PPs. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Colchester, UK, pp. 190–199.

Vanderwende, L. (1994). Algorithm for automatic interpretation of noun sequences. In *Proceedings of the 15th Conference on Computational linguistics*, Kyoto, Japan, pp. 782–788.

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

Villada Moirón, B. (2005). *Data-driven identification of fixed expressions and their modifiability*. Ph. D. thesis, Alfa-Informatica, University of Groningen.

Villavicencio, A. (2005). The availability of verb-particle constructions in lexical resources: How much is enough? *Computer Speech and Language, Special Issue on Multiword Expressions 19*(4), 415–432.

Villavicencio, A., T. Baldwin, and B. Waldron (2004). A multilingual database of idioms. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, pp. 1127–1130.

Villavicencio, A., F. Bond, A. Korhonen, and D. McCarthy (2005). Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech and Language, Special Issue on Multiword Expressions 19*(4), 365–377.

Widdows, D. (2005). *Geometry and Meaning*. Stanford, USA: CSLI Publications.

Widdows, D. and B. Dorow (2005). Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In *Proceedings of the ACL 2004 Workshop on Deep Lexical Acquisition*, Ann Arbor, USA, pp. 48–56.

Xu, R., Q. Lu, and S. Li (2006). The design and construction of a Chinese collocation bank. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

Zhang, Y., V. Kordoni, A. Villavicencio, and M. Idiart (2006). Automated multiword expression prediction for grammar engineering. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, Sydney, Australia, pp. 36–44. Association for Computational Linguistics.