Multiword Expressions: A Pain in the Neck for NLP*

Ivan A. Sag[1], Timothy Baldwin[1], Francis Bond[2], Ann Copestake[3], and Dan Flickinger[1]

[1] CSLI, Ventura Hall, Stanford University
Stanford, CA 94305 USA
{sag,tbaldwin,danf}@csli.stanford.edu

[2] NTT Communication Science Labs., 2-4 Hikaridai
Seika-cho, Soraku-gun, Kyoto, Japan 619-0237
bond@cslab.kecl.ntt.co.jp

[3] University of Cambridge, Computer Laboratory, William Gates Building
JJ Thomson Avenue, Cambridge CB3 OFD, UK
Ann.Copestake@cl.cam.ac.uk

Abstract. Multiword expressions are a key problem for the development of large-scale, linguistically sound natural language processing technology. This paper surveys the problem and some currently available analytic techniques. The various kinds of multiword expressions should be analyzed in distinct ways, including listing "words with spaces", hierarchically organized lexicons, restricted combinatoric rules, lexical selection, "idiomatic constructions" and simple statistical affinity. An adequate comprehensive analysis of multiword expressions must employ both symbolic and statistical techniques.

1 Introduction

The tension between symbolic and statistical methods has been apparent in natural language processing (NLP) for some time. Though some believe that the statistical methods have rendered linguistic analysis unnecessary, this is in fact not the case. Modern statistical NLP is crying out for better language models (Charniak 2001). At the same time, while 'deep' (linguistically precise) processing has now crossed the industrial threshold (Oepen et al. 2000) and serves as the basis for ongoing product development in a number of application areas (e.g. email autoresponse), it is widely recognized that deep analysis must come

* The research reported here was conducted in part under the auspices of the LinGO project, an international collaboration centered around the LKB system and related resources (see http://lingo.stanford.edu). This research was supported in part by the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University. We would like to thank Emily Bender and Tom Wasow for their contributions to our thinking. However, we alone are responsible for any errors that remain.

In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002), Mexico City, Mexico, pp. 1-15.

to grips with two key problems, if linguistically precise NLP is to become a reality.

The first of these is **disambiguation**. Paradoxically, linguistic precision is inversely correlated with degree of sentence ambiguity. This is a fact of life encountered by every serious grammar development project. Though knowledge representation, once thought to hold the key to the problem of disambiguation, it has largely failed to provide completely satisfactory solutions. Most research communities we are aware of that are currently developing large scale, linguistically precise, computational grammars are now exploring the integration of stochastic methods for ambiguity resolution. The second key problem facing the deep processing program – the problem of **multiword expressions** – is underappreciated in the field at large. There is insufficient ongoing work investigating the nature of this problem or seeking computationally tractable techniques that will contribute to its solution.

We define multiword expressions (MWEs) very roughly as "idiosyncratic interpretations that cross word boundaries (or spaces)". As Jackendoff (1997: 156) notes, the magnitude of this problem is far greater than has traditionally been realized within linguistics. He estimates that the number of MWEs in a speaker's lexicon is of the same order of magnitude as the number of single words. In fact, it seems likely that this is an underestimate, even if we only include lexicalized phrases. In WordNet 1.7 (Fellbaum 1999), for example, 41% of the entries are multiword. For a wide coverage NLP system, this is almost certainly an underestimate. Specialized domain vocabulary, such as terminology, overwhelmingly consists of MWEs, and a system may have to handle arbitrarily many such domains. As each new domain adds more MWEs than simplex words, the proportion of MWEs will rise as the system adds vocabulary for new domains.

MWEs appear in all text genres and pose significant problems for every kind of NLP. If MWEs are treated by general, compositional methods of linguistic analysis, there is first an **overgeneration problem**. For example, a generation system that is uninformed about both the patterns of compounding and the particular collocational frequency of the relevant dialect would correctly generate *telephone booth* (American) or *telephone box* (British/Australian), but might also generate such perfectly compositional, but unacceptable examples as *telephone cabinet*, *telephone closet*, etc. A second problem for this approach is what we will call the **idiomaticity problem**: how to predict, for example, that an expression like *kick the bucket*, which appears to conform to the grammar of English VPs, has a meaning unrelated to the meanings of *kick*, *the*, and *bucket*. Syntactically-idiomatic MWEs can also lead to parsing problems, due to nonconformance with patterns of word combination as predicted by the grammar (e.g. the determinerless *in line*).

Many have treated MWEs simply as **words-with-spaces**, an approach with serious limitations of its own. First, this approach suffers from a **flexibility problem**. For example, a parser that lacks sufficient knowledge of verb-particle constructions might correctly assign *look up the tower* two interpretations ("glance up at the tower" vs. "consult a reference book about the tower"), but fail to

treat the subtly different *look the tower up* as unambiguous ("consult a reference book ..." interpretation only). As we will show, MWEs vary considerably with respect to this and other kinds of flexibility. Finally, this simple approach to MWEs suffers from a **lexical proliferation problem**. For example, light verb constructions often come in families, e.g. *take a walk*, *take a hike*, *take a trip*, *take a flight*. Listing each such expression results in considerable loss of generality and lack of prediction. Many current approaches are able to get commonly-attested MWE usages right, but they use ad hoc methods to do so, e.g. preprocessing of various kinds and stipulated, inflexible correspondences. As a result, they handle variation badly, fail to generalize, and result in systems that are quite difficult to maintain and extend.

Though the theory of MWEs is underdeveloped and the importance of the problem is underappreciated in the field at large, there is ongoing work on MWEs within various projects that are developing large-scale, linguistically precise computational grammars, including the ParGram Project at Xerox PARC (`http://www.parc.xerox.com/istl/groups/nltt/pargram/`), the XTAG Project at the University of Pennsylvania (`http://www.cis.upenn.edu/~xtag/`), work on Combinatory Categorial Grammar at Edinburgh University, and the LinGO Project (a multi-site collaboration including CSLI's English Resource Grammar Project — `http://lingo.stanford.edu`), as well as by the FrameNet Project (`http://www.icsi.berkeley.edu/~framenet/`), which is primarily developing large-scale lexical resources. All of these projects are currently engaged (to varying degrees) in linguistically informed investigations of MWEs.[1]

We believe the problem of MWEs is critical for NLP, but there is a need for better understanding of the diverse kinds of MWE and the techniques now readily available to deal with them. In Section 2, we provide a general outline of some common types of MWE in English and their properties. In Section 3, we survey a few available analytic techniques and comment on their utility, drawing from our own research using HPSG-style grammars and the LKB system. In the conclusion, we reflect on prospects for the future of MWE research.

2 Some Kinds of MWE

MWEs can be broadly classified into **lexicalized phrases** and **institutionalized phrases** (terminology adapted from Bauer (1983)). Lexicalized phrases have at least partially idiosyncratic syntax or semantics, or contain 'words' which do not occur in isolation; they can be further broken down into **fixed expressions**, **semi-fixed expressions** and **syntactically-flexible expressions**, in roughly decreasing order of lexical rigidity. Institutionalized phrases are syntactically and semantically compositional, but occur with markedly high frequency (in a given context). Below, we examine instances of each category and discuss some of the peculiarities that pose problems for both words-with-spaces and fully compositional analyses.

[1] We thank Chuck Fillmore, Aravind Joshi, Ron Kaplan, and Mark Steedman for discussions of this point.

2.1 Fixed Expressions

There is a large class of immutable expressions in English that defy conventions of grammar and compositional interpretation. This class includes *by and large*, *in short*, *kingdom come*, and *every which way*. Many other MWEs, though perhaps analyzable to scholars of the languages whence they were borrowed, belong in this class as well, at least for the majority of speakers: *ad hoc* (cf. *ad nauseum*, *ad libitum*, *ad hominem*,...), *Palo Alto* (cf. *Los Altos*, *Alta Vista*,...), etc.

Fixed expressions are fully lexicalized and undergo neither morphosyntactic variation (cf. **in shorter*) nor internal modification (cf. **in very short*)). As such, a simple words-with-spaces representation is sufficient. If we were to adopt a compositional account of fixed expressions, we would have to introduce a lexical entry for "words" such as *hoc*, resulting in overgeneration and the idiomaticity problem (see above).

2.2 Semi-Fixed Expressions

Semi-fixed expressions adhere to strict constraints on word order and composition, but undergo some degree of lexical variation, e.g. in the form of inflection, variation in reflexive form, and determiner selection. This makes it possible to treat them as a word complex with a single part of speech, which is lexically variable at particular positions. They can take a range of forms including non-decomposable idioms, and certain compound nominals and proper names. Below, we discuss some problematic instances of each, for which neither a fully compositional account nor simple string-type listing in a lexicon is appropriate.

Non-Decomposable Idioms Nunberg et al. (1994) introduced the notion of 'semantic compositionality' in relation to idioms, as a means of describing how the overall sense of a given idiom is related to its parts. Idioms such as *spill the beans*, for example, can be analyzed as being made up of *spill* in a "reveal" sense and *the beans* in a "secret(s)" sense, resulting in the overall compositional reading of "reveal the secret(s)". With the oft-cited *kick the bucket*, on the other hand, no such analysis is possible.

Based on the observation that this process of semantic deconstruction starts off with the idiom and associates particular components of the overall meaning with its parts, it has been recast as **semantic decomposability**. We distinguish between **decomposable idioms** such as *spill the beans* and *let the cat out of the bag*, and **non-decomposable idioms** such as *kick the bucket*, *trip the light fantastic* and *shoot the breeze*. We return to discuss decomposable idioms in Section 2.3.

Due to their opaque semantics, non-decomposable idioms are not subject to syntactic variability, e.g. in the form of internal modification (*#kick the great bucket in the sky*[2]) or passivization (**the breeze was shot*). The only types of

[2] We make the claim that *proverbial* as in *kick the proverbial bucket* is a metalinguistic marker, and thus does not qualify as an internal modifier.

lexical variation observable in non-decomposable idioms are inflection (*kicked the bucket*) and variation in reflexive form (*wet oneself*).

Adopting a words-with-spaces description of non-decomposable idioms is unable to capture the effects of inflectional variation and variation in reflexive form, except at the risk of lexical proliferation in describing all possible lexical variants of each idiom (with well over 20 lexical entries for *wet/wets/wetted/wetting myself/yourself/herself/himself/themselves/oneself/itself*). On the other hand, a fully compositional account may have no trouble with lexical variation, but it has troubles with idiomaticity (e.g. deriving the "die" semantics from *kick*, *the*, and *bucket*) and overgeneration (e.g. in generating **the breeze was shot*).

Compound Nominals Compound nominals such as *car park*, *attorney general* and *part of speech* are similar to non-decomposable idioms in that they are syntactically-unalterable units that inflect for number. For many right-headed compound nominals, a words-with-spaces handling can generally cope with number inflection by way of the simplex word mechanism of simply adding an *-s* to the end of the string, as in [*car park*]*s*. For left-headed compounds such as *attorney general*, *congressman at large* and *part of speech*, on the other hand, this would result in anomalies such as *[*congressman at large*]*s*. Admittedly, the lexical proliferation associated with listing the singular and plural forms of each compound nominal is less dramatic than with non-decomposable idioms, but still leaves a lot to be desired in terms of systematicity.

As for non-decomposable idioms, fully compositional approaches suffer from the idiomaticity and overgeneration problems.

Proper Names Proper names are syntactically highly idiosyncratic. U.S. sports team names, for example, are canonically made up of a place or organization name (possibly a MWE in itself, such as *San Francisco*) and an appellation that locates the team uniquely within the sport (such as *49ers*). The first obstacle for a words-with-spaces representation for U.S. team names is that the place/organization name is optionally ellidable (e.g. *the* (*San Francisco*) *49ers*), a generalization which cannot be captured by a single string-based lexical entry.

Additionally, U.S. sports team names take a definite reading. This results in the determiner *the* being selected by default when the team name occurs as an NP, as in *the* (*San Francisco*) *49ers* and *the* (*Oakland*) *Raiders*. When the team name occurs as a modifier in a compound noun (as in *an/the* [[(*Oakland*) *Raiders*] *player*]), however, the determiner is associated with the compound noun, and the team name becomes determinerless. Coordination also produces interesting effects, as it is possible to have a single determiner for a coordinated team name complex, as in *the* [*Raiders and 49ers*].

Lexical proliferation once again becomes a problem with a words-with-spaces approach to U.S. sports team names. We would need to generate lexicalizations incorporating the determiners *the* or *those*, as well as alternative lexicalizations with no determiner. And all of these would have to allow the place/organization name to be optional (e.g. *the San Francisco 49ers*, *those San Francisco 49ers*, *San*

Francisco 49ers, *the 49ers*, *those 49ers* and *49ers*). In addition, the words-with-spaces approach seems inconsistent with the internal modifiers we find in such examples as *the league-leading* (*San Francisco*) *49ers*. Full compositionality, on the other hand, runs up against gross overgeneration, as any place/organization name is allowed to combine with any appellation, yielding such non-denoting names as *the Oakland 49ers*.

2.3 Syntactically-Flexible Expressions

Whereas semi-fixed expressions retain the same basic word order throughout, syntactically-flexible expressions exhibit a much wider range of syntactic variability. We illustrate the types of variation possible in the form of verb-particle constructions, decomposable idioms and light verbs.

Verb-Particle Constructions Verb-particle constructions consist of a verb and one or more particles, such as *write up*, *look up* and *brush up on*. They can be either semantically idiosyncratic, such as *brush up on*, or compositional such as *break up* in *the meteorite broke up in the earth's atmosphere* (Bolinger 1972, Dixon 1982, Dehé et al. to appear).[3] In compositional usages, the particle(s) act as a construction and modify the spatial, aspectual, etc properties of the head verb, such as *up* transforming *eat* from an activity into an accomplishment in *eat up*. That is, the particle(s) generally assume semantics idiosyncratic to verb-particle constructions, but are semi-productive (cf. *gobble up* in the case of *up*).

Transitive verb-particle constructions take an NP argument either between or following the verb and particle(s) (e.g. *call Kim up* and *fall off a truck*, respectively). Certain transitive verb-particle constructions are compatible with only particle-initial realizations (consider **fall a truck off*), while others are compatible with both forms (e.g. *call Kim up* vs. *call up Kim*). Even with intransitive verb-particle constructions, adverbs can often be inserted between the verb and particle (e.g. *fight bravely on*). As a result, it is impossible to capture the full range of lexical variants of transitive verb-particle constructions as words-with-spaces.

As with other MWE types, a fully compositional approach is troubled by the idiomaticity and overgeneration problems. Even for seemingly synonymous verbs combining compositionally with the same particle, idiosyncrasies are observed (e.g. *call/ring/phone/telephone* vs. *call/ring/phone/*telephone up*: McIntyre 2001) which would be beyond the descriptive powers of a purely compositional account.

Decomposable Idioms Decomposable idioms, such as *let the cat out of the bag* and *sweep under the rug*, tend to be syntactically flexible to some degree.

[3] The combination *break up* also has semantically idiosyncratic senses including "adjourn" and "separate".

Exactly which types of syntactic variation a given idiom can undergo, however, is highly unpredictable (Riehemann 2001).

Because decomposable idioms are syntactically variable to varying degrees, it is hard to account for them using only syntactic selection. Instead, they act like they are composed of semantically linked parts, which thus suggests a semantic approach is appropriate (Nunberg et al. 1994). Because they are highly variable syntactically, decomposable idioms are incompatible with a words-with-spaces strategy; fully compositional techniques suffer from the idiomaticity problem.

Light Verbs Light-verb constructions (e.g. *make a mistake*, *give a demo*, **do a mistake*, **make a demo*) are highly idiosyncratic – it is notoriously difficult to predict which light verb combines with a given noun (Abeillé 1988). Although such phrases are sometimes claimed to be idioms, this seems to be stretching the term too far: the noun is used in a normal sense, and the verb meaning appears to be bleached, rather than idiomatic.

Light-verb constructions are subject to full syntactic variability, including passivization (e.g. *a demo was given*), extraction (e.g. *How many demos did Kim give?*) and internal modification (e.g. *give a revealing demo*). They thus cannot be treated as words-with-spaces. A fully compositional account, on the other hand, would be unable to model the blocking of alternative light verb formations (e.g. *give a demo* vs. **make a demo*), and thus would suffer from gross overgeneration.

2.4 Institutionalized phrases

Institutionalized phrases are semantically and syntactically compositional, but statistically idiosyncratic. Consider for example *traffic light*, in which both *traffic* and *light* retain simplex senses and combine constructionally to produce a compositional reading. Given this strict compositionality, we would expect the same basic concept to be expressible in other ways, e.g. as *traffic director* or *intersection regulator*. Clearly, however, no such alternate form exists, because the form *traffic light* has been conventionalized. The idiosyncrasy of *traffic light* is thus statistical rather than linguistic, in that it is observed with much higher relative frequency than any alternative lexicalization of the same concept. Other examples of institutionalized phrases are *telephone booth* (or *telephone box* in British/Australian English), *fresh air* and *kindle excitement*. We refer to potential lexical variants of a given institutionalized phrase which are observed with zero or markedly low frequency as **anti-collocations** (Pearce 2001).

One subtle effect observed with institutionalized phrases is that association with the concept denoted by that expression can become so strong as to diminish decomposability. *Traffic light*, for example, could conceivably be interpreted as a device for communicating intended actions to surrounding traffic. However, partly as a result of the existence of an institutionalized term for such a device (i.e. *turn*(*ing*) *signals*) and partly due to the conventionalization of *traffic light* to denote a stoplight, this reading is not readily available.

Note that we reserve the term **collocation** to refer to any statistically significant cooccurrence, including all forms of MWE as described above and compositional phrases which are predictably frequent (because of real world events or other nonlinguistic factors). For instance, *sell* and *house* cooccur in sentences more often than would be predicted on the basis of the frequency of the individual words, but there is no reason to think that this is due to anything other than real world facts.

As institutionalized phrases are fully compositional, they undergo full syntactic variability. Words-with-spaces approaches thus suffer from lexical proliferation, while fully compositional approaches encounter the idiomaticity and overgeneration problems.

3 Some Analytic Techniques

In this section we will introduce some analyses for MWEs using the constraint-based Head-driven Phrase Structure Grammar (HPSG) formalism (Pollard and Sag 1994, Sag and Wasow 1999). Most of these analyses have been implemented in grammars in the LKB grammar development environment (Copestake in press). Ultimately, we plan to include them all in the English Resource Grammar; at present some are being tested in smaller grammars.

The LKB grammar development environment is a general system for developing typed feature structure grammars which implements a particular typed feature structure logic. It is written in Common Lisp and currently runs under Linux, Solaris, Windows and MacOS. Grammar development is effectively a process of programming in a very high-level specialized language, and the system supports interactive grammar development as well as parsing and generation.

The LinGO English Resource Grammar (ERG) is a broad-coverage grammar of English described in a typed feature structure logic compatible with the LKB and several other systems. The grammar itself is written in HPSG, while the semantic representation used is Minimal Recursion Semantics (MRS hereafter – Copestake et al. 1999). An overview of the ERG (from a computational linguistic perspective) is given in Copestake and Flickinger (2000).

3.1 Analyzing Fixed Expressions

Truly fixed expressions, like *ad hoc* or *of course*, can simply be dealt with as words-with-spaces. In this case a list of words is given the same lexical type as a single word and associated with a single semantic relation. For example, in the current ERG, *ad hoc* is defined as having the type `intrans_adj_l` (intransitive adjective listeme,[4] which is also the type for simplex adjectives such as *pretty*). However, simply listing MWEs as strings, as in (1), is adequate only for expressions which allow no variability at all. The expression can be externally modified: *very ad hoc*, but not internally modified **ad very hoc*.[5]

[4] A listeme is a lexically-listed entity.

[5] This and subsequent feature structures are intended for illustrative purposes and are not as they appear in the ERG.

(1) `ad_hoc_1 := intr_adj_l &`
`[ STEM < "ad", "hoc" >,`
`SEMANTICS [KEY ad-hoc_rel ]].`

In practice, there is often an unfortunate side effect to allowing these expressions in an implementation: developers exploit this class to add entries that can vary, but don't often, in order to quickly achieve greater coverage.

3.2 Analyzing Semi-Fixed Expressions

When analyzing semi-fixed expressions, it is important to strike a balance between too weak a mechanism, which will not allow sufficient variability, and too strong a mechanism, which will allow too much. We make heavy use of existing features of our grammars, in particular multiple inheritance. We also introduce two new mechanisms: the ability to specify which words inflect in an otherwise fixed expression and the ability to treat a list of listemes as a single listeme.

Internal Inflection Some semi-fixed MWEs, such as *kick the bucket*, *part of speech* and *pain in the neck* differ from fixed expressions in that one word in them inflects, as though it were the phrasal head. In this case, it is still possible to treat the whole entry (a list of words) as a single listeme that is associated with a single semantic relation. We add a pointer showing which word to inflect (`INFL-POS` = inflection position, i.e. inflect the nth word in the `STEM` list). An entry for *part of speech*, where only the first word *part* inflects, is given in (2).

(2) `part_of_speech_1 := intr_noun_l &`
`[ STEM < "part", "of", "speech" >,`
`INFL-POS "1",`
`SEMANTICS [KEY part_of_speech_rel ]].`

The analysis can be extended to words with two inflecting parts, such as *wine and dine*, which we would like to treat as a single transitive verb, but with both *wine* and *dine* inflecting: *Kim wined and dined Sandy.*

In a deeper treatment of these expressions the list of words would be replaced with a list of listemes (`LEX-SIGNS`), so that the words can inherit their properties from existing listemes. In this case, the expression as a whole would, by default, inherit its lexical type from the designated inflecting word: thus *part of speech* would inherit from *part* and would be a count noun, while *fool's gold* would inherit from *gold* and would be a mass noun. This inheritance allows us to capture the generalization that a *performance artist* is a kind of *artist* though the use of *performance* is non-compositional.

Hierarchical Lexicon with Default Constraint Inheritance Default inheritance allows us to simplify the structure of the lexical types used. For example, by default, proper names in English take no determiner. In our analysis, we handle this by requiring the specifier (`SPR`) list to be empty, as in (3a). However,

some names, such as those of U.S. sports teams, normally take a definite determiner. Therefore, the constraint on `Name` is defeasible: it can be overridden in rules that inherit from it. The logic for defaults we assume follows Lascarides and Copestake (1999), where default values are indicated by '/'.

The type `USTeamName` overrides the default, in this case, by specifying that the specifier must be a definite determiner, and that the number defaults to plural, as shown in (3b):

(3) a `Name`: [`SPR` / ⟨ ⟩]
b `USTeamName`: [`SPR` ⟨ `Det`[`definite`] ⟩, `NUM` / `plural`]

The specifier is not given as the listeme *the*, but just as the specification `definite`. In the absence of other information this would normally be the definite article,[6] but other definite determiners are also possible: *How about those Raiders?*

The listeme for *the Oakland Raiders*, would thus be of the type `USTeamName` and described as a list of listemes, inherited from *Oakland* and *Raiders*. This analysis captures the fact that the first word is the same as the place *Oakland*. The structure is shown in (4), where `oakland_1` and `raiders_1` are listeme identifiers for the place *Oakland* and the appellation *Raiders*:[7]

```
(4) oakland_raiders_1 := USTeamName &
      [  LEX-SIGNS / < oakland_1, raiders_1 >,
         SEMANTICS < oakland_raiders_rel > ].
```

Note further that there are exceptions to the subregularity of sports team names. Certain teams have names that are combinations of determiner plus mass noun, such as *the* (*Miami*) *Heat*, *the* (*Philadelphia*) *Charge*, and *the* (*Stanford*) *Cardinal*.[8] Since mass nouns are singular, the appropriate constraint on the subtype `MassTeamName` overrides the defeasible [`NUM` / `plural`] specification in (3b).

The `USTeamName` type, as it is presented here, still does not capture (i) the optionality of *Oakland* and (ii) the fact that the first word in team names is typically a place or organization. Two analyses suggest themselves. In the first of these, the lexical type `USTeamName` licenses an optional second specifier, in addition to the determiner. This specifier would be the appropriate place name or organization. In the second possible analysis, an extremely circumscribed construction, inheriting from the noun-noun compound phrase rule, would license combinations headed by listemes of the type `USTeamName` with a modifier that must be a place or organization. It remains to be seen whether either of these proposals is viable.

---

[6] Obtainable by setting *the* to be the default definite determiner.

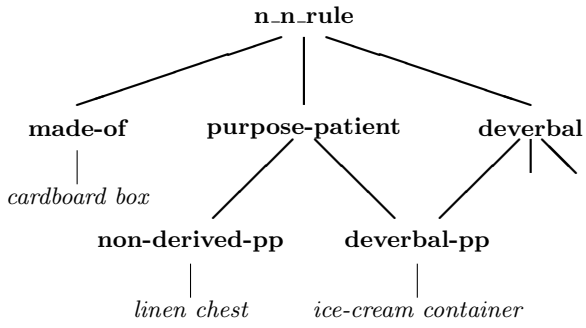
[7] Inheritance from identifiers diverges from standard HPSG practice, but see Copestake (1992) for formalization and motivation.

[8] This name refers to the color, not the bird.

3.3 Analyzing Syntactically-Flexible Expressions

Many of the syntactically-flexible MWEs can again be handled by existing mechanisms: the use of **circumscribed constructions** and lexical selection. We introduce a new mechanism to handle the most variable decomposable idioms, that allows us to check that all the idiomatic parts are there in the appropriate semantic relationships.

Circumscribed Constructions Inheritance hierarchies of constructions for noun-noun compounds can be used to capture some of the semi-productivity of syntactically-flexible expressions (Copestake and Lascarides 1997). The idea is that compounds like *spring beginning* (cf. (*the*) *beginning of spring*) are not completely blocked, but they are prevented from having any conventional interpretation, and will be interpreted as incoherent unless licensed by a specific discourse context. The diagram below shows a fragment of the compound nominal construction hierarchy adapted from that paper, with example compounds corresponding to the various categories at each leaf node:



This hierarchy allows generalizations about productive and lexicalized forms to be represented: for productive forms, the construction is interpreted as a grammar rule, while lexicalized forms stipulate the construction as part of their entry. The use of defaults allows generalizations about stress, for instance, to be expressed.

Lexical Selection Verb-particle constructions, conjunctions like *either... or...* and so on, where material intervenes between the elements of the phrase, can be accounted for by means of a lexical selection mechanism where a sign associated with one word of the phrase selects for the other word(s). For instance, in the existing ERG, there is an entry for *hand* which subcategorizes for *out*, as shown in (5):

(5)
```
hand_out_v1 := mv_prep_particle_np_l &
    [ STEM < "hand" >,
      SEMANTICS [ KEY hand_out_rel,
                  --COMPKEY out_rel ] ].
```

The semantics of the whole expression is given in the KEY relation (hand_out_rel); the verb *hand* then selects for the preposition whose KEY relation is given by COMPKEY (out_rel). This allows:

(6) Kim handed out chocolate to the kids.

A lexical rule permutes the subcategorization list to allow:

(7) Kim handed the chocolate out to the kids.

Combinations with prepositions, such as *rely on*, *fond of* or *report on/about* can be handled in a similar manner, by selecting for the semantic relation encoded by the preposition. Early HPSG accounts of preposition selection used a PFORM (PREPOSITION-FORM) feature for this (Pollard and Sag 1994). The atomic values of PFORM simply encoded the phonetic form of the preposition. The ERG uses the basic semantic KEY relations. Either analysis allows prepositions to be grouped together into regularized types, which allows natural classes of prepositions to be selected.

Light Verbs Light verbs, that is those verbs which cooccur with certain classes of nouns, can also be handled by selection. All nouns which can be used with a given light verb will have semantic types which inherit from the same type (for example mistake_rel inherits from make_arg_rel). The light verb *make* then has the selectional restriction that its direct object must be of the type make_arg_rel). Another light verb, such as *do*, does not select for make_arg_rel, and thus will not allow **do a mistake*. Nouns which can be used with more than one light verb multiply inherit from the relevant classes. The normal mechanisms of the grammar will allow for the selectional restrictions to be passed along through long distance dependencies such as in *the mistakes that he managed to make were incredible*.

Decomposable Idioms Selection works if the syntactic relationship of the various parts of the phrase is fixed, as it indeed seems to be for verb particle constructions, but the mechanism runs into problems with some idioms, for instance, where the relationship between the words may be very flexible.

We start from the assumption that the relationship between words in decomposable idioms can be captured using a partially semantic mechanism, essentially following the approach described by Nunberg et al. (1994). The flat MRS representation adopted in the ERG is especially suited to this. Riehemann (2001) describes one approach that uses MRS; here we sketch another, which builds directly on ideas first presented in Copestake (1994).

Consider, for instance, the idiom *cat out of the bag* which can be described as a phrase containing the semantic relationships in (8), where `i_cat` and `i_bag` are the meanings corresponding to the idiomatic senses of *cat* "secret" and *bag* "hiding place".

(8) [$\texttt{i\_cat}(x) \wedge \texttt{i\_bag}(y) \wedge \texttt{out}(x, y)$]

This semantic representation is flexible enough to cover the most common forms of this idiom. The problem is that matching this specification to a conventional semantic representation is arbitrarily complex, because of the possible contributions of quantifiers and so on. In order to get this sort of idea to work, Pulman (1993) proposes an approach which relies on a form of quasi-inference operating on a compositionally derived logical form. However, his approach fails to allow for any syntactic idiosyncrasy among idioms.

Copestake (1994) develops a treatment of decomposable idioms that is semantically based, but which uses a notion of **idiomatic construction** to accommodate syntactic flexibility. Instead of locating interpretational idiosyncrasy in idiomatic listemes (e.g. *let*) that select for other such listemes (e.g. *the* and *cat*), this approach allows listemes to combine constructionally by ordinary syntactic means. However, idiomatic constructions provide an independent dimension of phrasal classification where idiomatic interpretations are assigned just in case the right pieces (e.g. *the*, *cat*, *out*, *of*, *the*, *bag*) are all present and in the right predicate-argument relations. Because the account is based on MRS, where the semantics is represented in terms of bags of predications, rather than representations with complex embeddings, it becomes natural to state a constraint requring that a given set of predications be present and appropriately related (e.g. the argument of *cat*'s predication must also be the first argument of the *out* predication). In this way, quantification and modification of pieces of idioms are allowed, as is reordering of idiomatic elements from their canonical position. This constructional approach thus differs from earlier lexical approaches, but retains the notion that there is a dependency among the lexical parts of decomposable idioms.

3.4 Information about Frequency

The treatment of frequency is different in type from the analyses described above. The grammatical rules constrain the space of possible sentences and interpretations, while frequency-based probabilities allow us to predict which of these is the preferred interpretation or string. In order to use probabilities in both analysis (from strings to meanings) and generation (from meanings to strings), we need frequency information about both semantic relations and construction rules, in so far as they contribute to semantic interpretation. The necessity of semantic frequency information has been somewhat neglected in current NLP research, no doubt largely because it is difficult to collect.

Johnson et al. (1999) describe a potentially viable approach to developing probabilistic grammars based on feature structures; Hektoen (1997) suggests an

alternative model of semantic probabilities. Both of these are possible approaches to institutionalized phrases because of the fine granularity we assume for relations in MRS. For instance, `fine_rel` and `good_rel` are distinct, so the relative frequency of *fine weather* versus *good weather* could be considered in terms of their semantic relations.

The question of determining the preferred interpretation is sometimes regarded as outside the scope of a formal linguistic account, but we believe that frequency information should be regarded as part of a speaker's knowledge of language. In any case, its utility in natural language processing is beyond question.

4 Conclusion

In this paper we hope to have shown that MWEs, which we have classified in terms of lexicalized phrases (made up of fixed, semi-fixed and syntactically flexible expressions) and institutionalized phrases, are far more diverse and interesting than is standardly appreciated. Like the issue of disambiguation, MWEs constitute a key problem that must be resolved in order for linguistically precise NLP to succeed. Our goal here has been primarily to illustrate the diversity of the problem, but we have also examined known techniques — listing words with spaces, hierarchically organized lexicons, restricted combinatoric rules, lexical selection, idiomatic constructions, and simple statistical affinity. Although these techniques take us further than one might think, there is much descriptive and analytic work on MWEs that has yet to be done. Scaling grammars up to deal with MWEs will necessitate finding the right balance among the various analytic techniques. Of special importance will be finding the right balance between symbolic and statistical techniques.

References

Abeillé, Anne: 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*.

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

Bolinger, Dwight, ed.: 1972, *Degree Words*, the Hague: Mouton.

Charniak, Eugene: 2001, 'Immediate-head parsing for language models', in *Proc. of the 39th Annual Meeting of the ACL and 10th Conference of the EACL (ACL-EACL 2001)*, Toulouse.

Copestake, Ann: 1992, 'The representation of lexical semantic information', Ph.D. thesis, University of Sussex.

Copestake, Ann: 1994, 'Representing idioms', Presentation at the HPSG Conference, Copenhagen.

Copestake, Ann: in press, *Implementing Typed Feature Structure Grammars*, Stanford: CSLI Publications.

Copestake, Ann & Dan Flickinger: 2000, 'An open-source grammar development environment and broad-coverage English grammar using HPSG', in *Proc. of the Second conference on Language Resources and Evaluation (LREC-2000)*, Athens.

Copestake, Ann, Dan Flickinger, Ivan Sag & Carl Pollard: 1999, 'Minimal recursion semantics: An introduction', (`http://www-csli.stanford.edu/~aac/papers/newmrs.ps`), (draft).

Copestake, Ann & Alex Lascarides: 1997, 'Integrating symbolic and statistical representations: The lexicon pragmatics interface', in *Proc. of the 35th Annual Meeting of the ACL and 8th Conference of the EACL (ACL-EACL'97)*, Madrid, pp. 136–43.

Dehé, Nicole, Ray Jackendoff, Andrew McIntyre & Silke Urban, eds.: to appear, *Verb-particle explorations*, Mouton de Gruyter.

Dixon, Robert: 1982, 'The grammar of English phrasal verbs', *Australian Journal of Linguistics*, **2**: 149–247.

Fellbaum, Christine, ed.: 1998, *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press.

Hektoen, Eirik: 1997, 'Probabilistic parse selection based on semantic cooccurrences', in *Proc. of the 5th International Workshop on Parsing Technologies (IWPT-97)*, MIT, pp. 113–122.

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

Johnson, Mark, Stuart Geman, Stephan Canon, Zhiyi Chi & Stefan Riezler: 1999, 'Estimators for stochastic "unification-based" grammars', in *Proc. of the 37th Annual Meeting of the ACL*, University of Maryland, pp. 535–541.

Lascarides, Alex & Ann Copestake: 1999, 'Default representation in constraint-based frameworks', *Computational Linguistics*, **25**(1): 55–106.

McIntyre, Andrew: 2001, 'Introduction to the verb-particle experience', Ms, Leipzig.

Nunberg, Geoffery, Ivan A. Sag & Thomas Wasow: 1994, 'Idioms', *Language*, **70**: 491–538.

Oepen, Stephan, Dan Flickinger, Hans Uszkoreit & Jun-ichi Tsujii: 2000, 'Introduction to the special issue on efficient processing with HPSG: methods, systems, evaluation', *Natural Language Engineering*, **6**(1): 1–14.

Pearce, Darren: 2001, 'Synonymy in collocation extraction', in *Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, CMU.

Pollard, Carl & Ivan A. Sag: 1994, *Head Driven Phrase Structure Grammar*, Chicago: University of Chicago Press.

Pulman, Stephen G.: 1993, 'The recognition and interpretation of idioms', in Cristina Cacciari & Patrizia Tabossi, eds., *Idioms: Processing, Structure and Interpretation*, Hillsdale, NJ: Lawrence Erlbaum Associates, chap. 11.

Riehemann, Susanne: 2001, 'A constructional approach to idioms and word formation', Ph.D. thesis, Stanford.

Sag, Ivan A. & Tom Wasow: 1999, *Syntactic Theory: A Formal Introduction*, Stanford: CSLI Publications.